Nützlichkeit und Nutzen der Programmevaluationen im Bereich der österreichischen FTI-Politik. Metaevaluation der Programmevaluationen 2003-2014.

Studie im Auftrag des Rats für Forschung und Technologieentwicklung (RFTE)

Endbericht

Dezember 2015

Günther Landsteiner

Der vorliegende Bericht wurde im Rahmen eines freien Auftragsverhältnisses erstellt von:

MMag. Günther Landsteiner
Unabhängiger Forscher-Berater und Evaluator
Zieglergasse 75 top 6
A–1070 Wien
mail: g.landsteiner@chello.at

Der Bericht stellt ausschließlich die fachlichen Ansichten und Meinungen des Autors dar, die von denen des Auftraggebers abweichen können.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Anhänge

Kurzbericht

1. Auftrag und Vorgehensweise der Metaevaluation

1.1. Motivation

In Österreich wurde über Jahre hinweg eine Evaluationskultur im forschungs-, technologie- und innovationspolitischen Bereich aufgebaut, wobei die Gründung der Plattform fteval 1997 ein entscheidendes Datum darstellt. Die Anzahl an Programmevaluationen, die seitdem durchgeführt wurden, ist im internationalen Vergleich für den Politikbereich jedenfalls als überdurchschnittlich zu erachten. Freilich sind mit der zunehmenden Entfaltung des Unternehmens FTI-Evaluation auch immer wieder Stimmen in der Akteursarena laut geworden, die Enttäuschungen gegenüber ursprünglich hohen Erwartungen thematisieren. Zugleich ist in den letzten Jahren eine internationale Entwicklung zu beobachten, die Frage nach der Wirksamkeit und dem Reflexionspotenzial von Evaluationen im FTI-Bereich aufzuwerfen. Parallel dazu haben über den FTI-Bereich hinaus Evaluationsgesellschaften die Frage der Nützlichkeit von Evaluation und des Wissens über die Evaluationspraxis auf die Tagesordnung gesetzt.

Der RFTE hat deswegen die vorliegende Studie zur Nützlichkeit und zum tatsächlich feststellbaren Nutzen der Programmevaluationen im österreichischen FTI-Bereich beauftragt, mit der der unabhängige externe Evaluator im Juli 2014 betraut wurde. Die Studie zielt darauf ab, evidenzbasierte Einschätzungen darüber zu gewinnen, wie sich die Nützlichkeit der Programmevaluationen gestaltet, und inwieweit etwaiges Verbesserungspotenzial besteht. Die Studie findet dazu ihre Basis in einem allgemein anerkannten, internationalen Wissenskorpus über qualitätsvolle Evaluation und spezifische Aspekte der Evaluationsnutzung. Die konkrete Herangehensweise wurde als diejenige einer Metaevaluation definiert, die um Komponenten einer Nutzungsforschung zur Evaluation angereichert ist. Neben der Frage, welche Grundlagen für die Entstehung von Evaluationsnutzen mit den durchgeführten Programmevaluationen gelegt wurden, wird auch die Frage nach nachvollziehbaren Nutzungsweisen der Programmevaluationen verfolgt. Analysiert wird die Evaluationspraxis im FTI-politischen Bereich über den zwölfjährigen Zeitraum 2003–2014, soweit sie Programmevaluationen betrifft.

1.2 Vorgehensweise und Methodeneinsatz

Metaevaluationen stellen systematische Analysen von einer oder mehreren Evaluationen dar, die der Bestimmung von Wert und Güte dieser Evaluationen gelten. Der vorliegenden Studie geht es somit nicht um die inhaltlichen Ergebnisse, die Programmevaluationen erbracht haben, sondern um Gestaltungsweisen der Evaluationsprozesse und Evaluationsberichte, die als essentielle Voraussetzungen und Merkmale der Nützlichkeit von Evaluationen zu erachten sind. Hier verkörpert sich ein Nutzungspotenzial, das die Grundlage für faktische Nutzungsweisen bildet und als solches hinsichtlich von Stärken und Schwächen analysiert werden kann. Nutzungsforschung zur Evaluation identifiziert Ausmaß und Art von Evaluatinsnutzen und Einflussfaktoren, die das Zustandekommen dieses Nutzens fördern oder behindern.

Mit den „Standards für Evaluation“ der deutsch-österreichischen Evaluationsgesellschaft DeGEval liegt ein hervorragender Bezugspunkt für eine neutrale und breit abgesicherte Thematisierung von Qualitätsaspekten einer Programmevaluation vor, die explizit mit deren Nützlichkeit verbunden sind. Die „Standards für Evaluation“ verstehen sich als Anleitung für zielgerichtete professionelle Evaluation in allen Politikfeldern und als fachliches Dialoginstrument für einen Austausch über die Qualität von professionellen Evaluationen. Sie zählen zu einer Standardfamilie, die auch die US-amerikanischen „Program Evaluation Standards“ und die schweizerischen SEVAL-Standards umfasst. Die Wahl dieses international hoch relevanten, mit der übergreifenden Gesamtentwicklung von Evaluationstheorie und –forschung verbundenen Bezugspunktes erfolgte in bewusster Abhebung von den Standards der Plattform fteval, die sich die Community der österreichischen FTI-EvaluatorInnen und Auftraggeberinstitutionen gegeben hat. Eine vergleichende Auseinandersetzung mit den fteval-Standards oder deren Kommentierung war nicht Bestandteil des Auftrags. Die Anwendung der Standards und die Vorgehensweise der Metaevaluation wurden in einem Workshop mit den beiden Hauptexperten für die DeGEval- und SEVAL-Standards erörtert. Auf der Grundlage von Überlegungen sowohl konzeptueller als auch pragmatischer Art wurde ein Kriterienset bestimmt, das sich zunächst aus 10 DeGEval-Standards zusammensetzte und während der Durchführung der Evaluation auf 11 Standards erweitert wurde, um erkannten Bedarfslagen noch besser gerecht zu werden.

Für einen summativen und explorativen Zugriff auf die langjährige Evaluationspraxis im Kontext begrenzter Ressourcen wurde eine Stichprobe von 20 publizierten Evaluationsberichten auf der Basis eines mit dem Auftraggeber abgestimmten *theoretical sampling* gezogen. Damit wurde knapp die Hälfte der publizierten Evaluationsberichte zu Programmevaluationen in diesem Zeitraum erfasst. Zur Ergänzung der Informationslage über Evaluationsprozesse und zur Gewinnung von Informationen über tatsächliche Nutzungen von Programmevaluationen wurden eine Online-Befragung unter FTI-EvaluatorInnen sowie Interviews mit AuftraggeberInnen im FTI-politischen Governancesystem und Hauptakteuren der Plattform fteval durchgeführt. Für die EvaluatorInnen-Befragung wurden Inhalte der Standards und der internationalen Nutzungsforschung zur Evaluation operationalisiert, für die Interviews wurden diese Inhalte einem Gesprächsleitfaden zugrunde gelegt. 273 EvaluatorInnen und MitarbeiterInnen von relevanten Instituten in Österreich und im deutsch- und englischsprachigen Ausland, die für die Durchführung einer FTI-Programmevaluation im Beobachtungszeitraum in Frage kamen bzw. bekannt waren, wurden in der Online-Befragung kontaktiert. 37 EvaluatorInnen aus dem In- und Ausland haben die Umfrage beantwortet, wobei rund die Hälfte häufig im österreichischen FTI-Bereich tätige Evaluatorinnen sind, sodass die erhaltenen Umfragedaten ein gut belastbares Bild der österreichischen FTI-Evaluationspraxis liefern. Halbstrukturierte Interviews im Umfang von bis zwei Stunden wurden mit 15 Personen im Bereich der Auftraggeber und HauptsdressatInnen von Programmevaluationen geführt, mit gleichmäßiger Abdeckung der relevanten Ressorts und Agenturen auf Bundesebene.

In der Berichtsanalyse wurde auch die Frage verfolgt, ob für Nützlichkeit relevante Qualitätsmerkmale einer zeitlichen Entwicklung unterlegen sind. Alle eingesetzten Methoden waren darauf ausgerichtet, eine langjährige und vielfältige Evaluationspraxis summativ auffassen zu können und dabei auch Weiterentwicklungen und kumulative Effekte sichtbar machen zu können. Die Ergebnisse aus den drei Erhebungsverfahren werden im Folgenden in synthetisierender, an der Extraktion von Hauptcharakteristika der Evaluationspraxis ausgerichteter Weise dargestellt. Die folgende Darstellung zentraler Ergebnisse und Schlussfolgerungen bezieht sich ausschließlich auf Programmevaluationen im österreichischen FTI-Bereich und kann nicht unmittelbar verallgemeinert werden, auch wenn sich FTI-Evaluation über die letzten beiden Jahrzehnte europaweit entfaltet hat und internationalisierte Verständnisweisen der evaluativen Vorgehensweise im FTI-politischen Bereich nicht von der Hand zu weisen sind.

2. Zentrale Ergebnisse

2.1 Nutzung von Programmevaluationen

Bisherige Programmevaluationen haben durchaus Nutzen genereiert. Aus den Auskünften der AuftraggeberInnen und HauptaddressatInnen der Programmevaluationen und denen der EvaluatorInnen ergibt sich hier ein hochgradig konsistentes Bild. Dabei stehen Nutzungsweisen im Vordergrund, die von der Evaluationsforschung als instrumenteller und konzeptueller Nutzen von Evaluation bezeichnet werden. Auf Basis von Datenlagen, Schlussfolgerungen und Empfehlungen wurden Entscheidungen über Programme getroffen und neue Sichtweisen gewonnen, die zur Nachschärfung von Programmen oder zur Bereinigung von Zielkatalogen geführt haben. Entscheidungen über evaluierte Programme betrafen vor allem Adjustierungen von weiterlaufenden Programmen. Derartige Entscheidungen können sowohl zu Umsetzungsaspekten in den betrauten Agenturen fallen, entsprechend ihres jeweiligen Pouvoirs zum evaluierten Programm, oder seitens der Programmeigentümer für eine nachfolgende Programmphase in Programmdokumenten niedergelegt werden. Hinzu kommen konzeptuelle Einsichten über Merkmale von FTI-Segmenten ebenso wie über das Management von Programmen, die häufig auch da eintraten, wo nicht für außenstehende BeobachterInnen leicht erkennbare Entscheidungen gefällt wurden.

Die Programmevaluationen haben immer wieder Lerneffekte erzeugt, in deren Gefolge Themen verankert und Annahmen über Funktionsweisen von Teilen des FTI-Systems und der darauf gerichteten Steuerungs- und Anreizsysteme verändert wurden. Auch Evaluationen von Programmen, die in der Folge nicht weitergeführt wurden, haben solche systemisch wertvollen Einsichten und Lerneffekte erbracht. Evaluative Erkenntnisse zu einzelnen Programmen haben sich auf Konzeption und Gestaltung thematisch benachbarter Programme ebenso ausgewirkt wie auf Gestaltungsweisen von anderen Programmen im Portfolio derselben Agentur.

Nicht zuletzt ist es in der Entwicklung der Evaluationskultur zu organisatorischen Anpassungen gekommen, die den Umgang mit Evaluationen und deren Ergebnissen unterstützen. Insbesondere wurde in einer Agentur rezent ein Managementprozess für den gezielten Umgang mit Evaluationsergebnissen eingeführt, und in einer weiteren Agentur während der Laufzeit der vorliegenden Metaevaluation die Planung, Durchführung und Präsentation von Evaluationen im Rahmen eines übergreifenden Monitoring- & Evaluationssystems weiter professionalisiert. Organisatorische Anpassungen, die die Evaluationskapazität erhöhen, liegen auch an anderen Systemstellen vor, in unterschiedlichem Ausmaß und insgesamt nicht auf einem gleichen Niveau.

In allen Gesprächen mit AuftraggeberInnen und in der EvaluatorInnen-Befragung wurde ersichtlich, dass die über die Jahre durchgeführten Programmevaluationen als wesentliche Beiträge zu einer Verbreiterung und Vertiefung der Wissensbasis eingeschätzt werden, auf die sich FTI-politisches Handeln gerade auch als aktualitätsbezogenes und voranschreitendes Handeln in dynamischen Umwelten stützt. Zugleich wird auch deutlich, dass es sich beim Eintreten von Nutzen aus Programmevaluationen um Gemengelagen handelt, sodass nicht nur eine Evaluation zu einer Nutzung führt, sondern multiple Effekte auftreten. Freilich handelt es sich bei Umsetzungen von Erkenntnissen aus Programmevaluationen nicht um Automatismen, sondern um Handlungsweisen in Multiakteurs-Konstellationen im Einzelfall, bei denen auch immer wieder Reibungsverluste auftreten. Evaluationsnutzungen sind deutlich von den Konfigurationen der Principal-Agent-Beziehungen geprägt, die sich für die parallel agierenden Segmente des politisch-administrativen Handelns im FTI-Bereich unterschiedlich darstellen. Ob und wie Evaluationsergebnisse in einem dieser Steuerungsbereiche auch über den Kreis der unmittelbar mit einem Programm befassten Personen breiter bekannt gemacht und aufgegriffen werden, erweist sich als unsystematisch und stark vom Engagement von Einzelpersonen abhängig. Zusätzlich erhöht wird die Komplexität der Nutzenentstehung im Überstieg zwischen der administrativen und der politischen Sphäre, wobei auch hier von einer beträchtlichen Variation von Einzelfall zu Einzelfall auszugehen ist.

Als deutlich weniger ausgeprägt erweist sich eine Nutzung von Evaluationsergebnissen, die über die Grenzen der jeweiligen Steuerungsbereiche mit ihren Principal-Agent-Verhältnissen hinaus reicht. Obwohl auch hier relevante Wissenszuwächse beschrieben werden und in den Steuerungsbereichen jeweils davon ausgegangen wird, dass interessierende Information aus anderweitig durchgeführten Evaluationen zumindest prinzipiell zugänglich ist, gibt es keinen systematischen Vorgang im FTI-politischen Governance-System, der das Aufgreifen von in anderen Zuständigkeitsbereichen erbrachten Evaluationsergebnissen und die Auseinandersetzung damit unterstützt. Ausstrahlungswirkungen von evaluativer Information auf interessierte Fachöffentlichkeiten bzw. auf Akteursgruppen in der FTI-Landschaft treten in noch geringerem Maß auf und müssen als volatil gelten, da sie abgesehen von der grundsätzlichen Verfügbarmachung derjenigen Evaluationen, zu denen die Berichte publiziert werden, in aller Regel nicht Gegenstand gezielter Vorgehensweisen sind.

Im FTI-politischen System dienen Programmevaluationen auch durchaus dazu, andere Akteure in der politischen Sphäre vom evaluierten Programm zu überzeugen oder Entscheidungen über Programme zu rechtfertigen („symbolischer Nutzen“). Hier geht es um eine Überzeugungsarbeit, die in der Multiakteurs-Arena eines von differenzierten Principal-Agent-Verhältnissen, Hierarchien und Kleinteiligkeit geprägten Systems stets zu leisten ist, wenn es um die Zukunft von FTI-Programmen bzw. Steuerunginstrumentarien geht. Zum Phänomenkomplex der Erzeugung von Evaluationsnutzen zählt freilich auch die Art der Verankerung der Evaluationsfunktion im rechtlich-institutionellen Rahmen der Bundesverwaltung. An der Schnittstelle zwischen Fachressorts und dem Bundesministerium für Finanzen (BMF) kommt den Programmevaluationen eine Funktion der Legitimation der Mittelausgaben zu. Diese Legitimationsfunktion ist im Motivbündel für die Planung und Durchführung von Programmevaluationen stets anwesend. Die Daten zeigen, dass eine symbolische Nutzung von Programmevaluationen andere Nutzenformen keineswegs ausschließt. Allerdings sorgt die gleichzeitige Anwesenheit von Lern- und Legitimationsfunktion für eine innere Spannung in jedem Evaluationsprojekt, die sich letztlich für eine Evaluationspraxis, die sich an der Erzeugung systematischer Wissenszuwächse im Governance-System orientiert, eher abträglich erweist.

Die Ergebnisse der Programmevaluationen können in ihrer Rolle als FTI-politische Informationsmittel als konkurrenzlos gelten, wenn sie auch oftmals nicht die alleinigen Grundlagen FTI-politischer Entscheidungen über Einsätze und Mittelzuweisungen sind. Im Verhältnis zu dieser Rolle von Evaluationsergebnissen ist das grundsätzliche Potenzial, bereits während der Evaluations-

durchführung und unabhängig von den Evaluationsergebnissen aus der Durchführung von Programmevaluationen unmittelbar zu profitieren („Prozessnutzen"), bislang nur wenig ausgeschöpft worden.

2.2. Einflussfaktoren auf die Nutzung von Programmevaluationen

Faktoren, die in der bisherigen Evaluationspraxis Einfluss darauf gehabt haben, ob und wie Evaluationsergebnisse auch genutzt wurden, siedeln sich sowohl im Bereich dessen an, was innerhalb eines Evaluationsprojekts durch dessen Ausgestaltung beeinflusst werden kann, als auch im Bereich des umgebenden Kontexts, der außerhalb dieses Einflussbereichs verbleibt. Anhand von Daten aus der EvaluatorInnen-Befragung können die 20 wesentlichsten Einflussfaktoren bestimmt und gereiht werden. Sie finden in Auskünften von AuftraggeberInnen ihre Entsprechungen, wobei naturgemäß auch Perspektivunterschiede existieren sind und die EvaluatorInnen auch Faktoren bewertet haben, über die AuftraggeberInnen so nicht gesprochen haben. Diese 20 wesentlichsten Einflussfaktoren verteilen sich zu gleichen Teilen auf intrinsische Evaluationsmerkmale und auf Kontextfaktoren.

Unter jenen Faktoren, die sich im Verantwortungsbereich einer einzelnen Programmevaluation ansiedeln, steht die Glaubwürdigkeit der EvaluatorInnen bei den AuftraggeberInnen an erster Stelle. Diese Glaubwürdigkeit wird im österreichischen FTI-politischen Bereich durch die Heranziehung von auf FTI-Evaluation spezialisierten Instituten im In- und Ausland sowie durch die kontinuierliche Kommunikation von österreichischen FTI-EvaluationsspezialistInnen mit den Auftrggeber-Insitutionen in der Plattform fteval hergestellt. Ähnlich wichtig ist die Klarheit der Berichterstattung (Klarheit der Berichtsaussagen, Vorhandensein eines Executive Summary und dessen Aussagekraft), die von den FTI-EvaluatorInnen ebenso wie von deren AuftraggeberInnen als zentral erachtet wird.

Unter den Kontextfaktoren rangiert die Erwartung der AuftraggeberInnen, dass ihnen die konkrete Programmevaluation von Nutzen sein wird, an erster Stelle. Die bereits genannte Kombination von Lern- und Legitimationszwecken in der institutionell-rechtlichen Verankerung der Programmevaluationen kann als ein wesentlicher Grund dafür angesehen werden, dass deutliche Unterschiede im Umgang mit verschiedenen Programmevaluationen zu bemerken sind und immer wieder Fälle eingetreten sind, in denen Programmevaluationen von vornherein von ihren AuftraggeberInnen als notwendige Übungen betrachtet und kaum mit Nutzenerwartungen verbunden wurden, was sich dann von der Evaluationsplanung weg bis hin zum Umgang mit den Ergebnissen niederschlägt. Ebenfalls bedeutend für Art und Ausmaß der Nutzung von Evaluationsergebnissen ist der Umstand, ob eine Programmevaluation in direktem Konnex mit einem aktuellen FTI-politischen Entscheidungsbedarf oder Problemdruck steht. Ein derartiger Konnex besteht primär durch einen vorgegebenen Evaluations- und Verhandlungsrhythmus für Programmvereinbarungen der Ressorts mit dem Finanzministerium (BMF), während aktuelle FTI-politische Themenstellungen oder .„Windows of Opportunity" kaum als solche zum Auslöser von direkt auf sie gemünzten Evaluationsaktivitäten werden. Aktuelle Informationsbedürfnisse der Ressorts und Agenturen werden innerhalb dieses Rahmens des Öfteren nur in eingeschränkter Weise befriedigt.

Eine starke Personenabhängigkeit der genaueren Umgangsweise mit einzelnen Programmevaluationen und ihren Ergebnissen tritt in der EvaluatorInnen-Umfrage mit vier Faktoren massiv zutage. Bei diesem „Human Factor" in der Evaluationsnutzung geht es um die persönlichen Denkstile der jeweils Evaluationzuständigen, um die Konsistenz der Evaluationsergebnisse mit ihren Sichtweisen und Erwartungen, um ihre Erfahrenheit mit Evaluation, und um ihre Rolle in der jeweiligen Institution. Des Weiteren kommt organisatorischen Anpassungen, Ressourcen und Erfahrungen der auftraggebenden Institutionen ein erheblicher Stellenwert zu. AuftraggeberInnen haben hierauf mindestens ebenso stark hingewiesen wie die EvaluatorInnen, für die diese Kontextfaktoren mit zu den einflussreichsten zählen.

Einige hoch relevante evaluationsmethodische Gesichtspunkte wie die Angemessenheit der Evaluationskriterien, eine ausgewogene Darstellung von Stärken und Schwächen des untersuchten Programms oder die Art des Evaluationsansatzes sind in den 20 wesentlichsten Einflussfaktoren auf eine Nutzenentstehung aus der Sicht der EvaluatorInnen enthalten. Sie fallen jedoch im Gesamtbild hinter einige stärkere Einflussfaktoren merklich zurück, die durch die Vorgehensweise einer Evaluation nicht beeinflusst werden können. Methodenaspekte im engeren Sinn, wie die Anwendung eines Methodenmix, Triangulation oder die Finesse, mit der bestimmte Methoden eingesetzt werden, kommen unter den 20 wesentlichsten Einflussfaktoren auf Evaluationsnutzung, so wie die EvaluatorInnen sie einschätzen, nicht vor. AuftraggeberInnen sind auf evaluationsmethodische

Aspekte nicht in einer vergleichbaren Detailliertheit eingegangen, haben aber doch gelegentlich auf Mängel hingewiesen, die in der Vergangenheit die Entstehung von Nutzen aus Programmevaluationen beeinträchtigt haben und in den Einzugsbereich der Methodenanwendung fallen. Insgesamt erhärtet sich das Bild, dass die traditionell vor allem in Methodendiskussionen verankerte FTI-Evaluation die tatsächliche Entstehung von Nutzen aus durchgeführten Evaluationen nur in untergeordneter Weise diesem Hauptfokus ihrer Thematisierung der evaluatorischen Vorgehensweisen verdankt.

2.3. Nützlichkeit der Evaluationsberichte und Evaluationsprozesse

Die analysierten Evaluationsberichte entsprechen den herangezogenen DeGEval-Standards auf einem im Großen und Ganzen mittleren Niveau, und mit voranschreitender zeitlicher Entwicklung zunehmend besser. Verbesserungspotenzial ist dennoch vorhanden, wenn es um bestmögliche Programmevaluationen geht, die hohe Nützlichkeit erzielen und das im "Unternehmen Programmevaluation" angelegte Potenzial bestmöglich ausschöpfen. Eine sehr gute Erfüllung eines der 11 herangezogenen Standards konnte nur in einigen wenigen Fällen attestiert werden. Ebenso selten ist zugleich eine völlige Nichterfüllung eines der Standards, die auch in den letzten Jahren nicht mehr auftritt. Während zu allen herangezogenen Standards grundsätzlich noch Verbesserungspotenzial besteht, erscheinen die folgenden Gesichtspunkte als die relevantesten, um künftig noch nützlichere Programmevaluationen zu erzielen.

Die analysierten Programmevaluationen waren mit Ausnahme einer ex post-Evaluation Zwischenevaluationen oder Teile von Begleitevaluationen. Sie waren in den meisten Fällen sehr breit angelegt, Fragen von der Relevanz der Programme über ihre Effektivität bis hin zu ihrer Wirkung sollten verfolgt werden (sogenannte Multi-Purpose Evaluationen). Es wurden Outputs, Outcomes, und erste Wirkungen der Programme untersucht, sodass Erkenntnisse über die Programme durchaus erzielt wurden. Die Beobachtbarkeit von Programmwirkungen war auf Grund der früh gewählten Evaluationszeitpunkte fast immer deutlich eingeschränkt. Jedoch ist auch hinsichtlich dessen, was zu den Evaluationszeitpunkten bereits grundsätzlich zu den Programmen beobachtbar war, festzustellen, dass in vielen Fällen nicht von einer umfassenden und gründlichen Aufarbeitung der Programme gesprochen werden kann.

Die umfangreichen Evaluationsvorhaben wurden anhand von erhältlichen Monitoringdaten und weiteren, innerhalb der einzelnen Programmevaluationen jeweils selbst erhobenen Daten durchgeführt, die allerdings des Öfteren doch keine analytisch konsequente Ausleuchtung aller Programmkomponenten zuließen. Die Gesamtebenen aller relevanten Programmoutputs und -outcomes, die schrittweise hin zur Erreichung der Programmziele führen sollen, und insbesondere die Verbindungen zwischen diesen Ebenen, wurden nur mit teils deutlichen Einschränkungen greifbar gemacht (Standard N4). Aufgrund dieser Ausschnitthaftigkeit haben die meisten der untersuchten Programmevaluationen letztlich doch den Charakter einer sogenannten „black box"-Evaluation, durch die die genaue Art und Weise, wie ein Programm die intendierten Wirkungen erzielt bzw. an der Erzielung dieser Wirkungen gehindert ist, nicht oder zumindest nicht vollständig erfasst wird. Es zeigen sich des Öfteren Schwierigkeiten mit einer konzisen Gliederung von Programmkomponenten und Umsetzungsschritten zu Zielen unterschiedlicher logisch-hierarchischer Stellung (unmittelbare, intermediäre und übergeordnete Programmziele) und hinsichtlich der Art der Erreichung von direkten und indirekten Zielgruppen. Zugleich haben einige Evaluationen auch Fragestellungen behandelt, die nicht als zentrale Gesichtspunkte für ein tieferes Verständnis des evaluierten Programms zu erachten sind. Wie die EvaluatorInnen angeben, waren Auswahl und Umfang der in den Programmevaluationen herangezogenen Informationen häufig nicht ausreichend, um alle mitgegebenen Evaluationsfragen gut behandeln zu können, und noch weniger, um auch unbeabsichtigte Wirkungen der Programm erfassen zu können. Einige Evaluationsberichte tragen Züge eines „evaluability assessment", in dem die Bedingungen für eine zielführende Evaluation des Programms erst geklärt werden.

Fast alle analysierten Programmevaluationen haben sich auch mit dem Kontext der evaluierten Programme auseinandergesetzt, in unterschiedlicher Intensität und mit unterschiedlichen Perspektivierungen. Vor allem auf der Basis von qualitativen Untersuchungsstrategien wurden von manchen Evaluationen essentielle Randbedingungen greifbar gemacht, unter denen das jeweilige Programm in seinen Zielgruppen Wirkungen erreichen konnte bzw. daran gehindert war. Etliche Kontextanalysen leiden jedoch darunter, dass zwar einige Faktoren untersucht und für skizzenhafte Bilder fruchtbar gemacht wurden, aber der systematische Stellenwert dieser untersuchten Faktoren

unklar bleibt bzw. keinen expliziten Bezug zu einer strukturierten und gesamthaft verstandenen Wirklogik des jeweiligen Programms aufweist (G2).

Die Evaluationsberichte geben trotz regelmäßig enthaltener Methodenbeschreibungen in der Mehrzahl doch nur unzureichend Auskunft darüber, was warum untersucht wurde, und als wie vollständig und tragfähig die erbrachten Ergebnisse eingeschätzt werden können. Im Verein mit nur sehr breiten und allgemein gehaltenen Angaben über die Untersuchungsschwerpunkte (Standard N2) und einer bemerkenswerten Enthaltsamkeit bei der Angabe von Evaluationsfragestellungen, die den jeweiligen Programmevaluationen zugrunde gelegt waren, ergibt sich so eine nur eingeschränkte Transparenz der Evaluationsergebnisse (Standard G3) und der Schlussfolgerungen, die aus ihnen gezogen wurden (Standard G8). Eine Transparenz der Vorgehensweise erscheint jedoch vor allem von Bedeutung, damit Evaluationsergebnisse auch von Akteuren aufgegriffen werden können, die nicht zum engen Kreis derjenigen Wenigen zählen, die unmittelbar mit der Konzeption und Umsetzung des untersuchten Programms und der dazu durchgeführten Evaluation befasst sind.

Ist anhand der Evaluationsberichte wegen ihrer Gestaltungsweise die Frage oft nicht gut beantwortbar, wie essentiell die erbrachten Ergebnisse im Hinblick auf die Gesamtlogiken der evaluierten Programme jeweils tatsächlich sind (N4, G3), so erscheint ebenso die Frage virulent, wie Programmen insgesamt Wert zugemessen wurde (N5). Hier offenbart sich ein „blinder Fleck“ eines stark datenorientierten und zugleich oftmals eher unsystematischen Zugangs. Während manche Evaluationen nachvollziehbare Bewertungsmaßstäbe in konsequenter Weise in Anschlag gebracht haben, die in einer klaren Verbindung zu den Programmzielen standen, haben andere eher für sich stehende Einzelbewertungen zu einzelnen Beobachtungen vorgenommen, ohne dass in der Kombination von „üblichen“ Betrachtungsweisen ein stringentes Gesamtkonzept greifbar würde. Es wird in der internationalen Evaluationstheorie allerdings davon ausgegangen, dass die Wahl der Bewertungsmaßstäbe ebenso eine tragende Säule jedes Evaluationskonzepts darstellt wie ihre Wissenschaftlichkeit und ihre gezielte Auseinandersetzung mit dem intendierten Nutzen.

Im Zusammenhang mit eingeschränkten Datenlagen waren die EvaluatorInnen immer wieder bestrebt, Lücken durch ihr Hintergrundwissen über das FTI-System und Annahmen über dessen Funktionsweisen oder Eigenschaften von Akteursgruppen wett zu machen (F3, G8). Dies beeinflusste oft merklich den Charakter von Schlussfolgerungen und Empfehlungen, die in unterschiedlicher Weise, aber doch teils recht deutlich, einen Zug von ExpertInnengutachten tragen, in denen das persönliche Wissen der AutorInnen zur Geltung gebracht wird. Dies deckt sich nicht mit dem Grundansatz der Evaluationsstandards, dass alle Aussagen einer Programmevaluation in transparenter Weise in von ihr herangezogenen Fakten und Quellen abgestützt sein sollten.

Festzustellen ist schließlich, dass in den Evaluationsprozessen zahlreiche Schritte, die den Standards zufolge vor allem im Planungsstadium einer Evaluation erfolgen können bzw. sollten, bislang nur ansatzweise wahrgenommen wurden. Hier zeigen sich unter Rückgriff auf Ergebnisse der EvaluatorInnen-Befragung unter anderem deutliche Verbesserungsmöglichkeiten bei der gezielten Auseinandersetzung damit, wie das Evaluationsprojekt auf eine konkret intendierte Nutzung zugeht (N8), und wie es entsprechend LernpartnerInnen einbindet (N1). Innerhalb von kurzen Vorbereitungsphasen der Evaluationen (Beantwortung von Terms of Reference und Hearing) kam es nur eingeschränkt zu einer Mitsprache der EvaluatorInnen, im Rahmen derer sie auf Basis ihrer Kompetenzen die Herangehensweise der Evaluation beeinflussen und schärfen konnten (N4). Über weite Strecken wurden Methoden zum Einsatz gebracht, die von den AuftraggeberInnen erwünscht waren oder von den EvaluatorInnen bzw. ihren Instituten regelmäßig eingesetzt werden. Lediglich in zwei der untersuchten Berichte wurden ungewöhnliche und innovative Methoden eingesetzt, die für die spezifische Aufgabenstellung der betreffenden Programmevaluation als produktiv erachtet wurden. Bewertungsmasstäbe zur Einschätzung der Programme und Kriterien zur Einordnung von Beobachtungen wurden nur selten zwischen EvaluatorInnen und AuftraggeberInnen vorab gemeinsam geklärt (N5). Die Wahl von Bewertungsmasstäben wurde oft den EvaluatorInnen überantwortet, und diese zogen entweder Massstäbe heran, die in ihren Augen denen der AuftraggeberInnen entsprachen, oder verhielten sich unabhängig von solchen Annahmen.

2.4 Aktuelle Herausforderungen in der FTI-politischen Arena

Früher gehegte Erwartungen an die Leistungskraft von Programmevaluationen wurden als unrealistisch erkannt. Die verfügbare Ressourcenausstattung von Programmevaluationen wird als wesentlicher Mitgrund dafür erachtet, dass immer wieder Informationsbedürfnisse nur eingeschränkt befriedigt werden konnten. Die Verankerung der Programmevaluationen als Bestandteile der Programmvereinbarungen erzeugt eine Spannung zwischen vorgegebenen Evaluationsfragestellungen und aktuellen Informationsbedürfnissen in einem hochdynamischen System, die wiederholt auf Kosten aktuell relevanter Erkenntnisse gegangen ist. Die Evaluationsfunktion ist im Governancesystem an feststehende Evaluationszeitpunkte und –budgets gebunden, die gemäß den Auskünften der AuftraggeberInnen zwar bisweilen mit einer gewissen Flexibilität gehandhabt werden können, im Großen und Ganzen aber jedenfalls enge Grenzen setzen. Evaluationsprojekte oder Studien evaluativen Charakters, die nicht in Programmdokumenten vorprogrammiert waren, wurden nur in seltenen Ausnahmefällen initiiert.

In allen Ressorts und Agenturen wurden Zuständigkeiten und Kapazitäten geschaffen, um Evaluationen durchführen und Evaluationsergebnisse auf einer strategischen Ebene handhaben zu können. Die Planung und Durchführung der Programmevaluationen, die primär in den Ressorts erfolgt, ist dort an die Fachzuständigkeiten für die evaluierten Programme gekoppelt. Abstimmungsprozesse intern und innerhalb der Principal-Agent-Beziehungen sind erforderlich, die im allgemeinen auf Grund vorhandener Kooperationsbereitschaft erfolgreich verlaufen, aber doch keiner institutionell klar verankerten Systematik folgen. Das Engagement, das für eine einzelne Programmevaluation aufgebracht wird, bemisst sich nicht zuletzt an den zum jeweiligen Zeitpunkt gegebenen Möglichkeiten der fachzuständigen Einzelpersonen im Rahmen auch anderer Agenden. Durchgehend wird dargestellt, dass im Rahmen der gegebenen Kapazitäten keine weiteren Spielräume mehr bestehen.

Die Weitergabe von Evaluationsergebnissen innerhalb der Hierarchien stellt sich als geregelter Vorgang dar. Dabei wird davon ausgegangen, dass Evaluationsergebnisse nur eine Informationsquelle unter vielen sind, auf die sich politische EntscheidungsträgerInnen stützen, und dass auch die politische Aufmerksamkeit für unterschiedliche Programme deutlich variiert. Eine Zirkulation von Evaluationsergebnissen hin zu anderen Fachabteilungen, die zur Stärkung der Wissensbasis in systemischer Hinsicht beiträgt, bemisst sich stark am Engagement von Einzelpersonen. In jüngster Zeit sind verstärkte Bemühungen zu beobachten, durch übergreifende hausinterne Präsentationen Evaluationsergebnisse in Umlauf zu setzen und Diskussionen zu initiieren, in denen auch nicht direkt mit dem evaluierten Programm befasste Abteilungen von den Evaluationsergebissen profitieren können und strategische Einschätzungen vorgenommen werden können. Eine institutionelle Verankerung derartiger wertvoller Vorgänge ist allerdings nicht gegeben, und eine durchgehende Systematik liegt nicht vor.

Im Rahmen der institutionellen Architektur bestehen einige wenige Berührungspunkte zwischen den Steuerungssegmenten im FTI-politischen Bereich, in denen zumindest potenziell Informationen über geplante und fertiggestellte Evaluationen ausgetauscht werden können. In erster Linie sind es jedoch Personen und Netzwerke, die einen übergreifenden Wissensfluss im Governance-System gewährleisten, sodass sich ein solcher Wissensfluss letztlich als akzidentiell darstellt.

Es besteht allseitiger Bedarf an verstärkt systemisch orientierten Erkenntnissen, durch die die Positionierung einer Maßnahme im breiten Kontext verschiedener Förderungs- und Steuerungsinstrumente ebenso aufgezeigt werden kann wie Optionen, in welcher Weise Bedarfslagen im systemischen Gesamtzusammenhang durch Einsatz und Konfiguration bestimmter Instrumente und Maßnahmen gezielt und in bestmöglicher Weise begegnet werden kann. In einzelnen Systemsegmenten besteht zudem Bedarf an Typen von Politikinformation, die mit den routinisierten Multi-Purpose-Evaluationen nicht gut abgedeckt werden können. Es geht hier (1) um intensivere Auseinandersetzungen mit Zielgruppen und Wirkungsweisen von Maßnahmen auf einer detaillierten Ebene, die in Richtung einer wissenschaftlichen Begleitforschung weisen, (2) um ein möglichst frühzeitiges Erkennen der Realitätshaltigkeit von Annahmen über die Wirkungsweise von programmen, und (3) um hoch reaktive und schlanke Studien evaluativen Charakters, die das FTI-politische Handeln in dynamischen Umwelten zeitnah unterstützen.

3. Empfehlungen

Es wird auf Basis der drei Datenquellen und deren integrierender Analyse ersichtlich, dass es sich bei der Frage der Evaluationsqualität in Bezug auf Nützlichkeit und tatsächlich zustande kommenden Evaluationsnutzen nicht um Einzelursachen handelt, sondern um Syndrome und Faktorenbündel von erheblicher Komplexität. Die bisherige Evaluationspraxis im FTI-Bereich erweist sich als gleichermaßen durch Gestaltungsmerkmale einzelner Programmevaluationen wie durch Kontextfaktoren bedingt. Damit ist auch nicht die *eine* Lösung greifbar, die eine entscheidende Weiterentwicklung über den bisher erreichten Stand hinaus bewirken könnte.

Limitierungen für die Gestaltung von Programmevaluationen und die Entstehung von Evaluationsnutzen ergeben sich aus Merkmalen des institutionellen Arrangements. Evaluationsberichte und die hinter diesen Produkten stehenden Evaluationsprozesse gehen auf das, was die Evaluationsstandards als optimale Schritte hin zu hoher Nützlichkeit bezeichnen, bislang nur bedingt zu. Damit lassen auch Evaluationsprodukte und –prozesse Nutzen entstehen, die die von den Standards empfohlenen bzw. als notwendig erachteten Evaluationseigenschaften nicht optimal verwirklichen. Unter dem Gesichtspunkt einer größtmöglichen Nützlichkeit auf der Basis hervorragender Evaluationsqualität muss es zweifellos angelegen sein, von einer strukturell kompromisshaften Situation zu verbesserten Bedingungen für die Planung, Durchführung, Kommunikation und Nutzung von Programmevaluationen zu gelangen. Die Evaluationsstandards sind als praktische Anleitung zur Bewältigung von Problemen bei der Nutzenentstehung konzipiert, doch können sie Probleme nicht lösen, die außerhalb der Reichweite eines konkreten Evaluationsprojekts liegen, und gute Lösungen entlang der Standards müssen von EvaluationsauftraggeberInnen in der Gestaltung der Evaluationsaufträge auch ermöglicht werden. Strukturell ermöglichte Potenziale für die Planung, Durchführung, Kommunikation und Nutzung von Programmevaluationen bleiben sodann in den jeweiligen Projekten auf der Basis von Kapazitäten und Kompetenzen auszufüllen.

Die Metaevaluation gelangt daher zu Empfehlungen, die sich sowohl auf einer evaluations-theoretischen Ebene als auch auf der Ebene der institutionellen Einbettung der Evaluationsfunktion ansiedeln. Die vorgelegten Empfehlungen sind an der Weiterentwicklung einer in sich dynamischen und systemevolutiven Evaluationspraxis orientiert. Da die bisherige Evaluationspraxis in nachvollziehbarer Weise bereits Nutzen erzeugt hat, setzen die Empfehlungen nicht auf eine radikal-disruptive Veränderung, die aus einer Orientierung an Governancemodellen anderer Länder grundsätzlich abgeleitet werden könnte, aber hinsichtlich tatsächlicher Transferierbarkeit und Eintreten der erhofften Effekte doch auch mit einigen Ungewissheiten einhergeht. Für die evaluationsmethodische Ebene würde eine Benennung aller denkbaren Verbesserungsoptionen freilich darauf hinauslaufen, den gesamten Gehalt der Standards zu referieren. Diesbezügliche Empfehlungen werden nur für diejenigen Gesichtspunkte ausgesprochen, die als die wesentlichsten erscheinen. Letztlich beruht eine hoch entwickelte Evaluationskultur auch auf gesellschaftlich-kulturellen Faktoren wie der Offenheit für sachlich fundierte Kritik und der Bereitschaft zur offenen Diskussion, die sich freilich einer gezielten Beeinflussung entziehen.

Die folgenden 20 Empfehlungen werden ausgesprochen:

1. Programmevaluationen sollten in Zukunft weiterhin durchgeführt werden, da sie in der Vergangenheit wertvolle Beiträge zur zielgerechten Umgestaltung und Neukonzeption FTI-politischer Maßnahmen erbracht haben, die noch über die Ebene der jeweils evaluierten Programme hinaus reichen. Um die Produktivität der Programmevaluationen über das bisherige Maß hinaus weiter steigern zu können, sollten sie mit den folgenden Empfehlungen benannten Schritten einhergehen.
2. Die derzeit gegebene Verankerung der Evaluationsfunktion bei den Institutionen, die für die Konzeption und Umsetzung von FTI-Programmen zuständig sind, sollte beibehalten werden. Entscheidende Kapazitäten für die Planung, Durchführung und Verwertung von Programmevaluationen wurden hier über Jahre hinweg aufgebaut. Die Verankerung bei den Programmverantwortlichen sorgt auch für ein Commitment zu den Programmevaluationen, das für in der Vergangenheit entstandenen Evaluationsnutzen wesentlich war. Eine Weiterentwicklung der Evaluationskultur im FTI-Bereich sollte als pfadabhängige Entwicklung auf dieser wertvollen Grundlage gedacht werden.

3. Programmevaluationen sollten in Zukunft mit denjenigen Ressourcen ausgestattet werden, die eine konzeptgemäße Analyse des evaluierten Programms unter Heranziehung aller für die Evaluationsschwerpunkte und –fragestellungen benötigten Informationsquellen tatsächlich ermöglichen und eine gute Durchführung gemäß dem Qualitätsverständnis der internationalen Standards für Programmevaluation gewährleisten.

4. Programmevaluationen sollten künftig stärker auf eingegrenzte Evaluationsschwerpunkte fokussiert werden. Dadurch können unter Bedingungen begrenzter Ressourcen intensivere und genauere Untersuchungen zu den gewählten Schwerpunkten durchgeführt werden. Jeweils nicht gewählte Evaluationsschwerpunkte können gegebenenfalls durch eine weitere Evaluation verfolgt werden. Dabei können dann auch andere Evaluationsteams zum Einsatz kommen, was zu einer Anreicherung der Sichtweisen auf das untersuchte Programm auf Basis unterschiedlicher Kompetenzen beitragen kann.

5. Programmevaluationen sollten verstärkt in ihrer Prozessqualität begriffen und auf dieser Ebene in Planung und Durchführung gestärkt werden. Die DeGEval-Standards mit ihrem Interpretationshintergrund der Joint Committee-Standards weisen auf Schritte hin, durch die im Planungs- und Durchführungsstadium von Programmevaluationen Qualität in unterschiedlichen Hinsichten gestärkt und sicherstellt werden kann. Die *Plattform fteval* sollte sich mit solchen Möglichkeiten auseinandersetzen, da sie Voraussetzungscharakter für die Erzielung späterer Evaluationsergebnisse und deren Nutzungspotenzial für verschiedene Akteursgruppen haben.

6. Evaluationsberichte sollten in jeder Hinsicht klar und in einer auch für Außenstehende gut verständlichen Weise abgefasst werden. Dies ist insbesondere als Voraussetzung dafür zu verstehen, dass es zu einer verstärkten Nutzung von Programmevaluationen in anderen FTI-politischen Bereichen und nach dem Denkprinzip einer vermehrten systemreferentiellen Selbststeuerung der FTI-Akteure kommen kann.

7. Alle Evaluationsberichte sollten systematisch ein Kapitel beinhalten, in dem die Gesamtvorgehensweise der Evaluation in methodischer wie organisatorischer Hinsicht konzis und vollständig dargestellt wird und auch auf Vor- und Nachteile der tatsächlich durchgeführten Analyse hingewiesen wird. Eine derartige kompakte Übersicht über die Gesamtvorgehensweise erscheint insbesondere hinsichtlich einer stärkeren Nutzung von Evaluationsergebnissen in einem gesamt-systemischen Zusammenhang relevant, damit auch Akteure, die mit den unmittelbaren AuftraggeberInnen nicht identisch sind, auf die erbrachten Evaluationsergebnisse gut zugreifen können. In der Darstellung der Vorgehensweisen sollte es auch Mut zum Ausweis von Lücken geben, da keine Programmevaluation alles beleuchten kann, was theoretisch zu einem Programm untersucht werden könnte. Auch ein abholbares Wissen darüber, was noch nicht intensiv untersucht werden konnte, sollte als produktiver Beitrag zum FTI-politischen Wissens- und Informationssystem betrachtet werden, damit dieses im Weiteren produktiv ausgestaltet werden kann.

8. Es sollte eine verstärkte Auseinandersetzung mit der gezielten Anwendung von Bewertungsmaßstäben auf die evaluierten Programme angestrebt werden. Dabei geht es nicht nur darum, wie Zielerreichungen gemessen und eingeschätzt werden, was oft zum Evaluationszeitpunkt in dieser Form noch gar nicht möglich ist, sondern auch und gerade um die wohlbegründete Einordnung der Beobachtungen zu Aspekten der Programmentfaltung. Die von einer Evaluationsstudie angewendeten Bewertungsmaßstäbe sollten als vitale Konzeptfrage begriffen und im Planungsstadium als integraler Bestandteil des übergreifenden Evaluationskonzepts vereinbart und festgelegt werden. Konsistente Bewertungsmaßstäbe verkörpern sich unter anderem in der Verfolgung von Kohärenz und Konsistenz von Programmzielen und Programmkomponenten in ihrer Umsetzung, in einer Festlegung, wie die Sichtweisen verschiedener Akteursgruppen auf das evaluierte Programm zur Gesamteinschätzung führen, in vorab festgelegen Kriterien zur Einordnung späterer Messergebnisse, in oder in der gezielten Bestimmung von Messgrößen (etwa bei einem Programmziel „Kooperation“ die Quantität von Kooperationsbeziehungen versus qualitative Eigenschaften von eingegangenen Kooperationen).

9. Dem Risiko eines Lock-Ins in üblichen Herangehensweisen an Evaluation, die mit Ermüdungserscheinungen der Evaluationspraxis in Zusammenhang stehen, sollte durch eine systematische professionelle Beratung von Evaluationsplanungen und –prozessen gegengesteuert werden. Eine solche Beratung wird vor allem dann ein probates Mittel darstellen, wenn sie nicht nur FTI-spezifische Kompetenzen heranzieht, sondern auch evaluationsmethodische Kompetenzen, die den Konnex zu Entwicklungen und Know-How anderer Bereiche herstellen.

10. Eine Intensivierung der Planungsphasen der Programmevaluationen sollte angestrebt werden, um das Risiko zu minimieren, dass beschränkte Ressourcen in letztlich ergebnisarme Untersuchungsschritte fließen. Dafür bietet sich das international anzutreffende Modell einer sogenannten *„Inception Phase"* am Beginn einer Programmevaluation an, in der sich die beauftragten EvaluatorInnen intensiv mit der Datenlage, methodischen Möglichkeiten im Rahmen der gegebenen Ressourcen, und der Beantwortbarkeit der vorgesehenen Evaluationsfragen auseinandersetzen. Diese genaue Abwägung bildet sodann die Grundlage für ein bestmögliches Evaluationsdesign, das der im Anschluss durchgeführten Evaluation zugrunde gelegt wird. Das Modell zielt darauf ab, so realistische Erwartungen wie möglich an eine Evaluation zu entwickeln und die für die Evaluation verfügbaren Ressourcen so gut wie möglich zu nutzen. Vergaberechtliche Voraussetzungen für die Nutzbarkeit dieses Modells bleiben zu prüfen.

11. Die institutionelle Verankerung der Evaluationsfunktion in den Ressorts und Agenturen sollte weiter gestärkt werden. In den auftraggebenden Ressorts und Agenturen existieren HauptansprechpartnerInnen für Evaluationsangelegenheiten und VertreterInnen der Institutionen in der *Plattform fteval*, doch ist bis heute keine dieser Personen ausschließlich mit Evaluationsangelegenheiten betraut, um sich dieser komplexen und anforderungsreichen Materie vollständig widmen zu können. Ressourcen von fachzuständigen MitarbeiterInnen für die Auseinandersetzung mit anderweitig erarbeiteten Programmevaluationen sind kaum vorhanden. Eine spezialisierte, hoch professionelle Evaluationsabteilung oder Stabstelle, die sich mit der Planung der Programmevaluationen, dem Evaluationsmanagement, einer Qualitätskontrolle und der Verwertung und Weitergabe der Evaluationsergebnisse für das ganze Haus befasst, stellt in diesem Zusammenhang das Idealbild dar, das einen entscheidenden Schritt zur Überwindung der Variabilität im Umgang mit einzelnen Evaluationen verkörpern würde.

12. Die Lernfunktion der Programmevaluationen sollte künftig durch eine Flexibilisierung der Auslösung und Intensität der einzelnen Evaluationen weiter gestärkt werden. Frei allozierbare Evaluationsbudgets könnten die Gestaltung von Programmevaluationen im Aktualitätsbezug sowie unter Gewichtung von Informationsbedarfslagen ermöglichen. Nicht alle Programme brauchen in gleicher Weise evaluiert zu werden, um in einem übergreifenden FTI-politischen Informationssystem wesentliche Erkenntnisse zu erzielen. Eine Flexibilisierung würde somit zu zielgerechten Investitionen in anspruchsvollere Evaluationen und Studien und zu einer effektiveren Nutzung der im System vorhandenen Ressourcen beitragen. Programmwirkungen könnten zu passenderen Zeitpunkten analysiert werden, als es bislang der Fall war. Thematische Evaluationen, etwa zu Programmfamilien oder Zielgruppen, und Instrumentenevaluationen könnten verstärkt durchgeführt werden.

13. Programmevaluationen sollten verstärkt als übergreifende und konzise Analysekonzepte verstanden und angelegt werden. Evaluationsmethodische Konzepte und Tools, die für eine möglichst zielführende Evaluation von Programmen über die letzten beiden Jahrzehnte international entwickelt wurden, sollten dabei herangezogen werden. Zu empfehlen ist eine Zuwendung zu Ansätzen, die unter dem Sammelbegriff der Theorie-basierten Evaluation (*theory-based evaluation*) bekannt sind. Diese Ansätze sind gezielt dafür konzipiert, die Einlösbarkeit von Programmannahmen in der realen Programmentfaltung zu beleuchten und geschärfte Umgangsweisen mit der Kausalitätsproblematik zu ermöglichen, wie und inwieweit ein Programm zu intendierten Veränderungen beiträgt. Mit der Zuwendung zu ihnen würde die Evaluationspraxis im FTI-Bereich Analysestrategien zur Anwendung bringen, die in anderen Politikbereichen auf internationaler Ebene und in internationalen Organisationen bereits eingesetzt werden. Die avancierten Ansätze der Realistischen Evaluation (*realistic evaluation*) und der *Contribution Analysis* könnten aufgegriffen werden, um zu einem vertieften Verständnis der Wirkungsweise von Programmen in ihrer Kontextabhängigkeit zu gelangen und komplexe Programme, Programmfamilien, Portfolien und Maßnahmenbündel zielführend und in pragmatischer Weise zu analysieren. Allseitige Ressourcen für die Arbeit mit qualitativen Daten und notwendige Interaktionen zwischen EvaluatorInnen und AuftraggeberInnen während der Evaluationsdurchführung sind freilich vorausgesetzt..

14. Programmdokumente sollten so eingehend wie möglich darlegen, wie Zielsetzungen systematisch gegliedert sind, welche Outputs die verschiedenen Programmaktivitäten erzeugen sollen, und welche Annahmen darüber gemacht werden, wie diese Outputs zu Outcomes und weiteren Entwicklungen hin zu Zielerreichungen führen. Eine möglichst gute Darstellung der intendierten Wirkungsweise der Programme durch die Programmeigentümer bei der Programmkonzeption bildet den Gegenpol zur evaluatorischen Aufarbeitung einer Programmlogik und deren Ausgestaltung in der Programmwirklichkeit. Die Konzeptualisierung der intendierten

Wirkungsweise der Programme kann im Planungsstadium durch ex ante-Evaluationen unterstützt werden. Freilich können ex ante-Evaluationen spätere Überprüfungen nicht ersetzen, wie sich Programmeffekte im realen Operieren des Programms herstellen oder mit Hindernissen konfrontiert sind.

15. Programmevaluationen sollten gemeinsam mit allen verwandten und ergänzenden Bestandteilen eines übergreifenden FTI-politischen Wissens- und Informationssystems durch Publikation verfügbar gemacht werden, um auch Synergien zwischen Studien unterschiedlichen Typs allgemein nutzbar zu machen. Die konkrete Bezeichnung von Programmevaluationen, Reviews, Assessments oder wissenschaftlichen Studien evaluatorischen Charakters sollte nicht zum Anlass werden, wertvolle Informationspotenziale zu beschneiden. Ein Repositorium für alle evaluativen und wissenschaftlichen Studien kann im Bedarfsfall in allgemein zugängliche Bereiche und Bereiche mit Zugangsbeschränkungen gegliedert werden. Nicht-Publikation ist gerechtfertigt und angebracht, wenn in einer systematischen Qualitätskontrolle zum Schluss gekommen wird, dass durch die Publikation unzuverlässige oder irreführende Information zur Nutzung freigegeben würde. Die Nutzbarkeit jedweder evaluativer Information wird von einer adäquaten Dokumentation über den genauen Charakter dieser Information abhängig bleiben. Im Verständnis der Evaluationsstandards ist jeder Nutzung von Evaluationsergebnissen eine umfassende Auseinandersetzung mit der genauer Vorgehensweise und Durchführungsqualität der betreffenden Evaluation vorausgesetzt. Eine bloße Verfügbarkeit von Datenbeständen, die unter nicht genau verstehbaren Ausgangsbedingungen in Bezug auf nicht genau bekannte Informationsbedürfnisse erarbeitet wurden, sollte nicht als ausreichend erachtet werden.

16. Jede Programmevaluation sollte bei ihrer Publikation von einer „Management Response“ begleitet werden, die die Kenntnisnahme der Evaluationsergebnisse auf Ebene des Top Managements bestätigt, eine Positionierung zu diesen Ergebnissen angibt, und damit auch Verbindlichkeit erzeugt. Dabei geht es nicht etwa um eine automatische Übernahme von Evaluationsergebnissen, sondern im Gegenteil um das Produkt einer aktiven Auseinandersetzung mit ihnen. Dieser Weg wird beispielsweise von der Deutschen Forschungsgemeinschaft (DFG) bereits beschritten und wurde neuerdings auch von einer Agentur im österreichischen FTI-Governancesystem eingeschlagen.

17. Der RFTE sollte die ihm zur Verfügung stehenden Mittel nützen, um in der evaluativen Wissensproduktion offen bleibenden Informationsbedarf durch gezielte Vergabe von Studien in aktualitätsbezogener und flexibler Weise zu befriedigen. Dies erscheint im Hinblick auf intensive Analysen zu Themen und Segmenten des FTI-Systems ebenso relevant wie im Hinblick auf übergreifende, systemisch ausgerichtete Analysen. Ein Charakter wissenschaftlicher Begleitforschung, die in den Ressorts und Agenturen keinen Ort hat, könnte dabei zum Tragen kommen. Im Hinblick auf den systemischen Stellenwert solcher Studien erscheint eine Abstimmung mit den relevanten FTI-politischen Akteuren sinnvoll und wichtig.

18. Eine Koordinationsfunktion für FTI-Evaluationen sollte geschaffen werden, die sich mit möglichen Synergiebildungen zwischen an verschiedenen Systemstellen angesiedelten Evaluationsaufgaben und -ressourcen befasst, um durch Abstimmungen und Beratungen die gegenwärtige Zersplitterung der Evaluationsaktivitäten und Kleinteiligkeit im Analytischen zu überwinden. Dadurch kann ein Potenzial ausgeschöpft werden, das aus einer Bündelung von Ressourcen und Erkenntnisinteressen resultiert. Ressort- und Agentur-übergreifende Abstimmungsleistungen könnten erbracht werden, deren Machbarkeit unter den gegebenen Bedingungen eingeschränkt ist. Erträge hinsichtlich stärker systemisch ausgerichteter Fragestellungen zum Stellenwert von einzelnen Maßnahmen und Steuerungen sind zu erwarten.

    Dies kann zugleich als sinnvolle Alternative zu ebenso seltenen wie schwer initiierbaren Großunternehmungen wie der Systemevaluation 2009 erachtet werden, indem systemische Fragestellungen zum Gegenstand eines rollenden Verfahrens werden. Eine solche Koordinationsfunktion ist jedenfalls mit hohen fachlichen Kompetenzen und adäquaten Ressourcen auszustatten. Es bleibt zu prüfen, ob eine Einrichtung möglich ist, ohne bestehende Rechtsbestände anzutasten. Die Konfiguration und Einrichtung sollte durch eine Studie vorbereitet werden, die sich mit internationalen Beispielen auch außerhalb des FTI-politischen Bereichs befasst.

19. Ein Diskussionsforum sollte geschaffen werden, das Evaluationsergebnisse an ein breiteres Fachpublikum heranträgt, das über den engen Kreis der in der Plattform fteval versammelten Akteure hinausreicht und ProgrammangerInnen und Programmverantwortliche an unterschiedlichen Systemstellen genauso anspricht wie Akteursgruppen im FTI-System. Hierdurch

können Wissensflüsse in Gang gesetzt und Diskussionen ausgelöst und angereichert werden, die für ein System systemreferentieller und selbstreflexiver Akteure relevant sind. Der derzeitigen starken Abhängigkeit von Wissensflüssen im FTI-politischen Governancesystem von Personen und Netzwerken würde damit gegengesteuert. Ebenso würde der Umstand, dass auf der Basis einer bloßen Publikation evaluative Information eine Holschuld für etwaige InteressentInnen bleibt, behoben. Ein solches Diskussionsforum kann optional mit der vorgenannten Koordinationsfunktion verbunden werden, aber auch eine getrennt angesiedelte Systemfunktion darstellen.

20. Hinsichtlich einer substantiellen Stärkung der Lernfunktion der Programmevaluationen ist die derzeitige Kombination der unterschiedlichen Evaluationszwecke des Lernens und der Rechenschaftslegung, die für die Programmevaluationen durch deren institutionell-rechtliche Verankerung als Schnittstellenfunktion zwischen Fachressorts und Finanzressort stets gegeben ist, nicht als produktiv zu erachten. Nachdem mit der Wirkungsorientierten Folgenabschätzung (WFA) eine andersartige Evaluationsfunktion im Bezug auf Rechenschaftslegung geschaffen wurde, könnte überlegt werden, inwiefern die Lernfunktion der Programmevaluationen von Zwecken der Rechenschaftslegung künftig getrennt werden kann. Zwecke der Programmdokumentation könnten verstärkt in die Hände der Agenturen gelegt werden, die bereits jetzt wesentliche Teile der Datenbasen erarbeiten, die in Programmevaluationen verwendet werden. Im Gegenzug könnten Evaluationen dann verstärkt Analyseschritte setzen, die nicht der Gefahr eines Lock-Ins in vorab festgelegten Datenstrukturen ausgesetzt sind.

1. Motivation und Anlage der Studie

1.1 Motivation und Auftrag

In Österreich wurde über Jahre hinweg eine Evaluationskultur im forschungs-, technologie- und innovationspolitischen Bereich aufgebaut, wobei die Gründung der *Plattform fteval* als spezifisches Forum für Kommunikation und Kapazitätsaufbau für FTI-Evaluation 1997 ein entscheidendes Datum darstellt. Die Anzahl an Programmevaluationen, die seitdem durchgeführt wurden, ist im internationalen Vergleich für den Politikbereich jedenfalls als überdurchschnittlich zu erachten. Freilich sind mit der zunehmenden Entfaltung des Unternehmens FTI-Evaluation auch immer wieder Stimmen in der Akteursarena laut geworden, die Enttäuschungen gegenüber ursprünglich hohen Erwartungen thematisieren. Zugleich ist in den letzten Jahren eine internationale Entwicklung zu beobachten, die Frage nach der Wirksamkeit und dem Reflexionspotenzial von Evaluationen im FTI-Bereich aufzuwerfen (Edler et al 2008, Elg & Hakanson 2012, EPEC 2011, Hyvärinen 2011, MIOIR et. Al. 2010, MIOIR 2013, Barjak 2013). Parallel dazu haben über den FTI-Bereich hinaus Evaluationsgesellschaften die Frage der Nützlichkeit von Evaluation und des Wissens über die Evaluationspraxis auf die Tagesordnung gesetzt.

Der RFTE hat deswegen die vorliegende Studie zur Nützlichkeit und zum tatsächlich feststellbaren Nutzen der Programmevaluationen im österreichischen FTI-Bereich beauftragt, mit der der unabhängige externe Evaluator im Juli 2014 betraut wurde. Die Studie sollte unter Bezugnahme auf ein allgemein anerkanntes internationales Wissenskorpus über qualitätsvolle Evaluation und spezifische Aspekte der Evaluationsnutzung eine übergreifende Analyse der mehrjährigen Evaluationspraxis bieten. Sie sollte in einer an konkreten und eingegrenzten Fragestellungen ausgerichteten Weise zu einer Erhellung von Nützlichkeitsmerkmalen und Nutzungsbedingungen von FTI-Evaluationen beitragen, indem sie eine systematisch-gesamthafte Sicht auf Anlage, Durchführung und Verwertung der verschiedenen Evaluationsprojekte schafft. Die konkrete Herangehensweise wurde als diejenige einer Metaevaluation definiert, die um Aspekte der Nutzungsforschung zur Evaluation angereichert ist.

Der Evaluationszweck der vorliegenden Studie besteht darin, evidenzbasierte Einschätzungen zu erbringen, ob und inwiefern Verbesserungspotenzial für FTI-Evaluationen in Österreich unter den spezifischen Bezugspunkten der Nützlichkeit und Nutzung besteht, das von Akteuren des Feldes in der Beauftragung und Durchführung von Evaluationen in Zukunft aufgegriffen werden kann. Als Hauptnutzer der Studie war der auftraggebende Rat für Forschung und Technologieentwicklung (RFTE) vorgesehen, dem mit dem vorliegenden Bericht die Ergebnisse und Empfehlungen zur Formulierung von eigenen Empfehlungen des RFTE an die politisch-administrative Akteurslandschaft zur Verfügung stehen. Mit der Publikation des Berichts werden die Ergebnisse auch für die zentralen Akteure der Evaluationspraxis und weitere interessierte Öffentlichkeiten verfügbar.

Evaluationsgegenstand ist die österreichische Evaluationspraxis zu Programmevaluationen im FTI-Bereich auf Bundesebene, die über den Zeitraum der letzten 12 Jahre hinweg betrachtet wird. Dadurch wird die betreffende Evaluationspraxis seit den Strukturreformen im FTI-Bereich am Beginn der 2000er-Jahre abgebildet. Dabei können Programmevaluationen nicht nur als Einzelereignisse bzw. –produkte betrachtet werden, sondern auch Weiterentwicklungen der Evaluationspraxis nachvollzogen und kumulative Wirkungen einer multiplen Evaluationstätigkeit erfasst werden.

1.2 Konzeptuelle Grundlagen

1.2.1 Metaevaluation

Metaevaluationen stellen systematische Analysen von einer oder mehreren Evaluationen dar, die der Bestimmung von Wert und Güte dieser Evaluationen gelten. Dabei geht es nicht um die inhaltlichen Ergebnisse, die eine Evaluation erbracht hat, sondern um die Gestaltungsweise der Evaluationsberichte (unter Umständen auch mehrerer Teilberichte) und der hinter diesen Produkten stehenden Evaluationsprozesse. Metaevaluation unterscheidet sich von der anders gelagerten Herangehensweise einer Evaluationssynthese, die eine Integration der in verschiedenen Evaluationen erbrachten Ergebnisse intendiert (vgl. z.B. Widmer 2001, Widmer & De Rocchi 2012:160-162*)*. Mit der vorliegenden Studie wird somit nicht der Frage nachgegangen, welche inhaltlichen Diagnosen und Empfehlungen durch österreichische Programmevaluationen im FTI-Bereich erbracht wurden,

sondern wie die Gestaltung der Evaluationsberichte und der dahinter stehenden Evaluationsprozesse unter dem spezifischen Gesichtspunkt ihrer Nützlichkeit einzuschätzen ist.

Grundlegend für die Durchführung einer Metaevaluation ist die Überlegung, dass Evaluationen ausreichende Qualität aufweisen sollten, um eine gute Grundlage für ihre Nutzung zu bilden, und dass Evaluationsqualität gezielt gefördert werden kann, wenn Stärken und Schwächen von einzelnen Evaluationsprojekten oder von Gruppen von Evaluationen erkannt und gezielt reflektiert werden. „Metaevaluation is the process of delineating, obtaining, and applying descriptive information and judgmental information about an evaluation's utility, feasibility, propriety, and accuracy and its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility to guide the evaluation and publicly report its strengths and weaknesses." (Stufflebeam 2001: 183) Metaevaluationen können Evaluationsprojekte im Planung- und Durchführungsstadium unterstützten (formative metaevaluation), oder sie können ex post Qualitätsdimensionen durchgeführter Evaluationen reflektieren (summative Metaevaluation). Eine solche systematische summative Reflexion kann sich einer Reihe von Evaluationen auch über eine längere Zeitspanne hinweg widmen, um die spezifischen Merkmale einer bestimmten Evaluationspraxis zu erkennen und damit Grundlagen für eine Weiterentwicklung der Evaluationskultur bzw. -kapazität zu schaffen (Lipsey et al. 1985, Bickman 1997, Cooksy/Caracelli 2005). Die Metaevaluation kann dabei von den AuftraggeberInnen der untersuchten Evaluationen ebenso in Auftrag gegeben werden wie durch die EvaluatorInnen oder durch Dritte, was ihre Unabhängigkeit stärkt (Stufflebeam 2001). Die vorliegende Studie unternimmt eine summative Metaevaluation im Auftrag eines von Auftraggebern und EvaluatorInnen unabhängigen Gremiums, das mit der Materie der Evaluationsgegenstände der analysierten Evaluationen befasst ist, um essentielle Züge der Evaluationspraxis übergreifend zu identifizieren und auch eine etwaige Entwicklung der grundsätzlich als dynamisch verstandenen Evaluationspraxis im Zeitablauf nachzeichnen zu können.

Wie im vorangegangenen Zitat zur allgemeinen Charakterisierung von Metaevaluation bereits anklingt, sind in einer Thematisierung von Evaluationsqualität unterschiedliche Qualitätsdimensionen im Spiel. Damit ergibt sich eine in sich doppelte Perspektive auf Evaluationsqualität: Der Gedanke, dass nur Evaluationen ausreichender Qualität zur Grundlage von Nutzungsprozessen werden sollten, bzw. dass Auftraggeber und andere potenzielle Nutzergruppen sich über die Qualität der ihnen vorliegenden nutzbaren Produkte im Klaren sein sollten, verbindet sich mit einem Interesse an holistisch verstandener Evaluationsqualität in allen Hinsichten. Einem solch umfassenden Qualitätsverständnis verdankt sich unter anderem die Anlage einiger Metaevaluationen, die in der Schweiz durchgeführt wurden (z.B. Widmer 1996, Lehmann/Balthasar 2004). Andererseits lässt der Umstand, dass sich verschiedene Qualitätsaspekte des Evaluierens bzw. der Evaluationsberichte differenzieren lassen, auch eine Fokussierung auf bestimmte Qualitätsdimensionen zu, die in den Mittelpunkt des Erkenntnisinteresses einer konkreten Metaevaluation gestellt werden. Stufflebeam (2001) und Cooksy/Caracelli (2005) gehen davon aus, dass eine Metaevaluation – wie jede andere Evaluation auch – konkrete Evaluationsfragestellungen entwickeln und verfolgen wird und so einen konkreten Zuschnitt erfährt, der das zugrunde gelegte Qualitätsverständnis unter Bezugnahme auf Auftraggeber-Interessen fokussiert. Die vorliegende Studie fokussiert auf Merkmale der Evaluationen bzw. der Evaluationspraxis, von denen in begründeter Weise davon ausgegangen werden kann, dass sie die Nützlichkeit, also das in ihnen angelegtes Nutzungspotenzial, von Evaluationen erhöhen.

In einer Analyse von 18 internationalen Metaevaluationen primär US-amerikanischer Provenienz wurde festgestellt, dass solche Verfahren in ihren Konzeptionen der Qualität von Evaluationen variieren, da die betreffenden Auffassungen auch mit paradigmatischen Haltungen und Präferenzen für bestimmte Evaluationsmodelle einhergehen (Cooksy/Caracelli 2009). Teilweise haben Metaevaluationen mit emergenten Kriterien gearbeitet, die im Verlauf der Auseinandersetzung der EvaluatorInnen mit den Evaluationsberichten entstanden. Demgegenüber finden vorstrukturierte Einschätzungen ihre Grundlage in allgemein anerkannten Dokumenten über „gute Evaluation" mit handlungsleitendem Charakter, wie sie insbesondere mit Evaluationsstandards vorliegen, die auch wegen ihrer Neutralität als Grundlage einer durch vorab geklärte Kriterien gestützten Metaevaluation ausdrücklich empfohlen werden (vgl. u.a. Stufflebeam 2001). Können umfassende Metaevaluationen im Betrachtungswinkel eines holistisches Qualitätsverständnisses alle Standards heranziehen, so kann ebenso in einer fokussierten Metaevaluation ein eingegrenztes, auf die Fragestellungen der jeweiligen Metaevaluation zugeschnittenes Set von Standards zur Anwendung gebracht werden, das für die angestrebte Analyse einen klaren Kriterienraster vorgibt (so etwa Lynch et al 2003). Die vorliegende Studie folgt der Vorgehensweise einer an vorab festgelegen Analysekriterien ausgerichteten Metaevaluation, die sich auf ein spezifisch zugeschnittenes Set an Evaluationsstandards stützt.

Was die Informationsquellen einer Metaevaluation anbelangt, so stellen Evaluationsberichte den ersten und zentralen Ansatzpunkt dar. Metaevaluationen, die der gesamthaften Auffassung einer breiteren Evaluationspraxis gelten, können ausschließlich auf vorliegende Evaluationsberichte gestützt werden.(Cooksy/Caracelli 2005:31). Freilich erscheint es für eine Metaevaluation darüber hinaus auch angeraten, sich zusätzliche Informationsquellen an die Hand zu geben, um die Evaluationsprodukte eingehender als Ergebnisse von Prozessen, die mit der Auslösung und Planung von Evaluationen beginnen und in einen konkreten Kontext eingebettet sind, zu verstehen. Als relevante Datenquellen gelten beispielsweise Interviews mit Prozessbeteiligten, Beobachtungen oder Surveys (Stufflebeam 2001, Cooksy/Caracelli 2005). Letztlich ideal erscheint eine breite und detaillierte Bezugnahme auf eine Vielzahl von Materialien wie Pläne und Protokolle, Budgets, Terms of Reference, Unterlagen der EvaluatorInnen zur Datenerhebung und –auswertung, Zwischenmeldungen an die Auftraggeber und Reaktionen auf diese Zwischenmeldungen sowie weitere Kommunikationen zwischen Auftraggebern und EvaluatorInnen,, und Unterlagen über die evaluierten Programme (vgl. z.B. Stufflebeam 2001, Widmer 1996). Freilich steht einer solchen Herangehensweise ein außerordentlicher Ressourcenaufwand entgegen. Die Machbarkeit des Zugangs ist selbst bei hohem Ressourceneinsatz nicht völlig gesichert, da sie auch vom Charakter angelegter Dokumentationen und dem Erinnerungsvermögen beteiligter Akteure abhängig bleibt.[1] Schließlich bedürfte sie einer Vorab-Vereinbarung zur Kooperation und Herausgabe relevanter Unterlagen mit den betroffenen Auftraggebern und EvaluatorInnen, sodass die Festlegung des metaevaluatorischen Analysekonzepts erst im Anschluss an die betreffenden Klärungen erfolgen könnte, und die Metaevaluation sich auch auf diejenigen Ausschnitte der Evaluationspraxis zu beschränken hätte, wo ihr der detaillierte Zugriff gewährt wird. Für die vorliegende Studie war eine Abwägung des zentralen Interesses der übergreifend-umfassenden Betrachutng der mehrjährigen Evaluationspraxis unter Gesichtspunkten der pragmatischen Machbarkeit ausschlaggebend, auf eine detaillierte Rekonstruktion einzelner Evaluationsprozesse von vornherein zu verzichten. Die vorliegende Metaevaluation wurde so angelegt, dass sie ihren zentralen Ansatzpunkt an vorliegenden Evaluationsberichten findet, die sie einer qualitativen Analyse unterzieht, und die dadurch ermöglichten Erkenntnisse durch zwei ergänzende Erhebungsverfahren von übergreifend-summativem Charakter im Bezug auf die gesamte zur Debatte stehende Evaluationspraxis ergänzt.

Die Evaluationsstandards, die als Grundlage einer qualitativen Berichtsanalyse herangezogen werden, befassen sich mit der Planung, Durchführung und Präsentation von Programmevaluationen. Hier geht es um Gestaltungsweisen der Evaluationsprozesse und Evaluationsberichte, die als essentielle Voraussetzungen und Merkmale der Nützlichkeit von Evaluationen zu erachten sind. In den Evaluationberichten und -prozessen verkörpert sich ein Nutzungspotenzial, das hinsichtlich von Stärken und Schwächen analysiert werden kann und die Grundlage für faktische Nutzungsweisen bildet. In der Benennung spezifischer nützlichkeitsfördernder Qualitätsaspekte von Evaluationen beziehen sich die Standards nicht nur auf Merkmale gut nutzbarer (nützlicher) Evaluationsberichte, sondern zugleich auf die Evaluationsprozesse, durch die nützliche Evaluationen ermöglicht werden und die dann in aller Regel in Evaluationsberichten als ihren zentralen Produkten kulminieren, aber doch nur bedingt diesen Berichten entnommen werden können. Diesbezüglich wurde der Einsatz von ergänzenden Erhebungsverfahren als notwendig erachtet, die als ein strukturierter Survey mit einigen offenen Fragen sowie Interviews konfiguriert wurden. Die Evaluationsberichte sagen ferner nichts über die tatsächliche Nutzung aus, die in ihrem Gefolge zustande gekommen ist, und sie bilden auch keine Kontextfaktoren ab, die das Zustandekommen von Nutzungen ebenfalls beeinflusst haben können. Die vorliegende Metaevaluation stützt sich deswegen auf drei parallel eingesetzte Verfahren, um ein gesamthaftes Bild der Evaluationspraxis zeichnen zu können.

Mit der vorgenommen Analyse werden konzeptgemäß nicht Evaluationsstudien oder EvaluatorInnen in der Art eines Audits überprüft, sondern es werden übergreifende Einsichten in wesentliche Aspekte und Charakteristika der österreichischen Evaluationspraxis im FTI-Bereich greifbar gemacht. Die Studie stützt sich dabei auf ein in der Evaluationsforschung fundiertes Verständnis des Evaluationsprozesses, demzufolge jede Evaluation unter Rahmenbedingungen zustande kommt, unter

[1] So hat etwa die mit Mitteln der Forschungsförderung finanzierte Metaevaluation von Widmer (1996), die 10 Evaluationsstudien in größtmöglicher Vollständigkeit bewertet (Anwendung aller Standards, Einbeziehung möglichst aller verfügbarer Unterlagen des Evaluationsprozesses seitens der Auftraggeber und der EvaluatorInnen, Kontextualisierung, umfangreiche Dokumentation zu jeder behandelten Evaluationsstudie), einen Umfang von mehr als 800 Seiten. Zudem entstand erheblicher unerwarteter Mehraufwand, während sich ein Teil der geplanten Analyse dennoch als undurchführbar erwies.

denen sie sodann in unterschiedlicher Weise erfolgreich ist. Die Analyse bietet konzeptgemäß keine Antworten auf Fragen wie etwa die nach den besten oder schlechtesten Evaluationsinstituten, oder welche Programmevaluation die beste oder schlechteste war. Alle Schritte der Metaevaluation wurden auf das Prinzip der Anonymisierung gestützt, auch um die Qualität von Auskünften zu erhöhen und gute Gangbarkeit in einem in verschiedenen Hinsichten nicht unproblematischen Feld zu gewährleisten.

1.2.2 Internationale Evaluationsstandards

Ein international verankerter und abgesicherter Blickwinkel auf die Evaluationspraxis im österreichischen FTI-Bereich ergibt sich durch eine Bezugnahme auf hochrangiges Expertenwissen zu Programmevaluation, das in zwei Formen vorliegt. Zum einen vollzieht sich Theorieentwicklung und Austausch von Forschungsergebnissen über Evaluation in differenzierten ExpertInnen-Debatten in der primär englischsprachigen Fachliteratur. Zum anderen kondensiert der zentrale Gehalt dieser Debatten in professionellen Standards von nationalen Evaluationsgesellschaften, die als handlungsleitend im Hinblick auf eine möglichst gute und zielführende Evaluationspraxis gedacht sind.

Evaluationsstandards sollen die Qualität von Evaluationen als Dienstleistungen erhöhen, indem sie Anleitung für zielgerichtete professionelle Evaluation geben. Sie sollen Planung, Durchführung und sachgerechte Kritik anleiten und richten sich dabei nicht nur an EvaluatorInnen selbst, sondern auch an Auftraggeber und an die interessierte Öffentlichkeit, die Evaluationen nutzt. Sie sind Dialoginstrument und fachlicher Bezugspunkt für einen Austausch über die Qualität von professionellen Evaluationen (DeGEval 2008, DeGEval 2015). Die Funktion der Reflexion von Evaluationspraxis ist durch das Vorhandensein eines eigenen Standards zur Metaevaluation ausdrücklich verankert. Die Standards bilden demgemäß eine breit anerkannte Beurteilungsbasis für die Überprüfung laufender oder abgeschlossener Evaluationen.

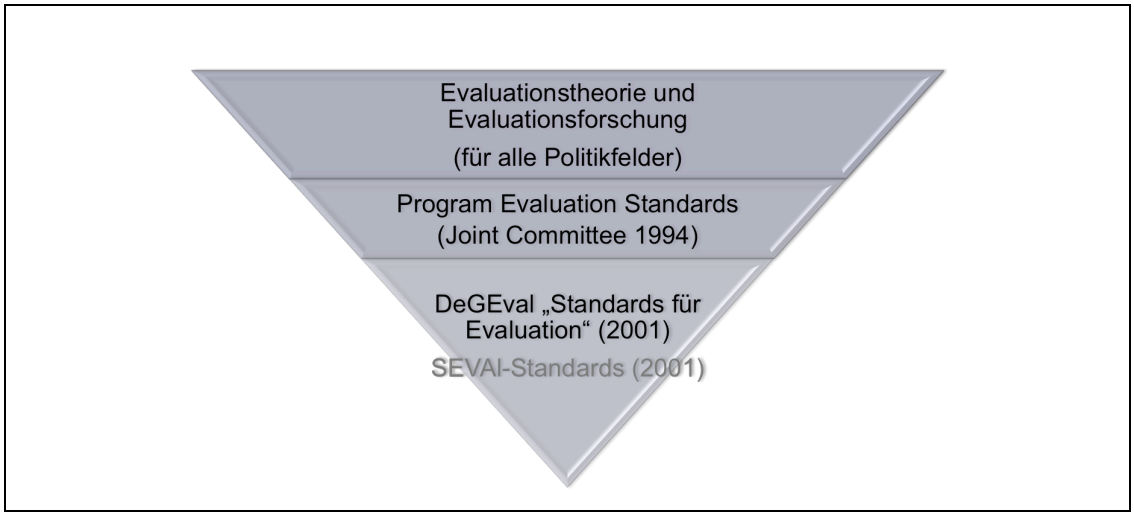
Der Ursprungspunkt der DeGEval-Standards liegt in den USA und in der Beantwortung der Nutzungskrise der Evaluation, die auch die Entstehung und Intensivierung der Nutzungsforschung zur Evaluation motiviert hat. Die *Program Evaluation Standards* wurden im US-amerikanischen Kontext durch einen umfangreichen Prozess von 1975 bis 1981 entwickelt, der zahlreiche EvaluatorInnen, Fachgesellschaften und Auftraggeber einbezog, um Defizite der damaligen Evaluationspraxis zu beheben und Evaluationen beurteilbar zu machen. Die Standards verstehen sich dabei als übergreifender Konsens, der in der Formulierung von Qualitätsaspekten verschiedene Entwürfe von Evaluationstheorien und -modellen miteinander verbindet und den gemeinsamen konzeptuellen Kern des „Unternehmens Evaluation" und dessen Qualität ausweist. Sie sind zugleich als ein Dokument konzipiert, das Weiterentwicklungen des Qualitätsverständnisses gegebenenfalls durch eine Überarbeitung aufnehmen soll (vgl. Madaus/Scriven/Stufflebeam 1984: xi-xiii). 1994 erscheinen die US-amerikanischen *Program Evaluation Standards* in ihrer zweiten Entwicklungsstufe (Joint Committee & Sanders 1994, im Folgenden „JC-Standards"). In dieser Fassung werden sie sodann auch als nationaler US-amerikanischer Standard durch die ANSI akkreditiert.[2]

Die Schweizerischer Evaluationsgesellschaft SEVAL und die für Deutschland und Österreich gebildete Evaluationsgesellschaft DeGEval übernehmen, kurz nach ihrer Entstehung, 2001 die JC-Standards in leicht adaptierter Form (Widmer et. al. 2001, DeGEval 2001). Die DeGEval- und SEVAL-Standards werden dabei vor allem unter pragmatischen Überlegungen kürzer gefasst, doch in der Absicht, die Anschlussfähigkeit an die JC-Standards zu erhalten. Zusätzlich wurden die JC-Standards im vollen Originalwortlaut 2001 auf Deutsch verfügbar gemacht und erschienen 2006 in zweiter Auflage (Joint Committee & Sanders 2006). Die JC-Standards bilden somit für die DeGEval-Standards einen "breiten fachlichen Hintergrund, der für Beauftragung, Planung, Durchführung und Evaluation von Evaluationen konsultativ genutzt werden kann" (Beywl & Taut 2000: 366). Differenzen zwischen den DeGEval-Standards und den JC-Standards existieren in Gestalt der Weglassung eines Einzelstandards, um den Anwendungsbereich der Standards zu erweitern, und in

2 Die US-amerikanischen Program Evaluation Standards wurden seitdem in einem umfangreichen Prozess neuerlich überarbeitet und liegen seit 2011 in ihrer dritten Fassung vor (Yarbrough et al 2011). Diese Fassung wurde bislang nur ansatzweise ins Deutsche übertragen, und die enthaltenen Veränderungen wurden von DeGeval und SEVAL bislang nicht aufgegriffen.

der Ergänzung eines Einzelstandards, um den betreffenden Aspekt der Evaluationsplanung und -durchführung deutlicher zu betonen.

Abbildung 1: DeGEval-Standards in übergreifender Perspektive



Die Evaluationsstandards fassen „Güte und Wert" von Evaluationen in vier Qualitätsdimensionen: Nützlichkeits-, Durchführbarkeits-, Korrektheits- und Genauigkeitsstandards. Sie beschreiben damit das Ideal „guter" oder „gelungener" Evaluation, die umfassend als Prozess und Produkt verstanden wird, an dem sowohl EvaluatorInnen als auch Auftraggebende beteiligt sind, und gegebenenfalls auch noch weitere Rollenträger. Die einzelnen formulierten Ansprüche zeigen dabei einige Überlappungen und stehen teils auch in einem konkurrierenden Verhältnis zueinander. Es besteht Konsens, dass jede Evaluation in ihrem Kontext und ihrer spezifischen Situation eine bewusste Gewichtung zwischen den Standards vornehmen wird. Von keiner Evaluation kann erwartet werden, dass sie alle Standards in gleicher Weise erfüllt, und mit der Anwendung der Standards ist auch nicht die Absicht verbunden, eine Evaluation, bei der ein bestimmter Standard nicht auf eine ganz bestimmte Weise erfüllt wurde, von vornherein abzuwerten. Die Anforderungsebene kann viel eher dahingehend beschrieben werden, dass eine bewusste Auseinandersetzung mit der Handhabung der verschiedenen Qualitätskriterien erfolgt und Entscheidungen, wie im konkreten Fall Qualität erreicht werden soll, umsichtig getroffen werden. Evaluationsqualität im Sinne der Standards ist somit ein gemeinsames Produkt von AuftraggeberInnen und EvaluatorInnen. Dabei sind die Standards als diesbezügliche Maximalansprüche bzw. Zielvorstellungen konzipiert, nicht als Mindesterfordernisse. Dies macht sie besonders geeignet zur Thematisierung von Verbesserungspotenzial, wie sie von der vorliegenden Studie angestrebt wird (Beywl 2001, Beywl & Widmer 2006, Widmer & De Rocchi 2012: 160-162). Die primäre Nutzungsmöglichkeit liegt in der Verwendung als nützlicher Ratgeber zur Bewältigung anstehender Herausforderungen für die Evaluationspraxis (Widmer (2011, 26f.). Die Evaluationsstandards können als ein Rahmen verstanden werden, der Indikatoren für die Entwicklungsstufe eine Evaluationspraxis in einem Einsatzbereich von Evaluation oder in einem Land bereit hält (Beywl/Speer 2004).

Die Standards befassen sich mit Gesichtspunkten der sachgerechten Evaluationsplanung, –durchführung und -präsentation, die nicht mit Gesichtspunkte und Annahmen über die Gestaltung der evaluierten Politiken selbst verwechselt werden dürfen (Beywl &Taut 2000: 367). Sie dienen der Analyse und Verbesserung von Evaluationen, so wie sie von ihren AuftraggeberInnen und EvaluatorInnen in den erteilten Evaluationsaufträgen definiert werden, und bewegen sich damit auf einer anderen Ebene als normative Aussagen darüber, welche Evaluationsziele in einem Evaluationsprojekt verfolgt oder welche inhaltlichen Evaluationsfragen gestellt werden hätten sollen.

Eine eindeutige oder verbindliche Operationalisierung der Standards für Zwecke einer empirischen Meta-Untersuchung existiert nicht. Die Standards stellen keine Checklist dar, und das Handbuch der Evaluationsstandards warnt davor, die Betrachtungsweise in einer „Checklistenmentalität" zu sehr zu vereinfachen (JCSEE/Sanders 2006: 47). Die Standards erschließen sich erst vollständig, wenn auf sie

zur Gänze eingegangen wird. Abbildung 2 zeigt die konkrete Formulierungsweise, in der die Standards vorliegen.

Abbildung 2: Format eines Standards	
DeGEval (2008)	**Joint Committee (1994, dt. 2006)**
Nummer und Name: Zuordnung zur Gruppe, laufende Nummer, Benennung	
Standard-Formulierung: 1 bis 3 Sollens-Aussagen, die wünschbare Merkmale einer Evaluation konkretisieren	
Übersicht: Begriffsklärung/Einführung, Schlüsselbegriffe des Standards und Hinweise zu seiner Anwendung	
---	**Richtlinien**: Vorschläge für Verfahren, um den Standard zu erreichen bzw. Strategie zur Fehlervermeidung
---	**Fallstricke**: Hindernisse der Umsetzung bzw. Fehler, die unerfahrene EvaluatorenInnen machen
---	**Anschauungsbeispiele**: Fallbeispiele von Evaluationen, in denen die Anwendung gelang bzw. misslang, mit Analyse

Erst der Rückgriff auf die JC-Standards 1994 (dt. 2006) eröffnet also den Zugriff auf Elemente, die für eine Metaevaluation besonders wertvolle Zugriffspunkte bilden (Beywl 2006). Um eine möglichst produktive Auseinandersetzung mit der Nützlichkeit österreichischer Programmevaluationen im FTI-Bereich zu ermöglichen, verfährt die vorliegende Untersuchung so, dass die DeGEval-Standards (4. Auflage 2008) mit Rückgriffen auf die Program Evaluation Standards 1994 in der deutschsprachigen Fassung von 2006 herangezogen werden. Während in der internationalen Evaluationsforschung gelegentlich in der Anwendung der Standards mit Checklisten gearbeitet wurde (vgl. Cooksy & Caracelli 2009), wird hier im Interesse einer differenzierten Betrachtungsweise, die auch einen Entdeckungszusammenhang darzustellen vermag, eine qualitative Auswertung vorgezogen, die auch soweit Offenheit mit sich bringt, um auf Eigenschaften des untersuchten Materials reagieren zu können.

1.2.3 DeGEval-Standards und fteval-Standards

Zu den Leistungen, die die *Plattform fteval* für die Entwicklung der österreichischen Evaluationspraxis im FTI-Bereich erbracht hat, zählt auch die Formulierung der fteval-Standards, die mittlerweile in der dritten Fassung vorliegen (Plattform Forschungs- und Technologieevaluierung 2003, 2005, 2013). In den fteval-Standards werden Fragen der Nutzung von Evaluationen und das Desiderat eines ausformulierten Evaluationssystems, innerhalb dessen Evaluationen eine klare Position im institutionellen Arrangement zukommt, angesprochen. Eine eingehende Prüfung der fteval-Standards ergibt allerdings klare Einschränkungen ihrer Verwendbarkeit für eine Studie, die Fragen der Nützlichkeit von Evaluationen so gut wie möglich ausloten will: Zum Einen waren Nutzungsaspekte in den früheren Fassungen der fteval-Standards, wie sie während nahezu des gesamten Beobachtungszeitraums der vorliegenden Studie vorlagen bzw. in Geltung waren, nur abrisshaft und nicht auf der Differenzierungsebene der DeGEval-Standards angesprochen. Zum Zweiten zeigen die fteval-Standards in der Neufassung (2013) in der vertieften Einlassung auf Nutzungsaspekte eine klare Annäherung an die DeGEval-Standards. Sie bleiben jedoch auch hier hinter der Differenzierungsebene der DeGEval-Standards zurück, da sie weder die Erläuterungen enthalten, die die DeGEval-Standards geben, noch die Literaturhinweise zu verfügbaren Forschungsgrundlagen enthalten, die die DeGEval- und JC-Standards ausweisen.

Die DeGEVal-Standards als Bestandteil einer breiteren Programmfamilie verstehen sich als international abgesicherter Bezugspunkt, anhand dessen die österreichische Evaluationspraxis im FTI-Bereich sinnvoll, produktiv und neutral gespiegelt werden kann. Sie verstehen sich zugleich als ein sehr gut geeignetes Analyseinstrument durch ihre detaillierten Ausführungen zur Nützlichkeit von Evaluation. Es wurde kein direkter Vergleich zwischen den DeGEval-Standards und den fteval-Standards angestrebt. Eine Kommentierung der fteval-Standards war nicht Bestandteil des Auftrags.

Wenn die *Program Evaluation Standards* auch ursprünglich in Bezug auf Evaluationen im Bildungs-, Gesundheits- und Sozialbereich entstanden sind, so besteht heute breite Übereinstimmung, dass sie auf Programme aller Art in verschiedensten Politikbereichen angewendet werden können (vgl. Widmer & Beywl 2006). Die Standards sind ausdrücklich auch für FTI-Evaluationen gedacht (DeGEval 2008), wurden der Plattform fteval direkt vorgestellt (Beywl 2001), und wurden im FTI-Bereich auch bereits eingesetzt (Good 2006 und 2012). Den Standards liegt ein weiter Programmbegriff zugrunde[3] und es besteht kein Zweifel, dass die in der vorliegenden Studie betrachteten Evaluationen sich innerhalb dieser breiteren Definition des möglichen Anwendungsbereichs ansiedeln. Auch gehen aus den fteval-Standards keine Besonderheiten hervor, die auf eine nur eingeschränkte Anwendbarkeit der DeGEval-Standards im österreichischen FTI-Bereich schließen lassen würden.

Nicht zuletzt sind die DeGEval-Standards auch für diejenigen Mitglieder der *Plattform fteval*, die gleichzeitig Mitglieder der DeGEval sind, prinzipiell als richtungsweisend auf dem Weg der Selbstverpflichtung anzusehen (vgl. Astor et al 2014). Hinsichtlich der in den fteval-Standards angesprochenen Zielsetzung des Entwurfs von Evaluationssystemen lässt die in dieser Metaevaluation gewählte Vorgehensweise auch positive Beiträge erwarten, indem sie essentielle nutzungsbezogene Merkmale der Evaluationspraxis und darin wirksame Einflussfaktoren zutage fördert. Die Herangehensweise der vorliegenden Untersuchung verdankt sich der evaluativen Perspektive des verbesserungsorientierten Denkens, nicht derjenigen einer Kontrolle. Dafür wären auch die Voraussetzungen gar nicht vorgelegen, da die herangezogenen DeGEval-Standards in den Vertragsverhältnissen zwischen den AuftraggeberInnen und den EvaluatorInnen, die den untersuchten Berichten zugrunde lagen, keine Geltung besaßen.

Wenn es um eine sachgerechte Anwendung der Standards zu tun ist, dürfen freilich Gesichtspunkte und Annahmen, die die Gestaltung der evaluierten FTI-Politiken selbst betreffen, nicht mit Gesichtspunkten der sachgerechten Evaluationsplanung und -durchführung verwechselt werden (vgl. Beywl &Taut 2000: 367). Evaluation von FTI-Politiken wird in der Expertenliteratur als durch zwei Hauptfaktoren motivierte Praxis aufgefasst: Kommt FTI-Politik in Europa zunehmend unter den Druck einer Rechenschaftslegung für öffentliche Ausgaben, so wird sie zugleich zunehmend zu einem Einsatz von Staaten und Regionen zur Erhöhung ihrer Wettbewerbsfähigkeit und zur Erreichung übergeordneter sozio-ökonomischer Ziele. Während neue Politiken und Maßnahmen auf neue Entwicklungen, Bedarfslagen und wahrgenommene Möglichkeiten des FTI-Bereichs reagieren, soll in Evaluationen dieser Politiken und Maßnahmen möglichst unmittelbar erkannt werden, ob diese Politiken und Maßnahmen auch „funktionieren" („what works").

Wie FTI-Politik gesamthaft beschrieben werden kann, stellt eine unabgeschlossene Diskussion dar, die auch von politischen Initiativen immer wieder neue Anreize erhält. In diesem Zusammenhang wird immer wieder für breiter angelegte Portfolio- oder Systemanalysen votiert (z.B. Edler 2008), und ein Bedarf der Verfügbarkeit immer umfangreicherer Datenbasen angemeldet. Mit dem EU-Projekt „INNO-Appraisal" (MIOIR et. al. 2010) und dem sogenannten „NESTA-Compendium" (MIOIR ed. 2013) finden sich in letzter Zeit zwei groß angelegte Versuche, die europaweite Evaluationspraxis im FTI-Bereich systematisch zu beschreiben und auf zentrale Merkmale zu untersuchen. Die Analyse des "NESTA-Compendiums" ergibt, dass einer Unterstützung von Politiklernen trotz der Vielzahl an Arbeiten immer noch beträchtliche Grenzen gesetzt sind. Die Diagnose benennt Faktoren wie die Konzeptualisierung von Wirkungsweisen und eine weite Verbreitung von "Erfolgsmetriken", die intervenierende Faktoren oder unintendierte Effekte nicht erfassen können. Edler et al (2014) ziehen

[3] „Programme sind beschriebene und durchgeführte, intentional aufeinander bezogene Bündel von Interventionen, Maßnahmen, Projekten oder Teilprogrammen, die aus einer Folge von auf ausgewiesene Ziele hin ausgerichteten Aktivitäten / Interaktionen bestehen, welche auf der Basis von verfügbaren Ressourcen durchgeführt werden und darauf gerichtet sind, vermittels bereitgestellter Leistungen (outputs) bestimmte, bei bezeichneten Zielgruppen oder im sozialen System zu erreichende Ergebnisse (outcomes) auszulösen." (Beywl & Taut 2000: 362)

auf Basis dieser Materialien und deren Analyse den Schluss, dass Evaluation im FTI-Bereich gerade hinsichtlich Kenntnissen und Bemühungen zu Nutzen- bzw. Nützlichkeitsaspekten ein merkliches Defizit aufweist. Parallel dazu ergibt eine breite bibliometrische Analyse von Publikationen der Evaluationsforschung quer über verschiedene Politikfelder, dass FTI-Evaluation sich stark unabhängig von anderen Bereichen und Hauptdiskussionslinien der Theoriebildung der Evaluation entwickelt hat, und dass sich gerade in der Intensität der Befassung mit dem Thema des Nutzens bzw. der Nützlichkeit eine der größten Abweichungen des relativ isolierten FTI-Bereichs von den Hauptströmungen der Evaluation findet (Gök & Mollas-Gallart 2014).

Die österreichische Situation ist durch die Einbettung unterschiedlicher Evaluationsfunktionen in ein komplexes Governancesystem gekennzeichnet (vgl. Pichler 2009), das neuerdings um die Komponente der Wirkungsorientierten Programmplanung ergänzt wurde (Pichler 2013). Zinöcker & Dinges (2009) und Astor et. al. (2014) beschreiben in ihren Bestandsaufnahmen Veränderungsdynamiken in der FTI-Evaluation, in denen das Methodenrepertoire der EvaluatorInnen angereichert wurde, und zum Andern Auseinandersetzungen von Auftraggebern mit Evaluationsergebnissen zugenommen haben.

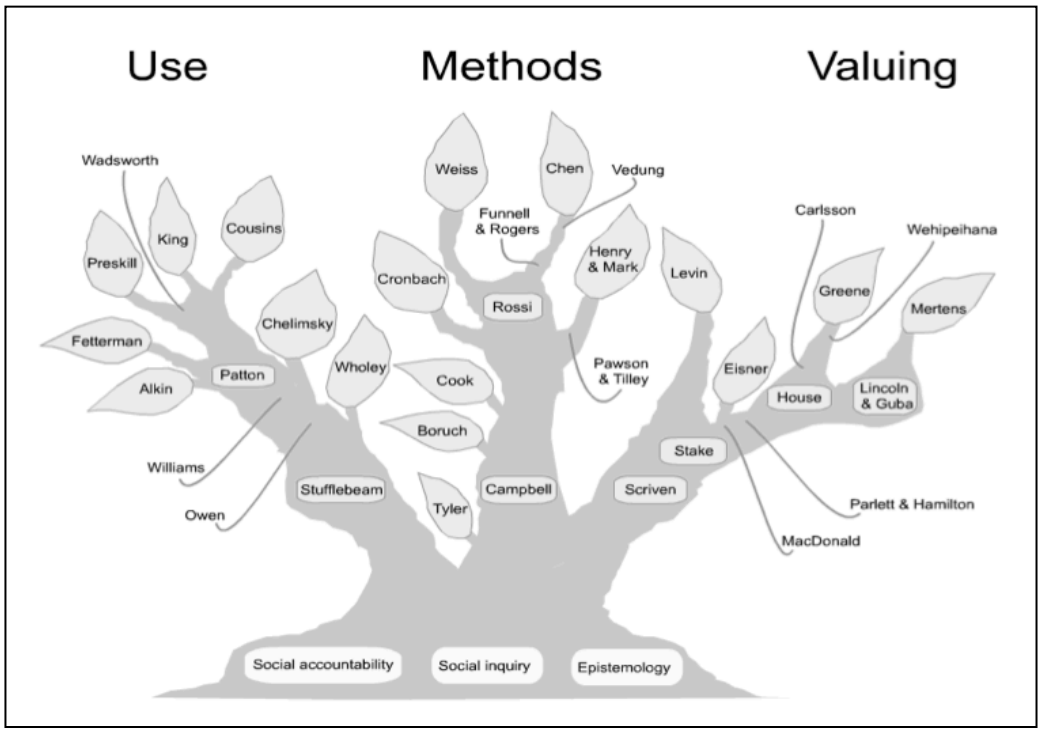
Die in diesen Reflexionen angedeuteten Bedarfslagen der FTI-Evaluation finden sich in den für das Analyseverfahren ausgewählten Standards abgedeckt. In der Anwendung der Standards bietet gerade die Interpretation im Rückgriff auf die JC-Standards und weiterhin mit damit verbundener Expertenliteratur die Möglichkeit, besonderes Augenmerk auf Gesichtspunkte zu legen, denen innerhalb der Evaluation und Politikberatung zu FTI spezielle Relevanz zugesprochen wird.

1.2.4 DeGEval-Standards im Verhältnis zur Evaluationstheorie

Spätestens anfang der 1990er-Jahre setzt sich in der Evaluationstheorie die Auffassung durch, dass jede gute Theorie Fragen des Nutzens bzw. der Nützlichkeit einbeziehen und behandeln wird (Shadish/Cook/Leviton 1991: 54ff). Evaluationen können zu einer Verbesserung der evaluierten Maßnahmen und in der Folge zur Verbesserung der sozio-ökonomischen Verhältnisse, auf die diese Maßnahmen zielen, nur beitragen, wenn sie auch genutzt werden. Verschiedene Evaluationsansätze und -modelle wurden entwickelt, in denen sich Überlegungen über die Erreichbarkeit von Evaluationszwecken mit nutzenbezogenen und epistemologischen Abwägungen verbinden. Klassifikationen der vorfindlichen Evaluationsmodelle wurden in unterschiedlicher Weise vorgenommen (so etwa Madaus et. al. 1984, Stufflebeam 2001, Stufflebeam & Shinkfield 2007, Widmer & De Rocchi 2012, Stufflebeam & Coryn 2014). Eine Analyseform, die in der Literatur als Alkin's Theoriebaum bekannt ist (Alkin 2012), verzeichnet mehr als 20 EvaluationstheortikerInnen danach, welchen Stellenwert dem Thema des Nutzens bzw. der Nützlichkeit in der jeweiligen Gesamtkonzeption von Evaluation zugemessen wurde. Diese gliedernde Bestandsaufnahme der evaluationstheoretischen Ansätze weist Nutzung (*use*), Methoden (*methods*), und Bewertung (*valuing*) als die Grundfragestellungen in der Theoriebildung zur Evaluation aus (vgl. Abbildung 3 auf der folgenden Seite).

Dieses Spannungsfeld generiert immer wieder neue Evaluationsansätze, die bestrebt sind, verbesserte Lösungen für Grundfragen der Evaluation anzubieten. Unterschiedliche Evaluationsansätze lassen sich demzufolge so verstehen, dass sie bestmögliche Qualität einer Evaluation in einer spezifischen Tarierung der Parameter Nutzung, Methoden und Bewertung aufsuchen. „Qualität" in der Evaluation erweist sich bei näherem Hinsehen als vielschichtiger Begriff, und die Standards gelten in der Art und Weise ihrer Formulierung in vier Standard-Gruppen eben diesem Umstand. Mit dem Begriff der Qualität einer Evaluation wird zunächst oft die Güte der Anwendung von meist sozialwissenschaftlichen Methoden assoziiert. Eine nähere Einlassung auf die Literatur der Evaluationstheorie und Evaluationsforschung zeigt jedoch, dass der Begriff der Qualität deutlich breiter gefasst wird, und dass damit weitere Parameter ebenso angesprochen werden, die aufgrund jahrzehntelanger Evaluationspraxis und konzeptiver Überlegungen von EvaluationstheoretikerInnen als unverzichtbare Grundbausteine für das Gelingen des „Unternehmens Evaluation" betrachtet werden. Das „Unternehmen Evaluation“ verdankt insgesamt seine kreative Weiterentwicklung gerade der aktiven Auseinandersetzung mit diesen Parametern, in der es einer wachsende Bandbreite von Evaluationsansätzen zahlreiche Differenzierungen zu Grundfragestellungen hervor gebracht hat. Zugleich wird davon auszugehen sein, dass man es hinsichtlich von „Qualität" mit einer lebendigen, und wohl auch nicht endgültig abschließbaren, Diskussion zu tun hat, die als solche die Weiterentwicklung der professionellen Evaluation vorantreibt.

Abbildung 3: Alkin's Theoriebaum



Verschiedene Evaluationstheorien bzw. Evaluationsansätze unterhalten einen expliziten positiven Bezug auf die Standards und verstehen sich als aktive Auseinandersetzung mit deren Ansprüchen (z.B. Patton 1997, Rossi/Freeman/Lipsey 1999, Owen/Rogers 1999). Lehrbücher zu Evaluation beziehen sich teilweise direkt auf die Standards (z.B. Stufflebeam/Coryn 2014). Ebenso formuliert die Nutzungsforschung zur Evaluation einen Bezug auf die Standards, vor allem dann, wenn sie Nutzungsforschung als einen empirischen Test von vorgelegten Theorien versteht (so etwa Kirkhart 2000, Stufflebeam 2001). Mit den Standards liegt ein dezidiert neutraler Bezugspunkt vor, um eine wohlbegründete Auseinandersetzung mit Qualitätskriterien in ihrer Vielschichtigkeit zu führen. Die vorliegende Studie zieht als ihre Bezugspunkte für die Thematisierung der Nützlichkeit von Evaluationen nicht einzelne ausgewählte Evaluationstheorien heran, die sich zu dieser Thematik in einer spezifischen Weise positionieren, wie etwa Pattons *Utilization-Focused Evaluation* oder seine *Developmental Evaluation*. Vielmehr sucht sie den neutralen, von spezifischen Evaluationstheorien und -modellen unabhängigen Boden auf, der in Gestalt der Evaluationsstandards vorliegt.

1.2.5 Evaluationsforschung zur Nutzung von Evaluationen

Die Evaluationspraxis in den USA erfährt bereits in den 1970er-Jahren eine Nutzungskrise, auf die sowohl in der Theoriebildung der Evaluation als auch durch verstärkte empirische Auseinandersetzungen mit der Nutzung von Evaluation reagiert wird. Wenn die ursprünglich von EvaluatorInnen gehegte Annahme offensichtlich nicht zutrifft, dass Evaluationsergebnisse von Programmverantwortlichen und PolitikerInnen direkt und unmittelbar genutzt werden, um Entscheidungen über die evaluierten Programm zu treffen, wie kann dann die Entstehung von Evaluationsnutzen dann begriffen und beschrieben werden, und wie können und sollten Evaluationen deshalb gestaltet werden? Da die Entstehung von Nutzen zentrales Anliegen von Evaluation ist, befasst sich eine umfangreiche empirische Nutzungsforschung intensiv mit der Frage, wie Nutzen aus Evaluationen in unterschiedlichen Settings tatsächlich entsteht. Forschungen zur Nutzenentstehung sind gerade auch dadurch motiviert, dass auch bei nützlichen Evaluationsprodukten eine Nutzung ausgeblieben ist (Hughes/Leviton 1981). Es entstehen zahlreiche Einzelstudien, die jeweils auf ihre Weise die Nutzung bestimmter Evaluationen oder Gruppen von Evaluationen verfolgen. Auf diese stützen sich wiederum Arbeiten, die die gefundenen Fakten zu systematisieren und zu theoretisieren

versuchen (u.a. Weiss 1977, Hughes & Leviton 1981, Cousins & Leithwood 1986, Preskill & Caracelli 1997, Shulha & Cousins 1997, Weiss 1998a). Die über Jahrzehnte betriebene Nutzungsforschung hat den Stellenwert, evaluationstheoretische Debatten zu informieren und die Integration nutzungsbezogener Erkenntnisse in neuen Entwürfen von Evaluationsansätzen und –modellen zu ermöglichen. Arbeiten der Nutzungsforschung zur Evaluation wie etwa Alkin, Daillak, & White (1979), Patton,et al. (1977) oder Weiss (1973) haben die evaluationstheoretischen Debatten entscheidend geprägt (vgl. Shulha/Cousins 1997). Mittlerweile bereits klassische evaluationstheoretische Reaktionen auf Nutzungsfragen sind etwa M.Q.Pattons ‚Utilization-Focused Evaluation' (1997) und die vom selben Autor stammende ‚Developmental Evaluation' (Patton 2010). Zugleich wird der Nutzungsforschung das Potenzial zugesprochen, Evaluationstheorien mit ihren Konzepten, wie eine wertvolle Evaluation zustande kommen kann und soll, einem empirischen Test zu unterziehen (z.B. Kirkhart 2000).

Die über Jahrzehnte betriebene Nutzungsforschung hat eine empirisch abgestützte und allgemein anerkannte Typisisierung von Evaluationsnutzen hervorgebracht. Fünf Typen von Nutzen werden dabei regelmäßig unterschieden:

- **Instrumenteller Nutzen**: Evaluationsergebnisse und/oder Empfehlungen werden direkt zur Entscheidungsfindung über das evaluierte Programm genutzt.
- **Konzeptueller Nutzen**: Evaluationsergebnisse helfen Programmbeteiligten, über den Evaluationsgegenstand zu lernen und neue Sichtweisen darauf zu entwickeln.
- **Symbolischer Nutzen**: Das Vorliegen eines Evaluationsberichts oder die Tatsache, dass überhaupt evaluiert wird, dient zur Rechtfertigung bereits zuvor getroffener Entscheidungen, oder zur formalen Untermauerung, dass mit dem Programm rational umgegangen wird, unter Umständen um andere Akteure in der politischen Sphäre vom Programm zu überzeugen.
- **Aufklärung**: Evaluationsergebnisse reichern das verfügbare Wissen an, das von Akteuren im Umfeld des Evaluationsgegenstands genutzt werden kann. Die Anreicherung verfügbaren Wissen kommt darüber hinaus auch Personen bzw. gesellschaftlichen Sphären zugute, die nicht unmittelbar am Programm beteiligt waren oder sind.
- **Prozessnutzen**: Kognitive, verhaltensförmige oder organisatorische Veränderungen treten bereits im Laufe des Evaluationsprozesses ein, bereits vor Vorliegen der Ergebnisse bzw. unabhängig von diesen.

Die Konzeptualisierung der Nutzungsformen wird des Weiteren ergänzt um eine Diskussion über Formen und Gründe der Nicht-Verwendung von Evaluation. Diese Diskussion weist darauf hin, dass Evaluationen auch mit Recht nicht genutzt werden, da sie in unzureichender Weise erstellt wurden. Sollte ein Auftraggeber zur Ansicht gelangen, dass eine Evaluation unzureichend durchgeführt wurde, so würde die Nutzung einen Missbrauch darstellen (Alkin & Taut 2003). Diese Problematik verweist auf Qualitätsüberprüfungen von Evaluationen durch ihre Auftraggeber, auf die erreichte Durchführungsqualität von Evaluationen auf verschiedenen Ebenen, und auch auf die Glaubwürdigkeit der Evaluation, die beim Auftraggeber bei ihrer Planung und Durchführung erzielt werden konnte.

Nachdem in einem Frühstadium der Entwicklung von Evaluation vor allem die direkte Nutzung von Ergebnissen im Mittelpunkt stand, verfügt die Evaluationsforschung heute über einen wesentlich breiteren Nutzungsbegriff. In den 1990er-Jahren verlagert sich die Aufmerksamkeit für die Entstehung von Nutzen zunehmend von der direkten Nutzung von Evaluationsergebnissen hin zum Prozessnutzen und dessen möglicher Unterstützung. Dazu kommt als weiterer wesentlicher Schritt eine Bezugnahme auf Organisationsmerkmale und Organisationslernen (vgl. Preskill and Caracelli 1997, Shulha & Cousins 1997). Dabei treten Zugangsweisen zu Evaluation in den Vordergrund, die in der Evaluationsforschung als „kollaborativ" und „partizipativ" angesprochen werden. Diese beruhen auf breiterer und intensiverer Interaktion mit den Auftraggebenden, Programmbeteiligten und von einem Programm Betroffenen (oft bezeichnet als „Klienten") und heben sich damit von einer „objektivistischen" Herangehensweise an Evaluation, die in erster Linie oder ausschließlich auf die „Wahrheit der Daten" setzt, ab.

Die vorliegenden übergreifenden Systematisierungen von Evaluationsnutzen benennen jeweils eine Reihe von Faktoren, die Einfluss auf die Förderung oder Behinderung von Nutzung haben. Cousins & Leithwood (1986) identifizieren auf Basis einer systematischen Untersuchung von 65 empirischen

Studien zu verschiedenen Evaluationsfeldern 12 Faktoren. Fleischer & Christie (2009) identifizieren in einer Umfrage unter allen Mitgliedern der American Evaluation Society, quer über alle Evaluationsfelder, 15 essentielle Faktoren. Johnson et al (2009) bestätigen in einer umfangreichen Analyse neuerer Forschungen die Faktoren von Cousins & Leithwood (1986) und reichern sie zugleich um weitere Faktoren an, die auf die Förderung von Prozessnutzen und Organisationslernen Bezug nehmen. Diese Ergänzung erfolgt jedoch in einer unsystematischen Weise und ist daher für die Strukturierung einer empirischen Untersuchung kaum geeignet (vgl. auch Hense/Widmer 2013).

Die voranschreitende Auseinandersetzung mit Evaluationsnutzung zeigt immer deutlicher, dass die Entstehung von Nutzen als eine voraussetzungsvolle und nur recht bedingt vorhersehbare Produktion von Evaluationseffekten begriffen werden muss und ohne Bezugnahme auf den Kontext, innerhalb dessen Nutzungsweisen zustande kommen, nicht auskommen kann. Wie Evaluationen mit dieser Kontextabhängigkeit der Nutzenentstehung umgehen können bzw. sollen, wird zum Gegenstand der sogenannten Weiss-Patton-Debatte (zusammengefasst in Alkin 2003), die geschärfte Positionen zur Nutzungsproblematik, aber doch keine eindeutige bzw. eindeutig bessere Lösung hervorbringt. Fortschritte in der Nutzungsforschung und in der darauf reagierenden Evaluationstheorie drücken sich vor allem in einer Erweiterung der Aufmerksamkeitspunkte aus: „We used to do empirical studies to identify the correlates of use; we studied characteristics of studies, characteristics of potential users, and communication strategies that were associated with greater use of results. But we have come to a growing realization of how complicated the phenomenon of use is and how different situations and agencies can be from one another. We are also aware of the higher-order interactions among characteristics of the study, the evaluator, the setting, and the communication methods used. Because of these complexities, it is conceptually and theoretically difficult to reduce the elements to a set of quantitative variables. (...) If our understanding of the use of evaluations has advanced, it is partly because we have new ways of thinking about it." (Weiss 1998a: 23)

In den 2000er-Jahren wird der Begriff des "Evaluationseinflusses" (evaluation influence) geprägt. Dieser Begriff weist darauf hin, dass bislang übliche Fassungen des Nutzungs-Begriffs zu vereinfachend sein könnten und dass viele indirekte Wege eines nicht unmittelbar nachvollziehbaren Nutzens ebenfalls existieren. Henry & Mark (2003) und Mark & Henry (2004) entwickeln eine differenzierte Typologie, auf welchen Ebenen sich Prozesse der Veränderung ansiedeln lassen, die in der einen oder anderen Weise Nutzen verkörpern oder herbeiführen können. Sie differenzieren zwischen 4 Mechanismen und 3 Veränderungsebenen, auf denen diese Mechanismen jeweils wirken können. Diese finden sich in der folgenden Abbildung dargestellt.

Abbildung 4: Modell alternativer Mechanismen des Zustandekommens von Evaluationseinfluss

Type of Process/Outcome	*Level of Analysis*		
	Individual	*Interpersonal*	*Collective*
General influence	Elaboration Heuristics Priming Skill acquisition	Justification Persuasion Change agent Minority-opinion influence	Ritualism Legislative hearings Coalition formation Drafting legislation Standard setting Policy consideration
Cognitive and affective	Salience Opinion/attitude valence	Local descriptive norms	Agenda setting Policy-oriented learning
Motivational	Personal goals and aspirations	Injunctive norms Social reward Exchange	Structural incentives Market forces
Behavioral	New skill performance Individual change in practice	Collaborative change in practice	Program continuation, cessation, or change Policy change Diffusion

Quelle: Mark & Henry (2004): 41

Im Kern besagt diese Differenzierung von Mark & Henry (2004), dass alle Nutzungsarten einer Evaluation sowohl mit Ergebnisnutzen als auch mit Prozessnutzen einhergehen können. Zahlreiche, in spezifischen Settings jeweils unterschiedlich miteinander verknüpfte Faktoren können zur Wirkungsweise einer Evaluation beitragen, die zu einem bestimmten Zeitpunkt einen Stand erreicht, der einer der oben genannten Kategorien der Evaluationsnutzung zuzuordnen ist. Dieser jeweils erreichte Zustand braucht freilich nicht der Endzustand eines längerfristigen multifaktoriellen Einflussprozesses sein. Dabei wird es als unrealistisch erachtet, EvaluatorInnen für verantwortlich für Endzustände gemäß den traditionellen Beschreibungen von Nutzung halten zu wollen. Anwesend ist bei allen Differenzierungsbestrebungen bei Mark und Henry somit auch die Überlegung, die Verantwortlichkeit der EvaluatorInnen bzw. der Evaluationsgestaltung, die auch von den jeweiligen AuftraggeberInnen mit beeinflusst wird, auf kurzfristige Aspekte der komplexen Einflusspfade zu beschränken.

Die Typisierung von Mark & Henry (2004) wird als zu mikrologisch erachtet, um in einer praktischen Studie über eine größere Zahl von Evaluationen verfolgt werden zu können. Sie verlangt viel eher nach einer eingehenden qualitativen Untersuchung durch intensive Fallstudien. So haben etwa Ottoson & Martinez (2010) 23 Interviews zu einer einzelnen Evaluation durchgeführt, um auch noch verästelte und indirekte Einflusspfade zu verfolgen, die sie in Abhebung von den oben angeführten klassischen Nutzungstypen als „leveraged use“ bezeichnen. Die vorliegende Analyse stützt sich auf die Feststellung von Alkin & Taut (2003), dass bei allem Interesse für breitere Nutzenentfaltungen doch die Nutzung durch die HauptadressatInnen der Evaluation ("intended use by intended users") der erste Aufmerksamkeitspunkt sein muss, um die Arbeit von EvaluatorInnen adäquat zu reflektieren. Das Konzept des Evaluationseinflusses informiert jedoch die vorliegende Studie dahingehend, dass es sinnvoll und angebracht erscheint, nicht nur eher kurzfristige Nutzungsweisen einzelner Evaluationen, sondern gerade auch längerfristige und kumulative Wirkungen der Evaluationspraxis zu untersuchen. Es kann nicht übersehen werden, dass Evaluationen mit zunehmendem Aufbau von Evaluationskulturen und wiederholten Einsätzen in Politikbereichen nicht mehr nur ein Status als Einzelstudien zukommt, die einzelne Maßnahmen isoliert beleuchten, sondern sich verschiedene Evaluationseinsätze zumindest potenziell zu einem Wissensstrom verknüpfen (Rist/Stame 2006). Unter dieser Perspektive werden methodische Instrumente, die das Evaluationsgeschehen und seine Effekte in der Breite abzubilden vermögen (vgl. Stamm 2003, Balthasar 2007), als zielführend erachtet.

Ein wesentlicher Gesichtspunkt der neueren Auffassungsweisen von Evaluationsnutzen besteht im Hinweis, dass es unrealistisch ist, EvaluatorInnen für alleine verantwortlich für Endzustände der Nutzung halten zu wollen (Weiss 1998, Alkin & Taut 2003, Stamm 2003, Mark & Henry 2004). Damit tritt die Frage hervor, inwiefern sich Faktoren, die Nutzung beeinflussen, im Verantwortungsbereich der EvaluatorInnen ansiedeln lassen, und welche außerhalb ihres Verantwortungsbereichs liegen. Cousins & Leithwood (1986) und Johnson et. al. (2009) nehmen diesbezüglich eine Gliederung aller relevanten Faktoren in zwei Blöcke vor: erstens die Evaluations-Implementation, und zweitens das Entscheidungs- und Politiksetting. In Tabelle 1 auf der folgenden Seite wird die Aufschlüsselung und Gliederung dieser Faktoren wiedergegeben.

Nützlichkeit von Evaluationen stellt somit ein übergreifendes Konstrukt dar, das sich in der voranschreitenden Nutzungsforschung in eine Vielzahl ineinandergreifender Faktoren differenziert. Die Nutzungsforschung bezieht sich bei der Identifikation dieser Faktoren nicht direkt auf Evaluationsstandards, es findet sich jedoch in verschiedenen Arbeiten der Hinweis, dass die Standards die allgemeine Basis zu Fragen der Nutzung darstellen, und dass die Standards für die empirische Untersuchung eingesetzt werden können (so etwa Kirkhart 2000, Ottoson/Martinez 2010). Die empirische Basis für die Identifikation von wirksamen Faktoren stellt sich in der Vielzahl der Arbeiten über Evaluationsnutzen allerdings unterschiedlich dar. Insbesondere Johnson et. al. (2009) heben hervor, dass es daher angebracht ist, sich angesichts der Vielfalt von Evaluationssettings und Untersuchungsverfahren auf den Kern der Erkenntnisse zu konzentrieren bzw. zu beschränken, der über die verfügbaren Untersuchungen hinweg als gesichert gelten kann. Dieser Kern wird nach wie vor in der Kategorisierung von Cousins/Leithwood (1986) und Johnson et. al. (2009) erblickt (vgl. auch Balthasar 2007). Zusätzliche wesentliche Informationen ergeben sich aus Befragungen, die unter allen EvaluatorInnen der *American Evaluation Society* hinsichtlich ihrer Erfahrungen mit Evaluationsnutzung durchgeführt wurden (Cooksy/Caracelli 2005, Fleischer/Christie 2009). Die vorliegende Studie stützt sich in ihrer Operationalisierung von Einflussfaktoren auf die Entstehung von Evaluationsnutzen auf die in den genannten Arbeiten ausgewiesenen Faktoren. Indem sich die vorliegende Untersuchung auf Dimensionen des Nutzungsbegriffs konzentriert, die sich aus der auf

Nutzung fokussierenden Evaluationsforschung und deren Informationswert für den Umgang mit den Evaluationsstandards ergeben, führt sie die Bedingungen, die aus dem Entscheidungs- und Politiksetting erwachsen, in der ergänzenden Betrachtung durch Zusatzerhebungen mit. Denn: "There is no doubt that a lack of fit between the chosen institutional settings and the given purpose and proposed utilization reduces the possibility of the optimal use of evaluation." (Widmer et al 2004: 205).

Tabelle 1: Einflussfaktoren auf Evaluationsnutzung nach Cousins & Leithwood (1986) und Johnson et. al. (2009)

Evaluation implementation	
Evaluation Quality	characteristics of the evaluation process - sophistication of methods, rigor, type of evaluation model, type of approach to the evaluation problem, or the intensity of the evaluation activities
Credibility	of the evaluator and/or the evaluation process - objectivity, believability, appropriateness of evaluation criteria, evaluations have high face validity or are emphasized as important activities, collection of data perceived as inappropriate by decisionmakers
Relevance	of the evaluation to the information needs of the decisionmaker(s) in terms of the purpose(s) of the evaluation and the organizational location of the evaluator; evaluations reflect knowledge of the context, appealed to preferences of the decisionmakers, demonstrate insight into program operations and decisionmaking
Communication Quality	clarity of reporting results to the evaluation audience(s) in terms of style, evaluator advocacy of the results, and breadth of dissemination
Findings	positive/negative; consistent with evaluation audience expectations, value for decisionmaking, congruent with decisionmaker expectations, practical and conclusive, identifying alternative courses of action
Timeliness	in the dissemination of evaluation results to decisionmaker(s)
Decision or policy setting	
Information Needs	of the evaluation audience(s), including type of information sought, number of evaluation audiences with differing information needs, time pressure, and perceived need for evaluation.
Decision Characteristics	impact area, type of decision, program novelty, significance of the decision or evaluation problem,
Political Climate	political orientation of commissioners of the evaluation, dependence of the decisionmaker(s) on external sponsors, inter- and intraorganizational rivalries, budget fights, power struggles
Competing Information	from sources beyond the evaluation (personal observations, staff, peers, etc.) bearing upon the problem and competing with evaluation data
Personal Characteristics	defined in terms of the decisiosnmakers' organizational roles, information processing style, organizational experience, social characteristics, and so forth.
Commitment and/or Receptiveness to Evaluation	attitudes of the decisionmaker(s) toward evaluation, organizational resistance, open-mindedness, and so forth.

Die Nutzungsforschung bemüht sich um das Erkennen von Barrieren und förderlichen Bedingungen („utlilization enhancing conditions“, Alkin/Taut 2003) für die Entstehung von Nutzen und lenkt die Aufmerksamkeit auch auf die spezifischen Merkmale von Organisationen, die deren Kapazität zur Nutzung von Evaluationen bestimmen („organizational readiness for evaluation“, Cousins/Goh/Clark 2004). Die Vorhersehbarkeit der Folgewirkungen von evaluativer Evidenzproduktion ist jedoch in einer realistisch-nüchternen Sicht auf die Multidimensionalität von Nutzenentstehungen, die in den Evaluationsdebatten der 1990er-Jahre Platz greift und im Konzept

des Evaluatinseinflusses ihren stärksten Ausdruck findet, auch mit grundsätzlichen Grenzen konfrontiert. „However, it is difficult to foresee patterns of use. People who are indifferent to the evaluation at the start may get highly engaged by the results and use them to rethink assumptions, reorder their agendas, or alter program emphases or modes of implementation. Some potential users of the findings, enthusiastic at the point of initiation, may face serious distractions or political obstacles by the time results are ready, and proceed to ignore the evaluation. Some may have left their positions and moved on. Still, it is worthwhile for the evaluator to keep the potential for use in mind when choosing which questions the study will address." (Weiss 1998b:80)

Gerade für politische Nutzungskontexte wird darauf hingewiesen, dass hier neben der Verfügbarkeit von Evidenz auch stets anders gelagerte Faktoren wie feststehende Werthaltungen oder ein Interesse an Klientelen und Wählerstimmen im Spiel sein werden. Aufgrund entsprechender Erfahrungen gelangen zwei prominente Evaluatoren und Evaluationstheoretiker gegen Ende ihrer Karrieren zu den folgenden Schlüssen: „It has long been understood that how evidence factors into policy development is a function of the multidimensional and non-rational dynamics of the policy process." (Cousins 2006:1) „The best an evaluator can hope for is that the findings of evaluations are paid attention in decision making about the programs involved. In democratic decision making, many factors are involved, including evaluation findings." (Rossi 2013: 110). Es besteht Übereinstimmung in der Expertenliteratur, dass in Policy-Kontexten Evaluationen jedenfalls möglichst hohe Nützlichkeit und Qualität anstreben sollten, um beste Voraussetzungen für Nutzung zu schaffen, aber eine wirkliche Vorhersage von Nutzungsweisen und –intensitäten wegen der außerordentlichen Komplexität der Vorgänge nicht möglich ist. Für einen objektivistisch-distanzierten Evaluationsansatz unter punktueller Heranziehung von externen EvaluatorInnen, wie er für FTI-Evaluationen typisch ist, ergeben sich Grenzen, die die Harvardprofessorin C.H. Weiss zu dem Schluß gebracht hat, dass in einem solchen Setting EvaluatorInnen nur recht bedingt verantwortlich für Endzustände der Nutzenentstehung sein können: „Evaluators should not be held accountable for failures to use their results. Even when program staff know about the findings, understand them, believe them, and see their implications for improving the program, many factors can interfere with their using results for program improvement. Among the possible obstacles are conflicting beliefs within the program organization, with staff unable to agree on what the important issues are; conflicting interests between program units (...), rigidity of organizational rules and standard operating procedures that prevent adoption of improved strategies suggested by the evaluation; shifts in external conditions, such as budget cuts or changes in the political climate that make the organization unable to respond to the need for change revealed by evaluation, and so on." (Weiss 1998a: 22)

1.3 Schritte und Methodeneinsatz

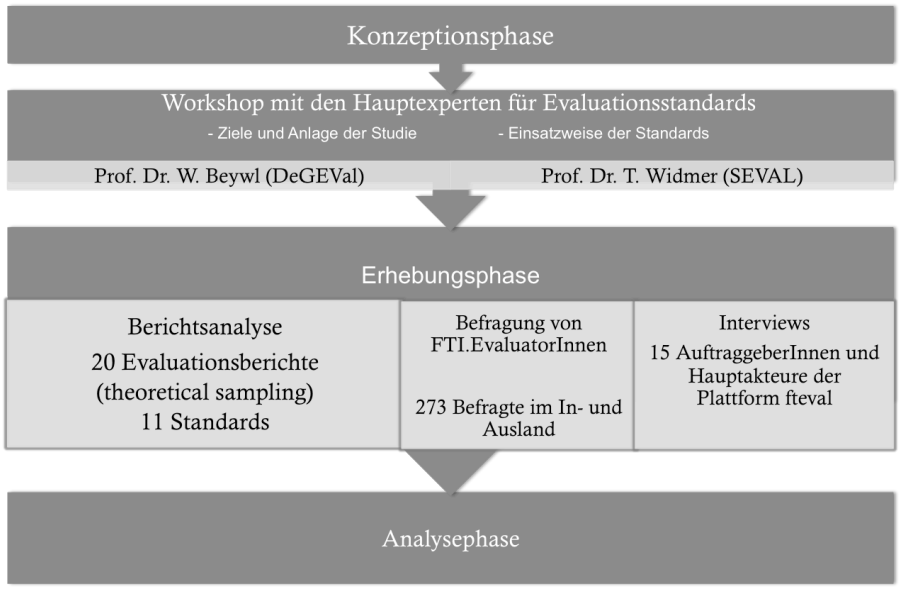
Aufgabe der Metaevaluation war, anhand eines Samples von Evaluationsberichten und zweier ergänzender Erhebungen eine allgemein-übergreifende Einschätzung der Evaluationspraxis für einen zwölfjährigen Zeitraum zu erbringen. Die der Untersuchung zugrunde gelegten Evaluationsfragen und die jeweils zu ihrer Beantwortung eingesetzten Methoden werden in der folgenden Tabelle 2 dargestellt.

Tabelle 2: Evaluationsplan

Evaluationsfragen	Methoden und Datenquellen
1. Inwiefern zeigen die Programmevaluationen im FTI-Bereich Eigenschaften, die gemäß dem internationalen Erkenntnisstand über Nützlichkeit von Evaluationen als wesentlich zu erachten sind? Welche Stärken und Schwächen liegen vor, und welche Verbesserungsmöglichkeiten lassen sich identifizieren?	• Qualitative Sekundäranalyse von Evaluationsberichten anhand der DeGEval-Standards • Semi-strukturierte Interviews mit AuftraggeberInnen bzw. potenziellen NutzerInnen • Befragung von EvaluatorInnen, die im Beobachtungszeitraum Programmevaluationen im in österreichischen FTI-Bereich erarbeitet haben
2. Wie werden Programmevaluationen im FTI-Bereich geplant, durchgeführt und genutzt? Welche Stärken und Schwächen liegen vor, und welche Verbesserungsmöglichkeiten lassen sich identifizieren?	• Semi-strukturierte Interviews mit AuftraggeberInnen bzw. potenziellen NutzerInnen • Befragung von EvaluatorInnen, die im Beobachtungszeitraum Programmevaluationen im in österreichischen FTI-Bereich erarbeitet haben

Die Metaevaluation war in vier Schritten organisiert. Sie begann mit einer Konzeptionsphase, die einen Workshop mit hochrangigen Experten umfasste und mit einer Akkordierung des weiteren Vorgehens mit dem Auftraggeber abgeschlossen wurde. Die leitenden Definitionen und methodischen Entscheidungen für die Durchführung der Metaevaluation wurden in einem Zwischenbericht dargestellt und vom Auftraggeber approbiert. Daran schloss sich eine Erhebungsphase, in der die für die beabsichtige Analyse erforderlichen Daten und Informationen beschafft wurden. Die ergänzenden Erhebungsinstrumente wurden dabei auch auf der Basis erster Zwischenergebnisse aus der begonnenen Berichtsanalyse konzipiert. Nach Vorliegen aller Ergebnisse wurden diese analysiert, wobei auch Querbeziehungen zwischen den Ergebnissen aus verschiedenen Methoden hergestellt wurden (Triangulation). Die folgende Abbildung stellt den Ablauf der Metaevaluation dar.

Abbildung 5: Schritte der Metaevaluation



In der Feindefinitionsphase des Projekts wurden zunächst eigene Recherchen zum internationalen Forschungsstand in Bezug auf Nützlichkeit und Nutzung von Evaluationen und zu Rolle und Anwendungsprinzipien der Evaluationsstandards angestellt. Die Anwendung der Standards und die Vorgehensweise der Metaevaluation wurden sodann in einem eintägigen Workshop mit den beiden Hauptexperten für die DeGEval- und SEVAL-Standards in der Geschäftsstelle des RFTE diskutiert. Dieser Workshop mit Prof. Dr. Thomas Widmer (Universität Zürich) und Prof. Dr. Wolfgang Beywl (Pädagogische Hochschule FHNW) fand am 23.10.2014 in den Räumen des Auftraggebers RFTE statt. Erörtert wurden mit den beiden Experten, die in direktem Konnex mit der US-amerikanischen Diskussion stehen und zentrale Rollen bei der Formulierung von Evaluationsstandards im deutschsprachigen Raum innehatten, die Zielsetzung der Studie in ihrem Kontext, das Evaluationsdesign mit seinen erwartbaren Ergebnissen, Schlussfolgerungen aus der Evaluationsforschung zur Evaluationsnutzung, und die Anwendungsprinzipien der Evaluationsstandards. Der Autor möchte an dieser Stelle den beiden Experten nochmals herzlich für ihr Engagement danken. Zentrale Ergebnisse des Experten-Workshops waren:

- Die DeGEval- und SEVAL-Standards stellen ein probates und erprobtes Instrumentarium zur Diskussion von Vor- und Nachteilen von Evaluationen dar. Die JC-Standards sind als Interpretationshintergrund für die DeGEval-Standards gültig.
- Grundsätzlich ist eine möglichst holistische Analyse anzustreben. Dies ergibt sich auch aus der Verflechtung von Einzelstandards, die in den JC-Standards an verschiedenen Stellen explizit vorliegt. Diese Analyse ist jedoch - wie jede Evaluation gemäß den Standards - in verschiedener Hinsicht gegenüber Machbarkeit, Evaluationskontext, und weiteren Faktoren abzuwägen. Es kann davon ausgegangen werden, dass mit dem gewählten Standard-Set die wesentlichen Parameter für die Fragestellung der Nützlichkeit adressiert werden, auch wenn eine noch umfassendere Analyse grundsätzlich denkbar gewesen wäre.
- Die möglichen Vorgehensweisen zur Erhellung von tatsächlicher Nutzenentststehung aus mit anerkannten Nützlichkeitsmerkmalen mehr oder weniger ausgestatteten Evaluationsprodukten und -prozessen siedeln sich im Spannungsfeld zwischen eher mikrologischen und vom Einzugsbereich her dann auch zwangsläufig eingegrenzten Fallstudien und einer gesamthaften Zugangsweise zur Evaluationspraxis durch entsprechend gestaltete Interviews und Surveys an. Eine gesamthafte Zugangsweise ist plausibel und stellt in Entsprechung zum formulierten Evaluationszweck bewusst die Möglichkeit der Erzielung übergreifend-gesamthafter Erkenntnisse über eine Forschungslogik, die eine Verfeinerung der Analytik zur Nutzenentstehung mit einer Beschneidung des Einzugsbereichs der Analyse zu bezahlen hätte.

Innerhalb der Feindefinitionsphase ersuchte der Auftraggeber RFTE auch die in der Plattform fteval organisierten AuftraggeberInnen und präsumtiven NutzerInnen von Programmevaluationen um Bekanntgabe etwaiger nicht auf der Homepage der Plattform fteval publizierter Berichte und bot ihnen auf Basis der institutionellen Verfasstheit des RFTE die Möglichkeit, zum Grundkonzept der geplanten Studie Stellung zu nehmen. In beiderlei Hinsicht wurden keine Hinweise erhalten, die eine Anpassung des Vorgehens angelegen erscheinen lassen hätten. Die in der Plattform fteval organisierten EvaluatorInnen wurden in einem getrennten Schritt durch den Metaevaluator über die Zwecke und Zielsetzungen der Studie informiert.

Im Folgenden wird auf die drei bereits in den Abbildungen 4 und 5 genannten Methoden zur Datenerhebung und –analyse eingegangen, auf die sich die Studie zur explorativen Auseinandersetzung mit Nützlichkeit und Nutzen der Programmevaluationen stützt.

Berichtsanalyse

Um angesichts der großen Anzahl von Programmevaluationen, die in Österreich durchgeführt wurden, einen summativen Zugriff im Kontext begrenzter Ressourcen zu ermöglichen, wurde eine Stichprobe von 20 Evaluationsberichten aus dem zwölfjährigen Zeitraum gezogen, über den die Evaluationspraxis - auch im Hinblick auf ihre Weiterentwicklung - betrachtet wird. Die Vorgehensweise zur Auswahl der zu analysierenden Evaluationen bzw. Evaluationsberichte, die im folgenden dargestellt ist, wurde mit dem Auftraggeber in allen Schritten abgestimmt.

Die Grundgesamtheit aller in Frage kommenden Berichte umfasst alle Evaluationsberichte, die FTI-Programme auf Bundesebene betreffen, von professionellen EvaluatorInnen bzw. Evaluationsinstituten erstellt wurden, ausdrücklich als Programmevaluationen bezeichnet sind, im Zeitraum 1.1.2003 - 30.9.2014 abgeschlossen wurden, und publiziert vorliegen. Als Datenbasis diente

das Online-Archiv der *Plattform fteval* (http://www.fteval.at/de/evaluation_studies). Diese Grundlage wurde in einem ersten Schritt um Berichte, die der genannten Definition nicht entsprechen, bereinigt und in einem zweiten Schritt ergänzt.[4] Durch diese Schritte wurde eine Datengrundlage von 46 Programmevaluationen, die den Definitionen der Studie entsprechen, erzielt. Die Stichprobe von 20 Evaluationsberichten, die im weiteren Verlauf der vorliegenden Studie analysiert wurden, umfassen nahezu die Hälfte der im Beobachtungszeitraum publizierten Programmevaluationen (43,5%). Die Liste der in der Metaevaluation analysierten Evaluationsberichte wird in Anhang 1 präsentiert.

Die Auswahl der zu analysierenden Evaluationsberichte erfolgte im Zufallsverfahren auf Basis theoretischer Kriterien (*theoretical sampling*). Leitend war dabei die Grundentscheidung, den zwölfjährigen Beobachtungszeitraum der Studie einerseits durchgehend abzubilden, dabei aber andererseits im Interesse der Aktualität der Ergebnisse Programmevaluationen jüngeren Datums stärker zu berücksichtigen. Zugleich sollten bei der Stichprobenziehung Strukturmerkmale der Evaluationspraxis, die sich aus den Auftraggeberschaften und aus den Tätigkeiten unterschiedlicher EvaluatorInnen ergeben, Berücksichtigung finden. Der Stichprobenziehung wurden somit folgende Parameter zugrunde gelegt:

- Der Beobachtungszeitraum als primäres Schichtungsmerkmal wurde in drei 4-Jahres-Abschnitte gegliedert. Diese Zeitabschnitte wurden sodann zugunsten einer genaueren Abbildung des Evaluationsgeschehens der jüngsten Jahre gewichtet. In der Stichprobe wird der Zeitabschnitt 1.1.2003-31.12.2006 mit 25% der zu analysierenden Berichte, der Zeitabschnitt 1.1.2007-31.12.2010 mit ebenfalls 25% der zu analysierenden Berichte, und der Zeitabschnitt 1.1.2011-30.9.2014 mit 50% der zu analysierenden Berichte abgebildet.
- Hinsichtlich der Auftraggeber der Evaluationen im Beobachtungszeitraum wurden vier Gruppen gebildet: BMVIT bzw. dessen Vorgängerorganisationen, BMWF bzw. dessen Vorgängerorganisationen, BMWFJ bzw. dessen Vorgängerorganisationen, sowie die Agenturen AWS, FFG und FWF, die als Auftraggeber von Programmevaluationen auf Bundesebene auftreten.[5] Auf Grund der geringeren Anzahl an verfügbaren Evaluationsberichten aus Evaluationen, die im Auftrag der drei Agenturen erstellt wurden, wurden diese zu einer Gruppe zusammengefasst. Programmevaluationen zu Tätigkeitsbereichen der Agenturen wurden im Beobachtungszeitraum sowohl von diesen selbst initiiert als auch von außen durch jeweils zuständige Ressorts beauftragt. Innerhalb der gewählten Sampling-Strategie wurde diesbezüglich der formalen Auftraggeberschaft Vorrang gegeben. Evaluationen, die von den Agenturen betreute Programme betreffen, befinden sich daher wegen der unterschiedlich gestalteten Beauftragungsverhältnisse im Governance-System sowohl in der Gruppe der Agenturen selbst als auch bei Ministerien.
- Evaluationsinstitute wurden so erfasst, dass wesentliche Gruppen entsprechend ihrer faktischen Rolle im betrachteten Evaluationsgeschehen über 12 Jahre hinweg abgebildet sind. Verschiedenste Evaluationsinstitute aus dem In- und Ausland sind in der Grundgesamtheit mit ein bis maximal vier Evaluationen über den gesamten Beobachtungszeitraum hinweg vertreten. Während eine genauere Abbildung der seltener tätigen Evaluationsinstitute zu einer kleinteiligen Struktur geführt hätte, innerhalb derer sich Entscheidungen über Berücksichtigung oder Nicht-Berücksichtigung in der Sampling-Struktur nur schlecht rechtfertigen lassen, wurde ein Anteil von mindestens 15% an allen Evaluationstätigkeiten als strukturprägend erachtet.

Nähere Angaben zum Sampling-Verfahren befinden sich in Anhang 3 des vorliegenden Berichts.

---

[4] Es wurden andere Segmente des Online-Archivs hinsichtlich etwaiger enthaltener Programmevaluationen gesichtet, der Datenbestand anhand der Forschungs- und Technologieberichte überprüft, und Informationen des Auftraggebers einbezogen. Wo nur Zwischenberichte zu Evaluationen vorlagen, wurden eigene Recherchen zur Auffindung von Endberichten unternommen. Dadurch konnte die Datengrundlage um drei anderweitig publizierte Berichte zu Evaluationen, zu denen zumindest ein Teilbericht im Online-Archiv der Plattform fteval vorhanden ist, ergänzt werden.

[5] 2014 wurden BMWF und BMWFJ zum BMWFW zusammengelegt. Da die Studie retrospektiv angelegt ist, liegen die Beauftragungen aller Evaluationen, die von ihr betrachtet werden können, vor dem Zeitpunkt dieser Umstrukturierung. Die während des Beobachtungszeitraums gegebene Strukturierung der Auftraggeber-Landschaft wurde daher berücksichtigt.

Im Anschluss an den Expertenworkshop wurde auf der Grundlage von Überlegungen sowohl konzeptueller als auch pragmatischer Art ein Set an DeGEval-Standards endgültig bestimmt, das die Kriterien für die Berichtsanalyse bildet. Neben den dezidiert als Nützlichkeitsstandards gekennzeichneten Standards interessieren auch Standards aus anderen Gruppen, die im verflochtenen, mit Querbezügen ausgestatteten Charakter aller 25 DeGEval-Standards einen übergeordneten Status haben. Methodische Robustheit ist unhinterfragt die grundlegende Basis einer guten und für eine Nutzung tauglichen Evaluation. Jedoch weisen die Standards klar darauf hin, dass weitere Aspekte des Evaluationsprozesses ebenfalls unverzichtbar sind, um Evaluationen Tauglichkeit und Güte zuzusprechen. Es sollte nicht durch zu starkes Fokussieren auf Methoden-Aspekte der Blick auf grundlegende Fragestellungen verstellt werden, wie etwa die Frage ob denn überhaupt wesentliche Information gesammelt und analysiert wurde (Cronbach 1984: 406f). Dieser Blickwinkel findet sich wiederum gerade in Nützlichkeits-Standards. [6] Wiederholt wurde darauf hingewiesen, dass methodische Genauigkeit als solche andere Aspekte der Nützlichkeit nicht ersetzen kann (so etwa Rossi/Freeman/Lipsey 1999, Beywl et. al. 2004). Es liegen aber auch Hinweise aus der Nutzungsforschung vor, dass methodische Genauigkeit nur eine untergeordnete Rolle in der tatsächlichen Nutzung von Evaluationen spielt (Fleischer/Christie 2009). Bisherige Anwendungen der Standards haben darüber hinaus gezeigt, dass die Schwierigkeiten, methodische Qualität im Sinn der Genauigkeitsstandards ex-post einzuschätzen, erheblich sind und selbst bei hohem Ressourceneinsatz nicht abschließend bewältigt werden können (vgl. Widmer 1996 ebenso wie Cooksy/Caracelli 2005). Es wurde diesbezüglich vorgeschlagen, eher auf Gesichtspunkte wie insbesondere die Transparenz der methodischen Berichterstattung Bezug zu nehmen. Die vorliegende Studie folgt diesem Vorschlag.

Das ausgewählte Set setzte sich zunächst aus 10 DeGEval-Standards und drei deskriptiven Kriterien zur allgemeinen Charakterisierung der Programmevaluationen zusammen. Dieses Standardset wurde in einem qualitativen Auswertungsverfahren zur Anwendung gebracht, das rekursiv zwischen den Prinzipien und Hinweisen der herangezogenen Standards und den einzelnen Evaluationsberichten hin und her ging, um auch etwaige zusätzliche Anforderungen erkennen zu können. Während der Durchführung der Evaluation wurde das in der Berichtsanalyse angewendete Standardset auf 11 Standards erweitert, um in der Auseinandersetzung mit dem Material erkannten Bedarfslagen noch besser gerecht zu werden.

Die folgenden Kriterien gelangen in der Analyse der ausgewählten Evaluationsberichte zum Einsatz:

Im Anschluss an die deskriptiven Kriterien 1 – 3 werden die entsprechenden DeGEval-Standards mit Nummer, Titel und Standardformulierung wiedergegeben. An zwei Stellen war zunächst angedacht, die jeweiligen Standards auch mit in einer weiterführenden strukturierten Analyse zu verbinden. Im Zuge der Durchführung der Berichtsanalyse erwies sich jedoch, dass diese Schritte auf Grund der Charakteristika der herangezogenen Evaluationsberichte nicht durchführbar waren und somit unterbleiben mussten.

1. **Evaluationstyp:** Eine grundlegende Unterscheidung im Bezug darauf, wann eine Evaluation ihren Evaluationsgegenstand betrachtet, was bis zu einem gewissen Grad auch ihre möglichen Zielsetzungen und Fragestellungen bedingt. Evaluationen können vor Beginn einer Intervention (Ex-ante-Evaluation) stattfinden, während der Durchführung (Zwischenevaluation), nach deren Abschluss (Ex-post-Evaluation), oder während aller Phasen (begleitende Evaluation).

2. **Evaluationsrolle:** Die älteste und gebräuchlichste Klassifikation dafür, wie eine Evaluation angelegt ist und was sie intendiert, ist diejenige zwischen formativer und summativer Evaluation.

[6] Einen Standard für „die richtige Methode“ auf der Ebene von Datengewinnungs- und Auswertungsverfahen wie Befragungen, Fokusgruppen, ökonometrischen Analysen etc. gibt es dabei nicht. Die Idee, dass sozialwissenschaftliche oder ökonometrische Einzelmethoden als solche sinnvoller oder weniger sinnvoll bzw. nützlicher oder weniger nützlich sein könnten, ohne dass ihr Einsatz im Rahmen eines umfangreicheren Settings der Evaluationsanalage reflektiert würde, ist den Standards fremd. Sie befassen sich mit übergeordneten methodologischen Kriterien wie Validität, Reliabilität, und konzeptueller Konsistenz. Die Umgangsweise mit Methodik manifestiert sich in den Standards analog zur evaluationstheoretischen Auffassung: „Not all methods are equally good for all tasks. So it is folly to prescribe one method for all evaluations, and evaluation theory must sort out the relative strengths and weaknesses of different methods for specific tasks.“ (Shaddish/Cook/Leviton 1991:44).

Von einer summativen Evaluation wird gesprochen, wenn eine abschließende Beurteilung des Evaluationsgegenstands erzielt und Entscheidungen zum Evaluationsgegenstand ermöglicht werden sollen. Formative Evaluation zielt darauf ab, die Gestaltung eines Evaluationsgegenstands zu begleiten, um Verbesserungen zu ermöglichen. Dies wird oft mit Programmphasen verbunden, wobei formative Evaluation frühzeitig Programmkonzepte testet und summative Evaluation nach längerer Laufzeit bzw. in einer Phase der Routineanwendung zum Einsatz gelangt. Dieses Verständnis hat sich inzwischen dahingehend erweitert, dass Evaluationen zugleich formativ und summativ sein können. Setzen sich formative Evaluationen vorrangig mit dem Operieren eines Programms auseinander, so erscheint dabei auch eine Einbeziehung der Wirkungsebene sinnvoll.

3. **Evaluationsschwerpunkt:** Evaluationen können unterschiedliche grundlegende Aspekte ihres Evaluationsgegenstands betrachten, und dabei ihre Analysen auch in unterschiedlicher Breite anlegen. Herangezogen wird hier die international anerkannte Gliederung nach OECD DAC (2010), die etwa auch in Leitmaterialien der Europäischen Kommission für die Evaluation der Struktur- und Regionalentwicklung sowie der Entwicklungszusammenarbeit direkte Entsprechungen findet: Relevanz (*relevance*) - Effektivität (*effectiveness*) - Effizienz (*efficiency*) - Wirkung (*impact*) - Nachhaltigkeit (*sustainability*).

4. **N1 Identifizierung der Beteiligten und Betroffenen**
   Die am Evaluationsgegenstand beteiligten oder von ihm betroffenen Personen bzw. Personengruppen sollen identifiziert werden, damit deren Interessen geklärt und so weit wie möglich bei der Anlage der Evaluation berücksichtigt werden können.

5. **N2 Klärung der Evaluationszwecke**
   Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen können und das Evaluationsteam einen klaren Arbeitsauftrag verfolgen kann.

6. **N4 Auswahl und Umfang der Informationen**
   Auswahl und Umfang der erfassten Informationen sollen die Behandlung der zu untersuchenden Fragestellungen zum Evaluationsgegenstand ermöglichen und gleichzeitig den Informationsbedarf des Auftraggebers und anderer Adressaten und Adressatinnen berücksichtigen.

7. **N5 Transparenz von Werten**
   Die Perspektiven und Annahmen der Beteiligten und Betroffenen, auf denen die Evaluation und die Interpretation der Ergebnisse beruhen, sollen so beschrieben werden, dass die Grundlagen der Bewertungen klar ersichtlich sind.

8. **N6 Vollständigkeit und Klarheit der Berichterstattung**
   Evaluationsberichte sollen alle wesentlichen Informationen zur Verfügung stellen, leicht zu verstehen und nachvollziehbar sein.

9. **N8 Nutzung und Nutzen der Evaluation**
   Planung, Durchführung und Berichterstattung einer Evaluation sollen die Beteiligten und Betroffenen dazu ermuntern, die Evaluation aufmerksam zur Kenntnis zu nehmen und ihre Ergebnisse zu nutzen.

10. **F3 Vollständige und faire Überprüfung**
    Evaluationen sollen die Stärken und die Schwächen des Evaluationsgegenstandes möglichst vollständig und fair überprüfen und darstellen, so dass die Stärken weiter ausgebaut und die Schwachpunkte behandelt werden können.

11. **F5 Offenlegung der Ergebnisse**
    Die Evaluationsergebnisse sollen allen Beteiligten und Betroffenen soweit wie möglich zugänglich gemacht werden.

12. **G3 Beschreibung von Zwecken und Vorgehen**
    Gegenstand, Zwecke, Fragestellungen und Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen genau dokumentiert und beschrieben werden, so dass sie identifiziert und eingeschätzt werden können.

13. G8 Begründete Schlussfolgerungen

Die in einer Evaluation gezogenen Folgerungen sollen ausdrücklich begründet werden, damit die Adressaten und Adressatinnen diese einschätzen können.

In Anhang 4 werden die herangezogenen DeGEval-Standards vollständig wiedergegeben. Es werden dabei auch die korrespondierenden, zur genaueren Interpretation angeratenen bzw. notwendigen JC-Standard (Joint Committee & Sanders 2006) in Auszügen wiedergegeben, die im Lauf der durchgeführten Berichtsanalyse zu Klärungen beigetragen haben bzw. im Umgang mit den Berichten schlagend wurden.

Für jeden Evaluationsbericht wurde ein Factsheet im Umfang von ca. 4 Seiten erstellt,[7] das die Erfüllung jedes herangezogenen Standards durch eine Einstufung auf einer fünfstufigen Skala bezeichnet und durch einen qualitativen Kommentar näher darstellt. Die numerischen Einstufungen verstehen sich dabei als Erzeugung einer groben Übersicht, die auch in einem Gesamtbild Grundzüge und Entwicklungen leicht erkennen lässt. Den eigentlichen Kern der Berichtsanalyse bildet jedoch die qualitative Analyse, wie jede individuelle Evaluation bzw. der Bericht über sie in spezifischer Weise Empfehlungen und Forderungen der Standards besser oder schlechter entspricht. Diese qualitative Betrachtungsweise bildet die Basis für die Identifikation von Merkmalen, die die österreichiche Evaluationspraxis im FTI-Bereich im zwölfjährigen Beobachtungszeitraum gekennzeichnet haben und die sodann für Schlussfolgerungen und Empfehlungen genutzt wird. Das Template der Factsheets wird auf der folgenden Seite wiedergegeben.

Abbildung 6: Factsheet für die Berichtsanalyse

Evaluationsbericht Nr.	**Beobachtungszeitraum:**	
Klassifikation	**Einordnung**	
Evaluationsrolle und -typ		
Evaluationsschwerpunkte		
Standard	**Kommentar**	**Einstufung***
N1 Identifizierung der Beteiligten und Betroffenen		
N2 Klärung der Evaluationszweck		
N4 Auswahl und Umfang der Informationen		
N5 Transparenz von Werten		
N6 Vollständigkeit und Klarheit der Berichterstattung		
N8 Nutzung und Nutzen der Evaluation		
F3 Vollständige und faire Überprüfung		
F5 Offenlegung der Ergebnisse		
G2 Kontextanalyse		
G3 Beschreibung von Zwecken und Vorgehen		
G8 Begründete Schlussfolgerungen		

* 1. Standard sehr gut erfüllt, 2. Standard gut erfüllt, 3. Standard teilweise erfüllt, 4. Standard ansatzweise berücksichtigt, 5. Standard nicht erfüllt. 0. unzureichende Information

Die Gruppe der für die vorliegende Analyse besonders relevanten Nützlichkeitsstandards enthält auch zwei Standards, die in Berichtsanalysen nicht bzw. nur mit sehr geringer Aussicht auf eine Vorfindlichkeit relevanter Angaben verfolgt werden können. Es sind dies *N3 Glaubwürdigkeit und Kompetenz des Evaluators/der Evaluatorin* und *N7 Rechtzeitigkeit der Evaluation.* Diese Standards wurden der Berichtsanalyse nicht zugrunde gelegt, jedoch in den ergänzenden Erhebungsverfahren

[7] Das bedeutet nicht, dass Berichtsanalysen anhand der Standards nicht grundsätzlich noch ausführlicher durchgeführt werden könnten, um auf spezifische Herausforderungen einer bestimmten Evaluation gezielt einzugehen. Im Zusammenhang der hier beauftragten Studie stand die Erzeugung eines Überblicks unter Maßgabe der verfügbaren Ressourcen im Vordergrund.

(EvaluatorInnenbefragung, AuftraggeberInnen-Interviews) verfolgt, sodass die vorliegende Studie auch Aufschlüsse über deren Relevanz für die Entfaltung von Evaluationswirkungen erbringen kann.

Interviews mit Hauptakteuren in den auftraggebenden Institutionen

Um die Sichtweisen von AuftraggeberInnen von Programmevaluationen in Erfahrung zu bringen, die zugleich die HauptadressatInnen der Evaluationsberichte und intendierte NutzerInnen sind bzw. Hauptakteure im Evaluationsgeschehen darstellen, wurden halbstrukturierte Interviews mit 15 Personen geführt, wobei auf eine gleichmäßige Abdeckung der Ressorts und Agenturen und auch auf eine Entsprechung zur Struktur und Geschichte der *Plattform fteval* geachtet wurde.[8] Die Interviews zielten auf folgende Fragekomplexe:

- Informationen über den Nutzen durchgeführter FTI-Programmevaluationen und diesbezügliche Einflussfaktoren;
- Informationen über Merkmale der Evaluationsprozesse;
- Informationen über die Rolle von Programmevaluationen als Bestandteile eines übergreifenden Wissens- und Informationssystems der FTI-Governance.

InterviewpartnerInnen und Interviewleitfaden wurden mit dem Auftraggeber abgestimmt. Die Liste der InterviewpartnerInnen findet sich in Anhang 6 des vorliegenden Berichts, der Interviewleitfaden ist in Anhang 7 beigegeben. Die Gespräche wurden in Form von Einzelinterviews und teilweise auch von Gruppeninterviews mit zwei GesprächspartnerInnen geführt, wo die InterviewpartnerInnen bestimmter Institutionen dies wünschten. Die Gespräche fanden in einer offenen und interessierten Atmosphäre statt und bewegten sich im Umfang von ein bis zwei Stunden.

In den semi-strukturierten Interviews wurde auf große Offenheit gegenüber den Thematisierungsweisen der GesprächspartnerInnen wert gelegt, um im Sinne eines erkundenden Verfahrens möglichst gut virulente Aspekte der Evaluationspraxis zu erkennen. Naturgemäß ergibt sich aus einer solchen offenen Gesprächsführung, dass nicht in allen Gesprächen alle Punkte des Interviewleitfadens in gleicher Weise und so systematisch abgehandelt wurden, wie es nur in einer strukturierten Befragung möglich gewesen wäre. Dass detaillierte retrospektive Darstellungen zu den zahlreichen und oft länger zurückliegenden Evaluationsverfahren nur eingeschränkt erfolgen würden, stand zu erwarten und war der Herangehensweise des Interviewleitfadens mit seinen übergreifenden Fragestellungen auch zugrunde gelegt. Indem die InterviewpartnerInnen ihre jeweiligen Erfahrungen stark unter Bezugnahme auf die aktuelle Situation darlegten, waren die Gespräche vor allem hinsichtlich übergreifender Charakteristika der Evaluationsprojekte in der heutigen Wahrnehmung sowie aktueller Bedarfslagen und Kontextfaktoren informativ. Die Gesprächsinhalte wurden nach Standards der qualitativen Sozialforschung anonymisiert ausgewertet. Freigaben zur Zitation wurden im Rahmen der meisten Gespräche erhalten. Wo InterviewpartnerInnen es wünschten, wurden wörtliche Zitate vor der Berichtsabfassung zur Abstimmung übermittelt. Vollständige Transkripte der Gespräche waren nicht vorgesehen und hätten deutlich höhere Projektressourcen erfordert. Bei der Wiedergabe von Zitaten erfolgt die Identifikation des betreffenden Interviews in Form der Angabe des Institutionentyps und einer fortlaufenden Nummerierung (z.B. A1, M2).

Online-Befragung von FTI-EvaluatorInnen

Zur Ergänzung der Informationslage über Evaluationsprozesse und zur Gewinnung von Informationen über tatsächliche Nutzungen von Programmevaluationen wurde eine Online-Befragung unter FTI-EvaluatorInnen durchgeführt. Die Entscheidung für die Breitenbefragung wurde getroffen, um zu einem möglichst umfassenden Bild einer in sich differenzierten Evaluationspraxis mit zahlreichen unterschiedlich konfigurierten Einzelprojekten zu gelangen. Ausschlaggebend war des Weiteren die Schwierigkeit, aus eine Vielzahl von EvaluatorInnen einige wenige InterviewpartnerInnen auszuwählen, die dann doch nur für denjenigen Ausschnitt der Evaluationspraxis sprechen hätten können, an dem sie jeweils beteiligt waren. Die den EvaluatorInnen gestellten Fragen umfassten drei Themenblöcke:

[8] Die Anzahl der tatsächlich geführten Interviews überschreitet die ursprünglich vorgesehene Zahl von 8-10 Interviews, da eine Einbeziehung der Erfahrungen und Sichtweisen auf eine größere Zahl von Akteuren im FTI-Governancesystem im Voranschreiten der Studie unter iterativer Auseinandersetzung mit ihren verschiedenen Informationsquellen als wichtig erachtet wurde.

- Informationen über den Nutzen durchgeführter FTI-Programmevaluationen anhand der in der internationalen Forschung über Evaluationsnutzen etablierten Kategorien, wobei auch Einflussfaktoren auf eine Nutzenentstehung erfasst werden, die so von AuftraggeberInnen nicht erfragt werden können;
- Informationen über Merkmale der Evaluationsprozesse, die in den herangezogenen Standards angesprochen sind, jedoch den Evaluationsberichten nicht entnommen werden können;
- Hintergrundinformationen über die antwortenden EvaluatorInnen, die die Einschätzung der Belastbarkeit der erhaltenen Angaben erlauben.

Bei der Konstruktion des Fragebogens wurde auf hohe Messgenauigkeit wert gelegt, da aus der internationalen Evaluationsforschung auch Hinweise vorliegen, dass angesichts einer nicht genau festgelegten Fachterminologie selbst von professionellen EvaluatorInnen Fragen zu ähnlichen Themen nicht immer genau und in gleicher Weise verstanden wurden, sodass die Ergebnisse einigen Interpretationsspielraum offen lassen. Zur Abbildung von Nutzen und Einflussfaktoren auf dessen Entstehung wurden die in Kapitel 1.2.5 dargestellten Klassifikationen eingesetzt und im spezifischen Bezug auf eine oft als methodenorientiert bezeichnete Evaluationspraxis um einige Items ergänzt. Hinsichtlich der Evaluationsprozesse wurden Inhalte der DeGEval-Standards und der Joint-Committee-Standards, die deren Interpretationshintergrund bilden, unter weitestgehender Beibehaltung der Originalterminologie operationalisiert. Der Fragebogen wurde sodann von einer hoch qualifizierten Evaluatorin, die außerhalb des FTI-Bereichs arbeitet, getestet. Da der Fragebogen auf Basis von Konzepten der Fachliteratur und der Standards mit dortigen Originalformulierungen konstruiert wurde, wurde hier, abgesehen von Vorabklärungen von Aufmerksamkeitspunkten, auf eine Abstimmung des Instruments mit dem Auftraggeber konsensual verzichtet.

Befragt wurden EvaluatorInnen, die im Beobachtungszeitraum an publizierten Programmevaluationen mitgearbeitet hatten oder an Instituten arbeiten, die im Bereich der Politikberatung und Evaluation tätig sind, sodass sie auch für eine Durchführung von unpublizierten Evaluationen in Frage kamen. Die Befragung wurde mit der professionellen Umfragesoftware *SurveyMonkey* im Sommer 2015 durchgeführt. 273 EvaluatorInnen und MitarbeiterInnen von relevanten Instituten in Österreich und im deutsch- und englischsprachigen Ausland wurden kontaktiert. Die Umfrage war 8 Wochen im Feld, drei Erinnerungen wurden versandt. Zusätzlich wurde auf die Umfrage durch die Geschäftsführung der *Plattform fteval* aufmerksam gemacht, wofür an dieser Stelle nochmals ausdrücklich gedankt sei.

49 EvaluatorInnen und Evaluatoren haben den Fragebogen aufgerufen. Allerdings liegen nicht von allen diesen antwortenden Evaluatorinnen auch Antworten vor, was auch damit zu tun haben kann, dass Beantwortungen nicht mit dem notwendigen Speicherbefehl abgeschlossen wurden, auf den im Anschreiben allerdings deutlich hingewiesen wurde.

Antworten liegen von 37 EvaluatorInnen liegen vor. 44,9% der Antwortenden sind häufig tätige EvaluatorInnen, die mindestens vier Programmevaluationen im Beobachtungszeitraum durchgeführt haben. Die Auskünfte über die Evaluationspraxis beziehen sich bei 73,5% der antwortenden EvaluatorInnen auf mehr als eine FTI-Programmevaluation. 85% führen seit 7 Jahren oder noch länger Evaluationen durch, die übrigen 15% sind seit mindestens 5 Jahren mit der Durchführung von Evaluationen befasst. Es kann somit davon ausgegangen werden, dass der Kern der faktisch relativ kleinen Gruppe von Hauptakteuren erfasst wurde und die erhaltenen Umfragedaten ein gut belastbares Bild der österreichischen FTI-Evaluationspraxis liefern.

In der Befragung wurde eine Vielzahl von potenziell wichtigen Aspekten und Faktoren im Sinne eines erkundenden Verfahrens abgefragt. In der Interpretation der Umfragedaten wird auf diejenigen Ergebnisse Bezug genommen, die als die Wesentlichsten erkannt wurden. Alle Umfrageergebnisse können dem Anhang 8 entnommen werden, wo sie in der Struktur des eingesetzten Fragebogens mit seinen Originalformulierungen wiedergegeben werden.

Allen InterviewpartnerInnen und EvaluatorInnen, die die Umfrage beantwortet haben, sei an dieser Stelle nochmals ausdrücklich für Ihre Auskunftsbereitschaft und für die Zeit, die sie sich genommen haben, gedankt.

Hinweise zur Textierung des vorliegenden Berichts

Der vorliegende Bericht setzt geschlechtsneutrale Formulierungsweisen ein. Es wird - in Übereinstimmung mit den Standards und der internationalen Fachliteratur - durchgehend der Begriff „Evaluation“ verwendet. Bedeutungsunterschiede zum in der österreichischen Community üblichen Begriff „Evaluierung“ sind damit nicht impliziert. In der Auswertung der Interviews werden hingegen

stets diejenigen Begriffe originalgetreu wiedergegeben, die die jeweiligen InterviewpartnerInnen selbst benutzt haben.

Da die DeGeval-Standards ihren historischen und sachlichen Hintergrund in den Program Evaluation Standards (Joint Committee/Sanders 2006 [1994]) haben, der zu ihrer Interpretation nicht nur legitim, sondern – wie die Erfahrungen in der Durchführung der vorliegenden Metaevaluation zeigen – auch im Umgang mit konkreten Evaluationen immer wieder notwendig ist, verfließen die beiden Bezugspunkte tendenziell miteinander. Im vorliegenden Bericht wird der gesamthaft begriffene sachliche Komplex der Standard-Inhalte bzw. –philosophien mit dem allgemeinen Begriff „Standard(s)" angesprochen. In den Factsheets wird im Sinne der Anonymisierung stets der Begriff „Programm" eingesetzt, auch wenn in der konkreten Originalbezeichnung der Maßnahme ein anderer Begriff verwendet wurde.

1.4 Limitierungen der Studie

Analysen zu FTI-Politiken und zu der auf sie bezogenen Evidenzproduktion, zu der Programmevaluationen jedenfalls zählen, sind vor allem von durch Politikwissenschaft und Ökonomie geprägten Analysestilen getragen. Im Unterschied dazu ist die vorliegende Studie in der Art ihrer Fragestellungen und in der Wahl ihrer Methoden evaluationstheoretisch verankert, was Differenzen zu in der FTI-Fachliteratur eingespielten Betrachtungswinkeln mit sich bringt und anders gelagerte Blickwinkel eröffnet. Sie verdankt sich der Bezugnahme auf einen spezifisch evaluationsbezogenen Wissensbestand, der übergreifend im Bezug auf unterschiedliche Einsatzfelder von Evaluation entstanden ist und in seiner Entstehungsgeschichte vor das Einsetzen der Evaluationstätigkeit zu europäischen FTI-Politiken zurückreicht. Die vorliegende Studie findet ihre Bezugspunkte in einem Fachwissen zu Evaluation, das vor allem in den USA akademisch-institutionell verankert ist, über eine eigene Landschaft an Fachpublikationen verfügt, sowie von mehreren Fachgesellschaften für Evaluation getragen wird (*evaluation science*). In Europa ist eine solche Verankerung von evaluationstheoretischem und –methodologischem Fachwissen bis heute weit weniger anzutreffen, wenn hier auch Evaluationsgesellschaften entstanden sind und in manchen Ländern einige akademische Positionen geschaffen und Fachjournale ins Leben gerufen wurden.

Metaevaluation stellt das spezifische Instrument dar, das die *Evaluation Science* für Zwecke einer Analyse von Qualitätsdimensionen einer oder mehrerer Evaluationen hervorgebracht hat und das für die Diskussion eines örtlich und zeitlich umrissenen Praxisfelds von Evaluation geeignet ist. In der vorliegenden Studie wird dieses Instrument erstmals zur Erhellung der FTI-Evaluationspraxis angewendet, die in Österreich unter konkreten Rahmenbedingungen entstanden ist, und um Aspekte der Nutzungsforschung zur Evaluation ergänzt. Ähnliche, wenn auch nicht völlig analog konzeptualisierte Metaevaluationen anhand der Evaluationsstandards und empirische Arbeiten zur Nutzungsforschung wurden bislang im Zuge des Aufbaus der dortigen Evaluationskultur in der Schweiz durchgeführt. Überschneidungspunkte einer solchen in Evaluationstheorie und -methodologie zentrierten Diskussion mit den für den FTI-Bereich tonangebenden Thematisierungsweisen von FTI-spezifischen Politikanalysen bieten sich durchaus an, und sie werden in der vorliegenden Arbeit auch in der Deskription des untersuchten Einsatzfeldes von Evaluation (Kapitel 1.2.3) und in ihren Schlussfolgerungen (Kapitel 6) aufgegriffen.

Parallel zu der von der Evaluationsforschung entwickelten Analytik von nutzungsrelevanten Faktoren, in einem „Kreislauf der Ideen" mit evaluationstheoretischen Arbeiten und den Evaluationsstandards steht, hat sich auch eine Zugangsweise entwickelt, die von institutionellen Faktoren und Merkmalen des politisch-administrativen Systems ihren Ausgang nimmt. Dieser politikwissenschaftliche Untersuchungstyp ist bestrebt, unter Bezug auf die Verfasstheit der politisch-administrativen Akteure, auf Eigenschaften der politischen Arena des Agenda-Setting bzw. der Aushandlung, oder anhand von Typologisierungen von Institutionen den Umgang mit evaluativer und andersartiger Information im politisch-administrativen System zu beleuchten (so etwa Bovens et. al. 2006, Leeuw 2006, Hannsen 2006, Balthasar 2007, Leeuw/Rist/Sonnichsen 2000, Hertting & Vedung 2012, Biegelbauer 2013). Im Unterschied zu diesen Ansätzen wählt die vorliegende Analyse einen Zugang, der sich detailliert mit Faktoren befasst, die in einer spezifischen evaluationstheoretischen Forschungstradition als relevant für Nützlichkeit und Nutzen von Programmevaluationen gelten. Sie ist dabei auch an Fragen der Entstehung von evaluativen Wissensströmen und an kumulativen Wirkungen des Evaluierens („streams", Rist/Stame) interessiert. Sie ist jedoch von ihrer Anlage her keine Studie zu einer vergleichenden Typologie von Evaluationssystemen (vgl. Leeuw/Furubo 2008), zu Evaluationssteuerung in demokratischen Governancesystemen (vgl. Hanberger 2013,

Vedung/Hansen/Kettunen_2012), oder zu Evaluation im Gefüge unterschiedlicher Formen von Politikanalyse (vgl. Bovens/'t Hart/Kuipers 2006). Mit den erbrachten Daten liegen zweifellos Möglichkeiten vor, sich in zukünftigen weiteren Schritten anhand solcher Konzepte mit der österreichischen FTI-Evaluationspraxis weiter auseinanderzusetzen, auch in komparativer Weise. Dies war jedoch nicht Bestandteil des vorliegenden Evaluationsauftrags und hätte den Rahmen der vorliegenden Untersuchung jedenfalls gesprengt.

Die Studie behandelt ausschließlich Programmevaluationen, die von ihren AuftraggeberInnen unter diesem expliziten Titel geplant und durchgeführt wurden. Sie untersucht keine Institutionenevaluationen oder Reviews, und somit nicht das gesamte Spektrum der evaluativen Wissensproduktion und strategischen Politikberatung, auf das sich das FTI-politische Governancesystem mit seinen verschiedenen, aufgefächerten Handlungsformen im Beobachtungszeitraum gestützt hat. Die Studie wirft jedoch ausdrücklich in den Auftraggeber-Interviews Fragen nach Rolle und Stellenwert der Programmevaluationen in einem breiteren Wissenssystem der FTI-politische Governance auf und erbringt auf dieser Basis einige Ergebnisse, die Programmevaluationen innerhalb dieses nicht vollständig ausgeleuchteten Systems positionieren.

Alle Entscheidungen zur Vorgehensweise und Methodik der vorliegenden Studie wurden an der Ermöglichung eines gesamthaften Zugriffs auf eine mehrjährige und aus zahlreichen Einzelfällen konstituierte Evaluationspraxis ausgerichtet. Diese Vorgehensweise impliziert Abstriche gegenüber einer eingehenden Analytik der Nutzenentstehung, die in einem alternativen Design mit Fallstudien grundsätzlich ebenfalls möglich gewesen wäre. Angesichts der verfügbaren Ressourcen hätten freilich nur drei bis vier Fallstudien durchgeführt werden können, was kaum einen umfassenden Blick auf die vielschichtige Evaluationspraxis mit unterschiedlich konzipierten und mit unterschiedlichen Ressourcen ausgestatteten Evaluationsprojekten zu unterschiedlichen Programmtypen eröffnet hätte. Fallstudien hätten auch nur zu rezenten Programmevaluationen durchgeführt werden können, die den beteiligten und betroffenen Akteuren noch ausreichend in Erinnerung sind, nicht jedoch zu länger zurückliegenden Programmevaluationen, die im zugrunde gelegten Konzept ebenso interessiert haben. Die vorliegende Analyse von Nutzungsweisen ist von ihrer methodischen Anlage her eher an Nutzungen im durch die Evaluationen direkt adressierten Auftraggeberbereich (*intended use by intended users*) und einer als klassisch zu bezeichnenden Auffassung von Nutzungsweisen von Evaluation orientiert, als an einer mikrologischen Ausleuchtung im Gefolge des Konzepts des Evaluationseinflusses (*evaluation influence*). Die Untersuchung tendiert damit auch dazu, Nutzen aus den Programmevaluationen insgesamt zu unterschätzen, der an nicht ausgeleuchteten Stellen des FTI-politischen Systems und in der FTI-Akteurslandschaft entstanden sein kann. Des Weiteren befasst sie sich nicht mit Formen einer mißbräuchlichen Nutzung (z.B. von aus dem Zusammenhang gerissenen und gegenüber dem Evaluationsbericht uminterpretierten Daten), deren Erfassung vor großen grundsätzlichen Herausforderungen steht und deshalb auch in der internationalen Nutzungsforschung ein offenes Desiderat bleibt.

Die vorliegende Studie verfährt in der Thematisierung von Nützlichkeit und Nutzen der FTI-Programmevaluationen auf der Basis eingeführter Konzepte in international abgestützten und vergleichbaren Kategorien. In der im Evaluationsauftrag verankerten Grundkonzeptualisierung war die Berichtsanalyse anhand der Evaluationsstandards als Kern der Untersuchung perspektiviert, um Verbesserungspotenzial in der Planung und Durchführung von FTI-Programmevaluationen identifizieren zu können, das von den handelnden Akteuren im Feld in Zukunft aufgegriffen werden kann. Die ergänzenden Erhebungsverfahren wurden eingesetzt, um das Verständnis der Evaluationspraxis anzureichern und der wesentlichen Frage nach dem Überstieg zwischen für Nützlichkeit als relevant zu erachtenden Evaluationsmerkmalen und der faktischen Entstehung von Nutzen nachgehen zu können. Mit dem gewählten Vorgehen erfolgt eine Kontextualisierung dessen, was die Standards als Gegenstand der Qualitätsreflexion in Evaluationsprojekten benennen, sodass anhand der Berichte einschätzbare Nutzungspotenziale mit der Art und Weise, unter welchen Bedingungen sich diese Potenziale ergeben und entfalten, relationierbar werden. Auf Grund der in der Durchführung der Untersuchung erhaltenen konkreten Daten haben Kontextfaktoren für die Planung und Nutzung von FTI-Programmevaluationen an Bedeutung für die Gesamteinschätzung der analysierten Evaluationspraxis gewonnen. Die Datenlage hat unter anderem die Abfassung eines eigenen Kapitels über strukturelle Herausforderungen im Auftraggeber-Bereich motiviert. Es war jedoch konzeptuell nicht vorgesehen, die durchgeführten Programmevaluationen ausschließlich an Auftraggeber-Interessen bzw. deren Verständnis von nützlichen Evaluationen zu bemessen.

Die Studie stützt sich bei der über die Berichtsanalyse hinausreichenden Beleuchtung von Evaluationsprozessen und Nutzungskontexten sowohl auf die Sichtweise von AuftraggeberInnen als auch auf die von EvaluatorInnen. Das Bild der Evaluationspraxis, das auf dieser Basis gezeichnet werden kann, bemisst sich freilich an den erhaltenen Daten. Für die Beleuchtung von Merkmalen der Planung, Durchführung und Verwertung von Programmevaluationen aus dem Blickwinkel von AuftraggeberInnen und von konkret beteiligten EvaluatorInnen wurden mit guten Gründen zwei unterschiedliche Datenerhebungsverfahren eingesetzt. In der strukturierten EvaluatorInnenbefragung wurden die Inhalte der herangezogenen Evaluationsstandards und der Nutzungsforschung operationalisiert und damit durchgängig systematische Daten erhalten. In den Interviews mit AuftraggeberInnen wurden entsprechende Leitfragen formuliert, aber Daten nur nach Maßgabe des Antwortverhaltens der jeweiligen GesprächspartnerInnen und somit in weniger systematischer Weise erhalten. Eine Rückführung der erhaltenen Aussagen in die Systematik, die in den Ergebnissen der EvaluatorInnenbefragung von vornherein sichergestellt ist, ist nur auf interpretativem Weg und in den Grenzen möglich, die sich aus dem Material selbst ergeben. Eine direkte Spiegelung der Sichtweisen von AuftraggeberInnen und EvaluatorInnen zu allen im explorativen Verfahren angelegten Analyseaspekten ist daher nicht möglich, und einzelne Abschnitte der angestellten Analyse der Evaluationspraxis können sich nur in unterschiedlichem Maß auf die Sichtweisen beider Seiten gleichzeitig stützen. Dies betrifft vor allem Auskünfte über genaue Merkmale von Evaluationsberichten und –prozessen, die retrospektiv eine faktische Nutzenentstehung beeinflusst haben, die von AuftraggeberInnen-Seite nur eingeschränkt erhalten werden konnten. Dieser Umstand kann auch im Zusammenhang mit dem in der Nutzungsforschung bekannten Phänomen gesehen werden, dass komplexere, längerfristige und ineinandergreifende Nutzungsweisen von Evaluation von den Beteiligen nur schlecht im Nachhinein einzelnen Evaluationen mit ihren jeweiligen Details zugeordnet werden können.

Die Metaevaluation mit ergänzenden Komponenten der Nutzungsforschung zielt auf eine problemzentrierte und im Sinne der Machbarkeit eingegrenzte Einschätzung des betreffenden Praxisfelds der Evaluation mit ihrem historisch-lokalen Entwicklungszustand. Sie reflektiert systematisch die Zugänge zu Programmevaluation, die mit Nützlichkeits- und Nutzungsfragen direkt in Zusammenhang stehen. Nicht angestrebt bzw. geleistet wurde eine umfassende Qualitätsklärung zu den in die Untersuchung einbezogenen Programmevaluationen, wie sie typischer Weise im Vorfeld einer Evaluationssynthese angestellt wird, um die Ausgangsbedingungen für die Nutzung der von den Evaluationen zur Verfügung gestellten Daten zu klären. Auf Grund der in der durchgeführten Analyse gemachten Erfahrungen ergibt sich gleichzeitig der für andere Untersuchungen interessante Hinweis, dass bereits die eingegrenzte Analyse anhand einiger Standards mit Hindernissen konfrontiert war und eine umfassende Qualitätsklärung ausschließlich anhand der Berichte kaum möglich erscheint.

Um die Gangbarkeit der Metaevaluation zu erhöhen, die weder durch die AuftraggeberInnen der untersuchten Programmevaluationen noch durch die beteiligten EvaluatorInnen, sondern durch Dritte beauftragt wurde, wurde sie in anonymisierender Form durchgeführt. Möglichkeiten der AuftraggeberInnen und EvaluatorInnen, sich mit der Einschätzung von in die Untersuchung einbezogenen Programmevaluationen anhand der DeGEval- und JC-Standards auseinanderzusetzen, sind dadurch eingeschränkt. Sollte Interesse an der Identifikation einer der im Anhang 1 gelisteten und in einem Factsheet in Anhang 2 beschriebenen Programmevaluation seitens der jeweiligen AuftraggeberInnen oder jeweils beteiligten EvaluatorInnen bestehen, so kann von ihnen das betreffende Factsheet vom Metaevaluator erfragt werden.

Eine abschließende Überlegung gilt einem möglichen Bias in den herangezogenen Daten. Die Berichtsanalyse basiert gezielt auf einem Samplingverfahren, das Verzerrungen in der Auswahl des betrachteten Ausschnitts aus allen Evaluationsberichten vermeiden soll. Die Auswahl erfolgt aus publizierten Evaluationsberichten. Da es eine bekannte Tatsache ist, dass nicht alle im Beobachtungszeitraum durchgeführten Programmevaluationen auch publiziert wurden, kann die zur Berichtsauswahl herangezogene Grundgesamtheit gerade Berichte nicht enthalten, die von ihren Auftraggebern nicht als zufriedenstellend eingestuft wurden. Es kann jedoch nicht pauschal davon ausgegangen werden, dass dabei immer Qualitätsmängel im engeren Sinn vorlagen, auch andere Gründe können für das Unterbleiben einer Publikation schlagend geworden sein. Sollte es sich bei der Grundgesamtheit der publizierten Berichte tatsächlich um eine Positivauswahl handeln, so würden in der vorliegenden Metaevaluation die besten Programmevaluationen analysiert, was durchaus aufschlussreich in Bezug auf den erreichten Stand und weitere Verbesserungsmöglichkeiten sein wird.

Bei Angaben aus Interviews und Befragung ist grundsätzlich denkbar, dass Angaben zu positiven Darstellungen tendieren, da die Antwortenden an der Materie ein Interesse haben. In der Befragung wurde dem durch genaue, teils komplexe Frageformulierungen in Originalformulierungen der Standards und der Fachliteratur gegengesteuert. Crosschecks zwischen den Antworten zu verschiedenen Fragen und Datenlagen aus unterschiedlichen Quellen wurden durchgeführt, und als wenig tragfähig eingeschätzte Ergebnisse werden nicht berichtet. Die Interviews verliefen durchwegs sehr offen und kritisch und waren von einem spürbaren Interesse an weiteren Verbesserungsmöglichkeiten getragen, was durch positiv verzerrte Darstellungsweisen wohl kaum ermöglicht wird. Angaben zum entstandenen Nutzen können zu einer Überschätzung tendieren, doch stellen die herangezogenen Datenquellen die relevantesten möglichen Informationsquellen dar, die in vergleichbaren Studien typisch herangezogen werden.

Der vorliegende Bericht präsentiert diejenigen Daten, die in einem triangulierenden Verfahren aus drei Datenquellen als die Wesentlichsten erkannt wurden. Mit den eingesetzten Datenerhebungsverfahren wurde eine umfangreiche Datenlage zu Nützlichkeitsaskpekten und Nutzungsparametern geschaffen, die durch im Rahmen der gegebenen Projektressourcen nicht möglichen weiteren, vertiefenden Analyseverfahren zugeführt werden könnten, um noch eingehendere und potenziell hoch relevante Erkenntnisse über Nützlichkeit, Nutzung und deren Konnex im Hinblick auf eine sowohl auf policy-Kontexte zugeschnittenen als auch qualitätsvollen Evaluation zu erbringen.

2. Nutzung von Programmevaluationen

Das vorliegende Kapitel konzentriert sich auf das summative Bild, das heute hinsichtlich von Art und Ausmaß des Nutzen gezeichnet werden kann, der den zahlreichen durchgeführten Programmevaluationen zugerechnet werden kann. Auf Faktoren, die Art und Ausmaß der Nutzungsweisen beeinflusst haben, wird sodann in Kapitel 3 eingegangen. Die Analyse geht von etablierten Kategorien des Evaluationsnutzens aus, die in der internationalen Nutzungsforschung zu Evaluation seit gut zwei Jahrzehnten tonangebend sind (vgl. Kapitel 1).

Das Kapitel stützt sich auf die Ergebnisse aus den Interviews mit AuftraggeberInnen bzw. HauptadressatInnen in den drei mit FTI-politischen Agenden betrauten Bundesministerien und Bundesagenturen sowie auf die Ergebnisse der EvaluatorInnen-Befragung. Beide Datenquellen beziehen sich auf die Gesamtheit aller Programmevaluationen, die im FTI-Bereich durchgeführt wurden. Das Bild, das anhand beider Datenquellen entsteht, ist hochgradig konsistent. Eine Quantifizierung der verschiedenen Phänomene der Evaluationsnutzung kann anhand der Interviews nicht vorgenommen werden. Die Ergebnisse der EvaluatorInnen-Befragung können hier weitere Aufschlüsse über größenordnungsmäßige Verhältnisse und Relationen der Nutzungsweisen zueinander geben. Die externen EvaluatorInnen sehen sich aus ihrer distanzierten Position heraus zu etwa einem Viertel nicht in der Lage, Einschätzung zur Entstehung von Nutzen aus den Programmevaluationen zu treffen, an deren Durchführung sie beteiligt waren. Ist grundsätzlich anzunehmen, dass externe EvaluatorInnen nur begrenzt Einblick in Evaluationsnutzungen haben, so verleiht doch der Umstand den Befragungsdaten gute Belastbarkeit, dass österreichische FTI-EvaluatorInnen in der Plattform fteval kontinuierlichen Austausch mit den AuftraggeberInnen pflegen und langjährig tätige, spezialisierte EvaluatorInnen nach wiederholten, unterschiedlichen Einsätzen auch die Rolle von SystemkennerInnen zukommt.[1]

Alle AuftraggeberInnen haben in den Interviews in überzeugender und nachvollziehbarer Weise dargestellt, dass durchgeführte Programmevaluationen genutzt wurden und werden. Die InterviewpartnerInnen haben auch an Beispielen erläutert, wie einzelne Programmevaluationen Nutzen erbracht haben und sich dabei teils auch auf Evaluationen bezogen, deren Berichte in der vorliegenden Studie der Berichtsanalyse anhand der Standards unterzogen wurden. Zugleich wurde deutlich auf Unterschiede zwischen verschiedenen Programmevaluationen hingewiesen, die zu verschiedenen Zeitpunkten mit unterschiedlichen Evaluationsdesigns unter unterschiedlichen Bedingungen beauftragt und durchgeführt wurden. Dies entspricht durchaus der Auffassungsweise der Standards, dass es „die richtige" Programmevaluation nicht gibt, sondern der Zuschnitt auf jeweils im Einzelfall vorliegende Bedürfnisse und Erfordernisse ins Zentrum zu stellen ist, um hohe Evaluationsqualität zu erreichen. Im breiten Blick über die Evaluationspraxis zweier Jahrzehnte ergibt sich seitens der AuftraggeberInnen an den politisch-administrativen Systemstellen, die zugleich die HauptaddressatInnen der Berichte und primären NutzerInnen der durchgeführten Programmevaluationen sind, das Bild eines grundsätzlich gelungenen und zufriedenstellenden Aufbaus einer Evaluationskultur:

„Ich glaube schon, dass es auch durch die Programmevaluierungen Veränderungen gibt." (A4)

„[Programmevaluation] gehört schon zu den Kernthemen. Wenn man weiß, wie früher Forschungsförderung betrieben wurde, dann hat sich schon Einiges gewendet. Transparenz, Nachvollziehbarkeit, dass Wirkung hinterfragt wird, das sind Kernthemen." (M2)

„[Wir sind] sicher einen ganz schönen Weg in den letzten Jahren gegangen, sodass wir in internationaler Sicht sicher nicht so schlecht liegen" (A1)

„Es ist bis zu einem gewissen Ausmaß und auf einem gewissen Niveau die Kultur der Evaluierung fest etabliert." (M2)

Anhand der Typologie von Evaluationsnutzen, die in der Evaluationsforschung zur analytischen Gliederung des komplexen und facettenreichen Geschehens „Nutzung" eingesetzt wird, lassen sich erhaltene Beschreibungen des entstandenen Nutzens näher einordnen. Zugleich zeigt sich bei dieser Gliederung, dass diese allgemein eingeführte Typologie in ihrer Abstraktheit einer Realität gegenübersteht, in der verschiedene Nutzungsformen ineinanderfließen und sich nicht immer klar trennen lassen.

[1] 44,9% der antwortenden EvaluatorInnen haben mindestens vier Programmevaluationen im Beobachtungszeitraum durchgeführt, insgesamt 73,5% zumindest zwei Programmevaluationen.

Instrumenteller Nutzen

Mit dem Begriff des instrumentellen Nutzens werden alle Reaktionen auf Evaluationsergebnisse (Datenlagen, Schlussfolgerungen, Empfehlungen) angesprochen, die im Gefolge zu einer Entscheidungsfindung über das evaluierte Programm führen. Dabei kommen sowohl Entscheidungen über Fortführung oder Einstellungen des Programms wie auch Adjustierungen eines weiterlaufenden Programms in Frage. Diese können wiederum sowohl unmittelbar zu Umsetzungsaspekten in den betrauten Agenturen entsprechend ihres jeweiligen Pouvoirs fallen, oder für eine nachfolgenden Programmphase in Programmdokumenten niedergelegt werden. Die AuftraggeberInnnen berichten vor allem über Nutzen auf der Ebene der Programmadjustierung, der sich mit zahlreichen durchgeführten Programmevaluationen verbunden hat.

„Jede Evaluation - mit einer Ausnahme - hat uns etwas gebracht, was das Design des Instruments betrifft, und hat uns als Institution etwas gebracht." (A3)

„Da gibt es im Detail dann immer wieder gute Vorschläge wo man sagt, da kann man nachschärfen, verbessern, man kann das nutzen. Das passiert dann schon. In die nächste Runde, das heißt Ausschreibung oder Programmphase, fließt das Wichtigste und Belastbarste dann schon ein." (A1)

„[Ich] kann ganz viele Fälle bestätigen, wo man aufgrund eines Ergebnisses die Richtung geändert hat, etwas anders gemacht hat, das ist dann halt nicht besonders spektakulär. Man hat dann nicht beendet oder etwas ganz Neues gemacht, sondern es geht dann um ganz andere Dinge, wo man auch wirklich hochqualitativen Input braucht, weil man weiß, dass man sich in hochsensible Entscheidungszonen begibt. (...) Das sind halt auch manchmal langsame schwerfällige Prozesse. Obwohl das für die Evaluatoren nicht so offensichtlich ist - dieses Ergebnis wurde sofort umgesetzt, so ist es halt nicht." (M2)

„Bei einer Zwischenevaluierung könnte sein, dass da etwas drin steht, wo man wirklich schnell reagieren muss. Aber im Normalfall wird man sagen, dort wo ein Programmdokument neu geschrieben wird, dort ist der Zeitpunkt, dass jemand etwas einbringt aus einer Evaluierung." (A3)

„Bei den Programmen ist es sicher unterschiedlich - die Agenturen machen ja alle in ihrem Bereich die Programmevaluierungen, und das hat sehr rasche direkte Auswirkungen, weil dann die Weiterentwicklung direkt erfolgt, nach Diskussion in den zuständigen Gremien." (M1)

„[Resultat der Programmevaluierungen ist,] dass man Manches besser versteht. Das ist sicher ein großer Zweck dabei. (...) Auf Ebene der Personen, die wirklich programmverantwortlich sind, würde ich sagen: Sie achten schon sehr darauf, dass sie etwas Vernünftiges machen mit dem Programm, das erlebe ich schon stark. (...) Die Lernfähigkeit ist da - absolut." (M1)

„Was ist der relevante Hebel, dass dieser Soll-Zustand [, der in der Programmkonzeption angestrebt wird,] tatsächlich zustande kommt? (...) Bisher hat das Programm so gegriffen, (...) aber wir wollen sehen, ob da etwas nicht in dem Ausmaß berücksichtigt wird, wie wir uns das vielleicht vorgestellt haben." (M1)

Nicht zuletzt wird berichtet, dass auf das Evaluationsergebnis hin, dass ein Programm mit Zielsetzungen überfrachtet war, in der Folge eine Zielbereinigung in Abstimmung mit den Programmeigentümern vorgenommen werden konnte. Aus einem Ministerium eines anderen Steuerungsbereichs wird ein ähnlicher Evaluationsnutzen in Bezug auf die verfolgten Zielsetzungen eines Programms dargestellt: *„[Es wurde erkannt,] dass bei hochgesteckten politischen Zielen das Ergebnis der Nicht-Umsetzung wenig überraschend ist, und dass uns auch nur bedingt bewusst war, dass das ja ein Wunder-Programm hätte sein sollen, in der politischen Formulierung in Bezug auf alle Politikbereiche, das war eher marketing-technisch interessant oder werbetechnisch, aber die EvaluatorInnen haben das sehr ernst genommen und uns erklärt: die ganzen Dinge sind gar nicht erreicht worden."* (M1) Auch an weiteren Systemstellen wurde kommuniziert, dass Klärungen zur Machbarkeit von politisch motivierten Zielformulierungen als wesentlicher Aspekt eines Evaluationsnutzens betrachtet werden. Im Zuge der voranschreitenden Entwicklung und Etablierung der Evaluationspraxis hat sich eine Reflexion des Programmanagements und der Programmkonfigurationen als zentrale Nutzungsform der Programmevaluationen herauskristallisiert. Der Wert, der der Durchführung von Programmevaluationen zugesprochen wird, ergibt sich im Rahmen einer reflektierend-überprüfenden Haltung gegenüber Programmen so *„ganz pragmatisch: ob die Zielerfüllung möglich war, und - wesentliche Fragestellung - ob das, was man mit den Instrumenten umsetzen möchte, möglich ist oder nicht." (M1)*

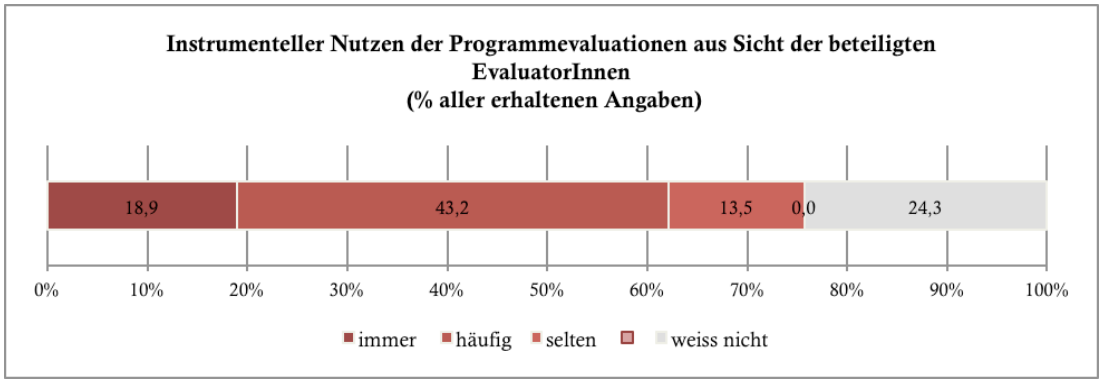
Gegenüber einer derartigen Evaluationsnutzung zur schrittweisen Aus- und Umgestaltung von Programmen hat die Fundamentalentscheidung, ob ein Programm überhaupt weitergeführt oder eingestellt wird, eher geringeren Stellenwert. Aus den erhaltenen Aussagen der InterviewpartnerInnen in ihren jeweiligen Aufgabenbereichen ergibt sich eine Tendenz, diese am deutlichsten sichtbare Form einer Evaluationsnutzung im Zuge immer besserer, auf immer mehr Erfahrungen und frühere Evaluationsergebnisse gestützten Programmkonzeptionen für eher unwahrscheinlich zu erachten. Sie

weisen darauf hin, dass die Fundamentalentscheidung über das Schicksal eines Programms den Charakter einer politischen Entscheidung hat, in die sie als AuftraggeberInnen und Haupt-adressatInnen der Evaluationen wenig involviert sind. Damit unterscheidet sich die österreichische Situation im FTI-politischen Bereich freilich kaum von vielfältigen internationalen Erfahrungen. „In practice, evaluation is most often called to help with decisions about improvig programs, projects, and components. Go/no-go, live-or-die decisions about programs are relatively rare, and reliance on evaluation data to make those decisions is rarer still." (Weiss 1998b: 32f.)

Aus den Auskünfte der AuftraggeberInnen ergibt sich, dass die in Programmevaluationen enthaltene Information im Gesamtinformationssystem, das der FTI-Governance zur Verfügung steht, keine direkte Konkurrenz in anderen Informationsmaterialien und –quellen hat. Dennoch handelt es sich bei Umsetzungen von Erkenntnissen aus Evaluationen weder um Automatismen noch um eindimensionale Ereignisse. Alle Auftraggeber-Institutionen versehen erhaltene Evaluationsergebnisse mit dem Vorbehalt, dass sie als Motivationen und Ansatzpunkte für Reflexionen erachtet werden, die ihre Sinnhaftigkeit gerade auch daraus gewinnen, dass mit erhaltenen Ergebnissen nicht „sklavisch" umgegangen wird. Beispielhaft für diese Haltung sind folgende Aussagen: *„Es gibt auch Evaluierungen, wo man aufgrund weiterer Überlegungen genau das Gegenteil macht von dem, was empfohlen wurde." (M2) „Die Existenzfrage [für die evaluierten Programme] hat sich selten gestellt (….), die ja eine enorm komplexe Frage ist. Es ist ein Baustein für eine Entscheidungsgrundlage, aber da gibt es andere Faktoren auch." (M2) „Eine Evaluierung kann zu ganz kritischen Ergebnissen kommen. Und die müssen sich in der Debatte durchsetzen. Natürlich gibt es bei jedem Programm auch irgendwie Interessen und Lobbying, die Arena wie Politik gemacht wird. Manchmal bleibt viel über, und manchmal weniger. Aber das, glaube ich, ist das Spiel, dem man sich stellen muss." (A1)*

Aus Sicht von drei Viertel der EvaluatorInnen, die die Umfrage beantwortet haben, ist instrumenteller Nutzen bei den von ihnen durchgeführten Programmevaluationen eingetreten. Ein Viertel sieht sich nicht in der Lage, eine Einschätzung abzugeben. Kein/e einzige/r EvaluatorIn gibt an, dass derartiger Nutzen bei den durchgeführten Evaluationen nie eingetreten ist. Aus Sicht von knapp einem Fünftel wurden Entscheidungen im Anschluss an das Vorliegen von Evaluationsergebnissen immer herbeigeführt, aus der Sicht von knapp der Hälfte der EvaluatorInnen häufig. Aus den Hinweisen von Seiten der GeprächspartnerInnen in den Institutionen, dass es sich hier auch um kleinteilige und von außen nur schlecht erkennbare Prozesse handelt, kann angenommen werden, dass die Effekte von den EvaluatorInnen tendenziell unterschätzt werden.

Abbildung 7: Instrumenteller Nutzen aus der Sicht der EvaluatorInnen



Wenn der Wert der Evaluationen augenscheinlich hochgradig in einem Wissenszuwachs besteht, auf dessen Grundlage Programme nachgeschärft bzw. in späteren Programmphasen besser ausgerichtet werden können, so können relevante Erkenntnisse auch späterhin und an anderen Einsatzpunkten des betreffenden FTI-Steuerungsbereichs niederschlagen. Dies kann in der späteren Formulierung von Programmen einer nachfolgenden Generation stattfinden, im Umgang mit anderen Steuerungsinstrumenten im eigenen Verantwortungsbereich, oder in der Schaffung von Aufmerksamkeit für Themenstellungen und Bedarfslagen, die fortan auf die Agenda gesetzt werden. Um derartige Nutzungsweisen geht es im folgenden Abschnitt.

Konzeptueller Nutzen

Unter konzeptuellem Nutzen versteht die Evaluationsforschung Lerneffekte, die aus Evaluationen hervorgehen, indem Programmbeteiligte neue Sichtweisen auf den Evaluationsgegenstand entwickeln, ohne dass sich dies mit einer unmittelbaren Entscheidung zum untersuchten Programm verbindet. Wie die Interviews zeigen, ist es gerade diese Form des Nutzens, die häufig eintritt und die der bisherigen Evaluationspraxis Wert verleiht. Eine Reihe von Aussagen von verschiedenen Akteuren bzw. VertreterInnen der relevanten Organisationen belegt dies, aus denen die folgenden Beispiele herausgegriffen sein sollen:

„Die Empfehlungen waren gut, sie wurden umgesetzt, und [das Programm] kommt so an und verfährt so weiter. Und [die Evaluation] hat bewirkt, dass das [Thema] in [Gremien und Institutionen] als wichtiges Handlungsfeld wahrgenommen wird und als Kernaufgabe. Ohne diese Evaluationen wäre es uns nicht so leicht gefallen, das so stark in [den befassten Gremien und Institutionen] zu verankern." (A3)

„[Programmevaluation ist] ein gutes Tool, um Sachen zu lernen. Da ging es um Programmanagement, um die Frage wie organisiert man neue Themen und ein Programmanagement das neue Themen entwickelt, das Communities entwickelt. Wir haben Dinge entdeckt, die wir im Haus nicht machen können.(...) Die wichtigen Elemente waren eher Lernprozesse wie ein Gesamtsystem funktioniert mit den verschiedenen Organisationen, die daran beteiligt sind, wie sind Ablaufprozesse." (M2)

„Wir haben sicher dazu gelernt, so richtig falsch gestrickte Programme, wie es sie noch vor 15 Jahren gegeben hat, [gibt es heute nicht mehr.] Es ist schon ein anderes Niveau jetzt." (A1)

Konzeptuelle Einsichten aus einzelnen Programmevaluationen können durchaus so eintreten, dass sie nicht unmittelbar mit hoch sichtbaren Ereignissen in der Geschichte des politisch-administrativen Handelns in Verbindung gebracht werden können, und sie können aus Programmevaluationen erwachsen, die auf der Ebene von instrumentellem Nutzen nicht sonderlich ertragreich erschienen sind. *„[Es] war der Eindruck der Fachabteilung, dass viel zusammengetragen wurde, das (...) sehr nützlich ist, da über das lang laufende Programm sehr viel an Informationsmaterial zusammengetragen wurde, weil Daten sehr gut aufbereitet wurden. (...) Die Ergebnisse im Großen waren für die Fachabteilung nicht überraschend, der Neuigkeitswert für die Fachabteilung auch nicht wirklich sehr hoch. Aber es fließt ein, die Ergebnisse fließen in die laufende Arbeit ein." (M1)* Zur selben Programmevaluation äußert sich ein/e andere GesprächspartnerIn so: *„[Hier] war es für uns teilweise durchaus spannend, was die Ergebnisse waren. Es bleibt immer eine Mischung: Es gibt immer Einiges, das man annimmt, das in der Luft liegt. Wenn das dann auch [mit harten Daten] unterlegt wird, ist das hilfreich bei Annahmen, die man im systemischen Ansatz hat. Es war durchaus interessant zu sehen, wie pointiert das Evaluationsteam gesehen hat, inwieweit [bestimmte Gründe die Erreichung einer Zielgruppe] beeinflusst haben." (M1)* Bezeichnend ist auch das folgende Attest eines wiederholten Eintretens von Lerngelegenheiten: *„Meistens ist es so, dass wir aus unserem Datenbestand die Daten an die Evaluatoren liefern, und die dann mit spezifischen Methoden, die uns manchmal neu sind, meistens aber nicht, über diesen Datenbestand drüber gehen und gewisse Schlüsse ziehen, und - das ist dann oft neu - mit anderen Datenbeständen, die wir vielleicht nicht gekannt haben, vergleichen und ins Verhältnis setzen. Das ist schon ein Nutzen, den man dann unmittelbar hat (...), das ist sicher sehr hilfreich." (A2)*

Konzeptuellen Nutzungen, die im FTI-politischen System eingetreten sind, ist auch ein Lernen über institutionell-organisatorische Konfigurationen und bestgeeignete Allokationen von Instrumentarien zuzurechnen: *„Was wir schon aus den Evaluierungen der Vergangenheit gemacht haben – das ist bei Weitem noch nicht abgeschlossen, da die Agenturen sich noch immer in einem bestimmten Wettbewerb befinden –, ist, dass man die Portfolios der Agenturen möglichst gut abstimmt." (M1)*

Die InterviewpartnerInnen stellen dar, dass auch Evaluationen, die zu letztlich nicht weitergeführten Programmen durchgeführt wurden, Einsichten und Lerneffekte hinsichtlich von Funktionsweisen des FTI-Systems bzw. der auf es gerichteten Steuerungs- und Anreizsysteme erbracht haben. Anhand der Weiterentwicklung von FTI-politischen Systemkomponenten, wie sie etwa mit neueren Programmgenerationen zur Wissenschafts-Wirtschafts-Kooperation oder mit Anpassungen in der Maßnahmenfamilie rund um den Innovationsscheck vorliegen, sind solche Nutzungen nachvollziehbar.

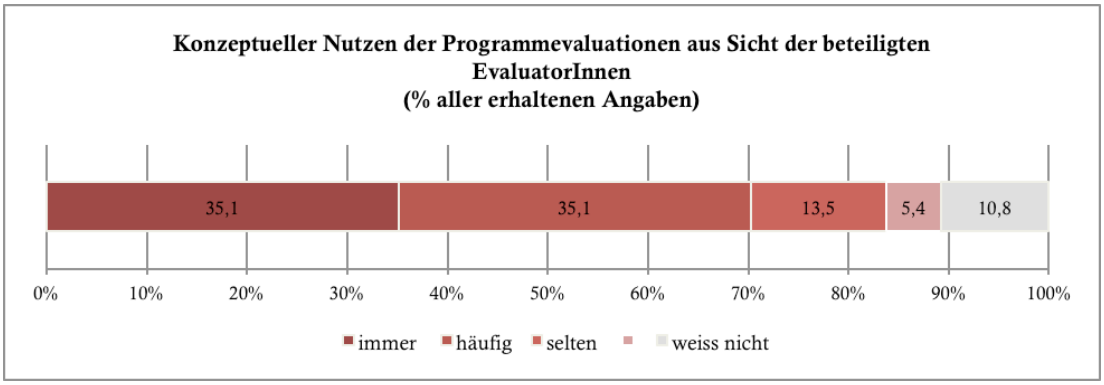
Zeigt sich also Programmevaluation aus Sicht aller Akteure, mit denen Interviews geführt wurden, als wesentlicher Beitrag zum Lernen in ihrem Handlungsbereich, so lagen und liegen doch auch Hindernisse vor, um derartige konzeptive Weiterentwicklungen auch auf operationaler Ebene zur Geltung zu bringen. So wird etwa zur Entwicklung neuer Sichtweisen auf den Evaluationsgegenstand gesagt: *„[Das] tritt häufig ein - abhängig von der Qualität der Evaluierung. Ob wir das Gelernte dann immer umsetzen können ... [steht auf einem anderen Blatt]." (A1)* Ähnliche Hinweise auf Einschränkungen und Reibungsverluste da, wo unmittelbar mit Evaluation befasste Akteure zu Einsichten gelangt sind, die ihnen aus ihrer Systemposition heraus wichtig erscheinen, liegen vor allem von Seiten der Agenturen

vor, aber auch von GesprächspartnerInnen in Ministerien. Dies überrascht insofern nicht, als es sich grundsätzlich um Aktivitäten in Multiakteurs-Szenarien handelt, die die Nutzungsforschung wiederholt als kaum vernachlässigbare Komponente von Settings der Evaluationsnutzung ausgewiesen hat. Auf Einschränkungen und Beeinträchtigungen im spezifischen Bereich der österreichischen FTI-Programmevaluation, die sich sowohl mit der wahrgenommenen Qualität von Evaluationen als auch mit Principal-Agent-Verhältnissen und Relationen zur politischen Sphäre der Entscheidungsfindung verbinden, wird in späteren Abschnitten des vorliegenden Berichtes zurückzukommen sein.

Insgesamt wurde in allen Gesprächen ersichtlich, dass die über die Jahre durchgeführten Programmevaluationen von den zentralen Akteuren als wesentliche Beiträge zu einer Verbreiterung und Vertiefung der Wissensbasis eingeschätzt werden, auf die sich FTI-politisches Handeln gerade auch als aktualitätsbezogenes und voranschreitendes Handeln stützt. Zugleich wird ersichtlich, dass es sich beim Eintreten von Nutzen aus Programmevaluationen um Gemengelagen handelt, in der nicht nur eine Evaluation zu einer Nutzung führt, sondern multiple Effekte auftreten. Im Resultat der durchgeführten Untersuchung erscheint es naheliegend, von einem instrumentell-konzeptuellen Komplex der Evaluationsnutzung im österreichischen FTI-Bereich zu sprechen. Insbesondere begleitende Evaluation kann eine wesentliche Systemfunktion erfüllen, um Programme in ihrer konkreten Entwicklung verstehen zu können und durch in Objektivität gegründetes Lernen sowohl etwaige angebrachte Nachadjustierungen vornehmen zu lassen als auch Reflexionen über zukünftige Maßnahmen-Konfigurationen zu unterstützen. Entsprechend werden die in der österreichischen FTI-Evaluationspraxis dominanten „Zwischenevaluierungen“an allen Stellen des relevanten Systems auf Bundesebene als wesentliche Mittel geschätzt, um zu relevanten Einsichten zu gelangen.

Die EvaluatorInnen, die die durchgeführte Umfrage beantwortet haben, attestieren zu zwei Drittel, dass ihrem Wissensstand nach die durchgeführten Programmevaluationen stets oder zumindest häufig konzeptuellen Nutzen erzeugt haben. Den Angaben der EvaluatorInnen zufolge handelt es sich beim Gewinnen neuer Sichtweisen auf den Evaluationsgegenstand um die am Häufigsten eingetretene Nutzungsweise unter allen in der verwendeten Typologie unterschiedenen Nutzungstypen von Programmevaluationen.

Abbildung 8: Konzeptueller Nutzen aus Sicht der EvaluatorInnen



Symbolischer Nutzen

Von symbolischem Nutzen spricht die Evaluationsforschung da, wo das Vorliegen eines Evaluationsberichts oder die Tatsache, dass überhaupt evaluiert wird, zur Rechtfertigung bereits zuvor getroffener Entscheidungen herangezogen wird, oder wo die Durchführung von Evaluationen zur formalen Untermauerung dient, dass mit dem Programm rational umgegangen wird, etwa um andere Akteure in der politischen Sphäre vom Programm zu überzeugen. Wie die Interviews zeigen, sind auch derartige Aspekte in der Evaluationspraxis des österreichischen FTI-Bereichs durchaus anwesend. Die Grenzen zu inhaltlichen Erträgen der Evaluationen sind dabei fließend und nur schwer festzumachen.

Hier geht es zunächst um die Überzeugungsarbeit, die in der Multiakteurs-Arena der politisch-administrativen Sphäre stets zu leisten ist, wenn es um Veränderung geht. *„Es ist eine kleine Untermauerung (...), zumindest eine Argumentationshilfe. Das heißt noch nicht immer, dass das Ergebnis der Evaluierung allen passt, das muss man dazu sagen, aber dort, wo das Mindset zusammenpasst, ist es eine*

sinnvolle Argumentationshilfe." (A2) *„Was für den politischen Prozess wichtig war, ist, dass wir mit der Evaluation zumindest argumentieren können, dass [die Agentur] eine Rolle in [diesem FTI-politischen Steuerungsbereich] haben sollte." (A3)* Gut ausgearbeitete Entscheidungsgrundlagen, die mit externer Expertise erstellt sind und damit auch in der politisch-administrativen Arena mit Anerkennung rechnen können, gelten Verantwortlichen für Steuerungs- und Maßnahmenbereiche der FTI-Politik als unabdingbar für Umsteuerungen von Programmen. Dies insbesondere da, wo geteilte Programmeigentümerschaft mehrerer Ressorts vorliegt: *„Wenn Sie da keine saubere Evaluierung haben, können Sie keine Richtungsänderung vornehmen." (M2).*

Für alle derartigen Fälle erscheint eine klare Trennung zwischen dem instrumentell-konzeptuellem Nutzenkomplex, der in den beiden vorangegangenen Abschnitten dieses Kapitels beschrieben wurde, und einer formal orientierten Erbringung von Rationalitätsuntermauerungen zu artifiziell. Zu berücksichtigen ist allerdings in einem realistischen Bild der Art und Weise, wie es im österreichischen FTI-System zu Umsetzungen und Impact von Evaluationsergebnissen kommt, dass innerhalb einer von Principal-Agent-Verhältnissen, Hierarchien und Kleinteiligkeit geprägten Situation zahlreiche Schnittstellen vorliegen und in Verhandlungs- und Genehmigungsprozessen zu bewältigen sind. Die Antizipation dieser Herausforderungen schlägt sich bereits in der Konzeptions- und Planungsphase von Programmevaluationen nieder. Aus der Sicht einer der beiden nicht völlig autonomen Agenturen stellt sich übergreifend die Erfahrung mit der Planung von Programmevaluationen so dar: *„Es ist ein komplexes Thema, weil da oft Interessen vom Auftraggeber drinnen stecken, oder Fragestellungen, wo der Auftraggeber intern noch eine Hierarchiestufe über sich hat, die er vielleicht auch noch einmal überzeugen möchte. Im Ministerium gibt es ja noch einmal Hierarchien, die möglicher Weise nicht ganz auf einer Linie sind, was ihre Interessenslagen angeht." (A2)* Gerade auch die Kleinteiligkeit der Strukturen sorgt hier für einige Unübersichtlichkeit, da Agenturen zahlreiche Schnittstellen zu verschiedenen Programmeigentümern haben bzw. Programmverantwortlichkeit über zahlreiche Systemstellen verteilt ist, die jeweils ihre eigenen Charakteristika aufweisen.

Schließlich kann die systemimmanente Notwendigkeit einer Überzeugungsarbeit, die sich mit der Einrichtung und Weiterführung von Programmen verbindet, auch in das der Evaluationsforschung gut bekannte Phänomen münden, dass Programmzuständige „ihre" Programme, an die sie glauben und die sich auch mit ihrem beruflichen Status und ihren Karrierechancen verbinden, durch die Vorgehensweise von Programmevaluationen zu verteidigen suchen (vgl. z.B. Weiss 1998b: 39f). Wie ein/e GeprächspartnerIn es ausdrückt, scheut man manches Mal vor einer *„Selbstbeschneidung der Möglichkeit der Programmgestaltung"* zurück (M1). Derartigen Phänomenen stehen zugleich all jene Äußerungen aus nahezu allen Gesprächen gegenüber, in denen politische Erwartungen an Programme oder Vorgaben thematisiert wurden, mit denen die administrativen Stellen auch auf dem Weg der Programmevaluationen umzugehen haben. Programmevaluationen kommt hier die Rolle zu, eine Einschätzung der Realitätshaltigkeit politischer Erwartungen an Programme zu ermöglichen.

Zum Phänomenkomplex der Erzeugung von symbolischem Nutzen zählt auch die Art der Verankerung der Evaluationsfunktion im rechtlich-institutionellen Rahmen der Bundesverwaltung. Programmevaluationen fungieren an der Schnittstelle zwischen Fachressorts und dem Bundesministerium für Finanzen (BMF), wo sie einen direkten Konnex zur Legitimation der Mittelausgaben haben. Darüber hinaus werden sie bei kontrollorientierten Betrachtungen der Bundesverwaltung durch den Rechnungshof herangezogen. Diese legitimierende Funktion *„schwingt immer mit"* bzw. ist *„beim Motivbündel immer dabei" (A1)*, das hinter der Planung, Durchführung und institutionellen Verwertung einer Programmevaluation steht.

Die legitimatorische Funktion der Programmevaluationen reichert die Gemengelage an Nutzungsweisen, die in der grundsätzlichen Anlage der Evaluationsprojekte immer schon mitgedacht sind, um eine weitere Komponente an und sorgt dadurch für eine innere Spannung, die jedes Evaluationsprojekt grundsätzlich durchzieht. Der konkrete, situative Umgang mit diesem Spannungsfeld zwischen Legitimationspflicht und lernorientierter Erkenntnis kann dazu führen, dass Programmevaluationen den Charakter einer *„Pflichtübung" (M1)* annehmen und Programmevaluationen in eine primär formalistisch wahrgenommenen Routine verfallen. *„Dieses kritische Hinterfragen, ich habe schon den Eindruck, dass das gemacht wird, aber es kommt sicher auch auf den Auftrag an. (...) Manchmal hat man schon das Gefühl, es steht halt im Programmdokument drinnen, es hat eine Zwischenevaluierung stattzufinden. Das ist sehr wohl eine Gefahr, dass das zu einer Pflichtübung degeneriert." (A2)*

Das Spannungsfeld zwischen Rechenschaftslegung, lernorientierter Auseinandersetzung mit dem Programm und politischer Willensbildung artikuliert sich in feinen Nuancen: *„Wenn wir offen sein können, da kein politisches Programm besteht, dann werden wir diese Evaluierungsergebnisse analysieren und anschauen. (...) Dann sagen wir: wenn wir das dahingehend abändern, dann würde es wieder Sinn machen. (...),*

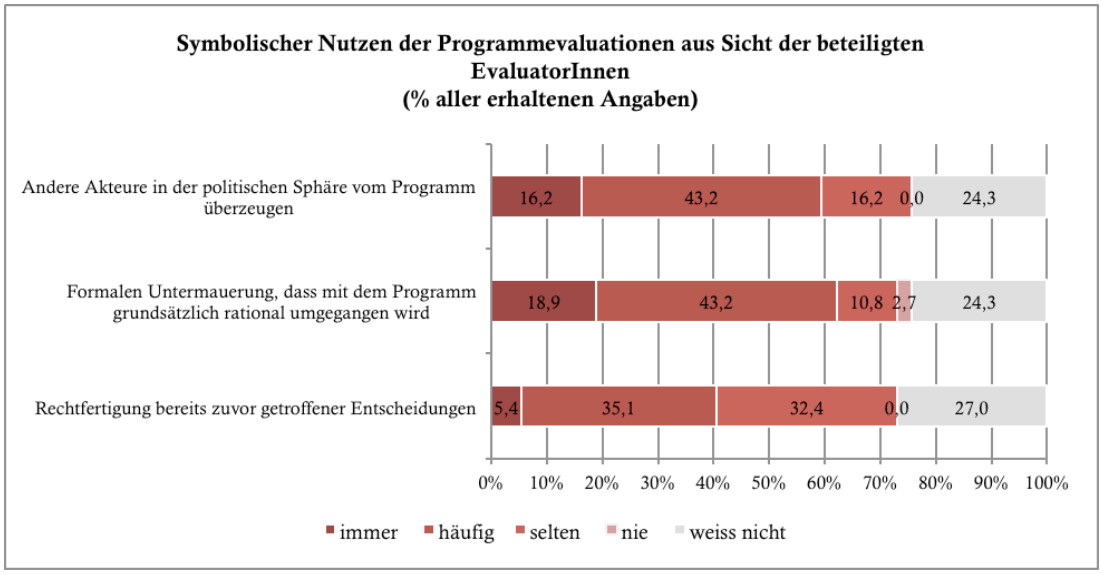
Sonst nehmen wir schon vielfach auch Rücksicht, wir haben ja nicht nur uns selbst als Bewerter, sondern das Finanzministerium schaut natürlich bei Programmfortsetzungen auf die Evaluierungen und sagt, von der Evaluierung her kann das nicht unsere Zustimmung finden." (M1)

Der Umgang mit dem Spannungsfeld kann aber auch dazu führen, dass die Evaluationsergebnisse unter Umständen nicht ausführlich reflektiert werden oder vorschnell genutzt werden. Aus einer Agentur wird etwa in diesem Sinn berichtet: *„Eine direkte Nutzung von Evaluationsergebnissen ist regelmäßig gegeben, allerdings wahrscheinlich oft in Arten und Weisen, dass man sagt, das und das hat die Evaluierung gebracht, und da müssen wir das machen, denn wenn der Rechnungshof fragt, müssen wir beweisen können, dass wir die Evaluierung berücksichtigt haben (...) Das formale Abhaken ist auch das eine oder andere Mal zu Lasten der Substanz gegangen." (A1)*

Wenn somit davon auszugehen ist, dass in der Vergangenheit nicht näher bezifferbare Fälle eingetreten sind, in denen Programmevaluationen von vornherein von ihren AuftraggeberInnen kaum mit Nutzenerwartungen verbunden wurden, zumindest nicht in Termini einer Option auf eine größere Umgestaltung des Programms, so stehen dem doch all jene Fälle gegenüber, in denen ein direkter Nutzen in der Programmadjustierung oder ein konzeptiver Nutzen für die Weiterentwicklung des FTI-politischen Systems erzielt wurde. Ein/e MinisteriumsvertreterIn äußert sich dazu folgender Maßen: *„Es ist tatsächlich so, Sie werden niemanden finden, der in unserem Feld für ein Programm verantwortlich ist und sagen würde: das ist egal oder das brauchen wir nicht. Es ist 100%iger Bestandteil. Auch, da immer mehr von außen verlangt wurde, von BMF oder Rechnungshof, da können Sie es sich schlichtweg nicht leisten, irgendein Programm zu machen, wo nicht drinnen steht wann das evaluiert wird. Was immer wieder kritisiert wurde, ist, dass insgesamt so ein Zugang herrscht, man macht das anstandshalber, so wirklich verändern tut es eigentlich nicht. Es ist aber nicht so leicht, das wirklich zu bewerten. Was ich in den letzten Jahren beobachten konnte, ist, dass es immer mehr zu einer teils chaotischen Ausdifferenzierung kommt. [Es] ist immer mehr die Funktion von Evaluation als verändernde Kraft wichtig geworden. [Wir haben] auch das Gefühl: jetzt haben wir durch die vergangenen Evaluierungen ein sehr genaues Bild bekommen von dem, was da ist und wie die Welt funktioniert." (M2)*

Betrachtet man die Aussagen der EvaluatorInnen, so wird hier symbolischer Nutzen ebenfalls als wesentliche Dimension der Evaluationsnutzung erkennbar, wenn auch mit vergleichsweise geringerem Stellenwert als inhaltliche Nutzungsweisen durch die HauptadressatInnen der Programmevaluationen. 62% der EvaluatorInnen meinen, dass die von ihnen durchgeführten FTI-Programmevaluationen immer oder häufig der formalen Untermauerung dienten, dass mit dem Programm grundsätzlich rational umgegangen wird. 60% meinen, dass ihre FTI-Programmevaluationen immer oder häufig dazu dienten, andere Akteure in der politischen Sphäre vom Programm zu überzeugen. Vergleichsweise seltener wird angegeben, dass die Evaluationen der Rechtfertigung bereits zuvor getroffener Entscheidungen gedient hätten, doch wurde auch dies von 41% der EvaluatorInnen immer oder häufig beobachtet.

Abbildung 9: Symbolischer Nutzen aus Sicht der EvaluatorInnen



Aufklärung

Mit diesem Begriff werden Anreicherungen des Wissens thematisiert, das über die enge Sphäre der unmittelbar mit einem Programm befassten Akteure hinaus auch für andere Akteuren verfügbar wird. Für die FTI-politische Sphäre geht es also um Wissensflüsse, die über die für ein Programm zuständigen Personen in Fachressorts, die aus dieser Position heraus zugleich als AuftraggeberInnen der Programmevaluationen fungieren, und die Agenturen, die mit der Programmumsetzung betraut sind bzw. je nach Autonomiestatus die Programmevaluationen auch selbst beauftragen, hinaus reichen. Dies betrifft einerseits Wissenszuwächse für weitere Akteure im Umfeld des jeweils untersuchten Evaluationsgegenstands, die an verschiedenen Stellen des FTI-politischen Systems mit thematisch verwandten Instrumenten befasst sind, wie etwa in anderen Fachabteilungen desselben Ministeriums oder in benachbarten Agenturen. Andererseits interessieren hier Informationsflüsse, die auch noch breitere politische oder gesellschaftliche Sphären erreichen, von Fachverbänden und diversen Akteursgruppen des nationalen Forschungs-, Technologie- und Innovationssystems bis hin zur wissenschaftlichen Diskussion.

Aus den Interviews mit den AuftraggeberInnen bzw. HauptadressatInnen der Evaluationen geht ebenso wie aus den weiter unten dargestellten Ergebnissen der EvaluatorInnen-Befragung hervor, dass die Generierung dieser Form des Nutzens einen durchaus noch ausbaufähigen Aspekt der bisherigen Evaluationspraxis im FTI-Bereich darstellt.

Zur Weitergabe von evaluativer Information innerhalb und außerhalb der auftraggebenden Institutionen existieren Vorgänge zur Informierung übergeordneter Hierarchiestufen, das Instrument der Berichtspublikation, und die Verfügbarkeit für die zugriffsberechtigte Beamtenschaft über den elektronischen Akt (ELAK). In den Forschungs- und Technologieberichten werden Programmevaluationen hinsichtlich ihrer wesentlichsten Ergebnisse vorgestellt. Mit der *Plattform fteval* steht ein Austauschforum zur Verfügung, in dem vor allem die mit Evaluation befassten VertreterInnen der beteiligten Institutionen und Organisieren kommunizieren. Daneben spielen institutionalisierte Kontakte zwischen den handelnden Personen an verschiedenen Punkten des Governance-Systems sowie, in augenscheinlich nicht unbeträchtlichem Ausmaß, informelle Kontakte zwischen Einzelakteuren eine Rolle für das Eintreten von Informationsflüssen. Es wird unter den InterviewpartnerInnen allgemein davon ausgegangen, dass in diesem Rahmen im Wesentlichen bekannt ist, welche Informationen existieren, und diejenigen Informationen auch erhalten werden können, die von Interesse sind – jedenfalls innerhalb der Akteursgruppe des FTI-politischen Systems, die sich direkt mit Evaluationen befasst. *„Die Information gibt es, und selbst wenn nicht ein strukturierter Austausch ist, erfährt man es." (M1)*

Aus den Auskünften der GesprächspartnerInnen wurde zugleich klar ersichtlich, dass die Möglichkeiten dafür, innerhalb der Ressorts Evaluationsergebnisse fachabteilungs-übergreifend zur Kenntnis zu nehmen und zu behandeln, eingeschränkt sind. Noch deutlich herausfordernder erscheinen die Möglichkeiten zur Herstellung von übergreifenden Wissensflüssen zwischen den drei FTI-politischen Steuerungsbereichen, da organisatorische Vorkehrungen nicht existieren, die eine Systematik von Austausch und übergreifender Reflexion gewährleisten könnten (vgl. dazu auch Kapitel 5).

Aus beiden heutigen Ministerien sowie aus Agenturen wird berichtet, dass es vor allem in jüngerer Zeit zunehmend Initiativen gegeben hat, Evaluationsergebnisse einer breiteren Gruppe von KollegInnen im Haus auch in Form von Präsentationen vorzustellen und Abteilungs-übergreifende Diskussionen zu veranstalten. Dies wird durchgängig als sehr produktiver Vorgang beschrieben. Allerdings wird die hausinterne Präsentation von einem/r InterviewpartnerIn auch mit dem einschränkenden Hinweis versehen, dass es hier darauf ankomme, ob eine solche Präsentation in Richtung eines fachlichen Austauschs angelegt ist, oder in erster Linie in Form einer Erfolgsmeldung zum untersuchten Programm gestaltet wird.

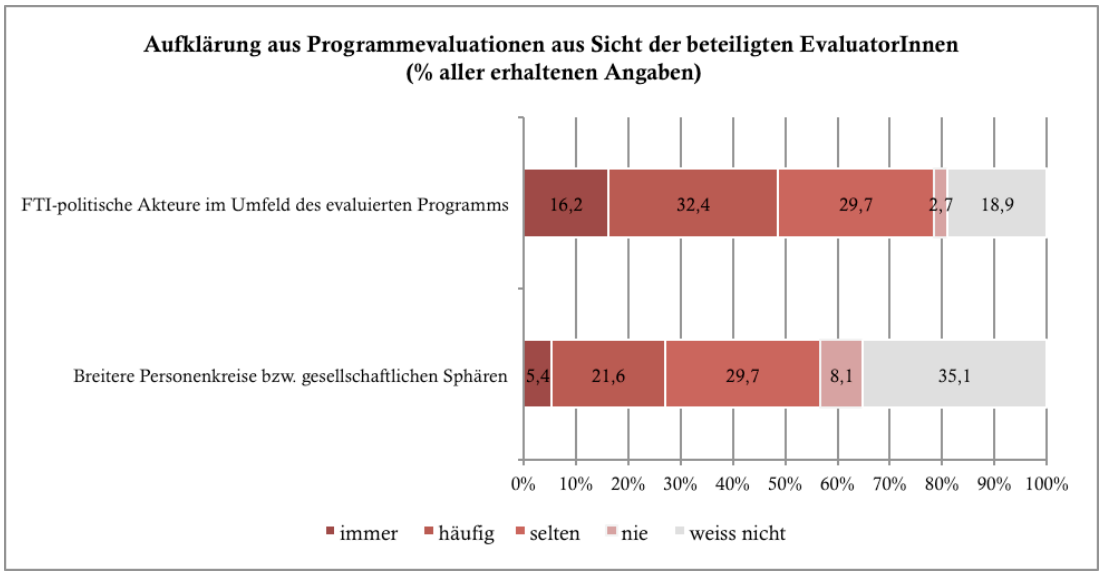
Hinsichtlich einer Außenkommunikation der FTI-politischen Systemstellen hin zur Akteurslandschaft des nationalen Innovationssystems wird noch umso mehr davon ausgegangen, dass Akteure, die mit den in den Evaluationen zur Debatte gestellten Instrumenten erreicht werden sollen oder im relevanten Steuerungs- und Maßnahmenbereich eine Rolle für die Entwicklung des jeweiligen FTI-Segments haben, die Informationen über Verlauf und Einschätzung von Maßnahmen grundsätzlich erhalten können, wenn sie daran interessiert sind. Aus einem Ministerium wird diesbezüglich berichtet, dass sich relevante Akteure wie z.B. Fachverbände im Anschluss an Evaluationen einstellen, um in Austausch über die Maßnahmen bzw. davon betroffenen Themen und Segmente der FTI-Landschaft zu treten. Von Seiten einer Agentur wird berichtet, dass Evaluationsergebnisse in der

Vergangenheit in unterschiedlichem Ausmaß, je nach untersuchtem Programm, auch an andere institutionelle Akteure herangetragen wurden.

Was die Publikation der Evaluationsberichte anbelangt, ist für die Vergangenheit von einer nur unvollständigen Verfügbarmachung aller erarbeiteten evaluativen Information auszugehen.[2] Innerhalb des von der vorliegenden Metaevaluation betrachteten Zeitraums haben sich die in der Plattform fteval vertretenen Akteure darauf verständigt, einer Publikationspflicht zu folgen. Die Haltungen der unmittelbar evaluationsverantwortlichen Stellen zur Publikation „ihrer" Programmevaluationen erweisen sich als unterschiedlich. Während sich die meisten Akteure heute zur systematischen Publikation bekennen, heißt es von einer Seite: *„Es sind keine geheimen Dinge. Wir haben uns sicher nicht aktiv darum gekümmert. Wenn sich der Auftraggeber nicht darum kümmert, [so] glaube ich doch nicht, dass dem etwas in den Weg gelegt würde."* Von einer weiteren Seite werden freilich massive Bedenken hinsichtlich des Umfangs der rezenten Publikationstätigkeit angemeldet. Hier wird darauf hingewiesen, dass sich die Erarbeitung evaluativer Information aufgefächert hat in klassische Programmevaluationen, Assessments, Reviews und wissenschaftliche Studien mit evaluatorischem Charakter, deren Ergebnisse jüngst nicht alle verfügbar gemacht worden seien.

Aus der Sicht der EvaluatorInnen, die mit der Umfrage erreicht wurden, stellen sich übergreifende Wissensflüsse im FTI-System folgender Maßen dar: 49% gehen davon aus, dass Evaluationsergebnisse in der Vergangenheit bei allen von ihnen durchgeführten Programmevaluationen, oder zumindest häufig, auch das verfügbare Wissen angereichert haben, das Akteure im Umfeld des evaluierten Programms nutzen konnten bzw. können (z.B. andere Abteilungen desselben Ministeriums oder derselben Agentur, andere mit FTI befasste Ministerien). Dass das durch die Evaluationen erzeugte Wissen auch weiteren Personen bzw. gesellschaftlichen Sphären zugute kam, wird lediglich von 27% der EvaluatorInnen als stets oder zumindest häufig eingetretener Effekt bezeichnet.

Abbildung 10: Nutzenform Aufklärung aus Sicht der EvaluatorInnen



Eine vollständige Publikation der Endberichte zu den von ihnen erarbeiteten Programmevaluationen, einschließlich aller Anhänge ohne irgendeine Abänderung, sehen die antwortenden EvaluatorInnen nur zu 14% immer verwirklicht. 75% berichten Abstriche von einer vollumfänglichen Publikation, wobei immerhin 11% angeben, dass sie bei keiner der von ihnen durchgeführten Programmevaluationen erfolgt ist.

[2] So zeigte sich etwa im Zuge der in dieser Metaevaluation angestellten Berichtsanalyse, dass auch auf der Homepage der *Plattform fteval* verfügbare Evaluationsberichte teils auch lediglich Kurzberichte zu Programmevaluationen darstellen, zu denen die Langberichte nicht herausgegeben wurden. Ein ursprünglich auf Basis des Samplings für die Analyse vorgesehener Evaluationsbericht musste deswegen durch einen anderen ersetzt werden.

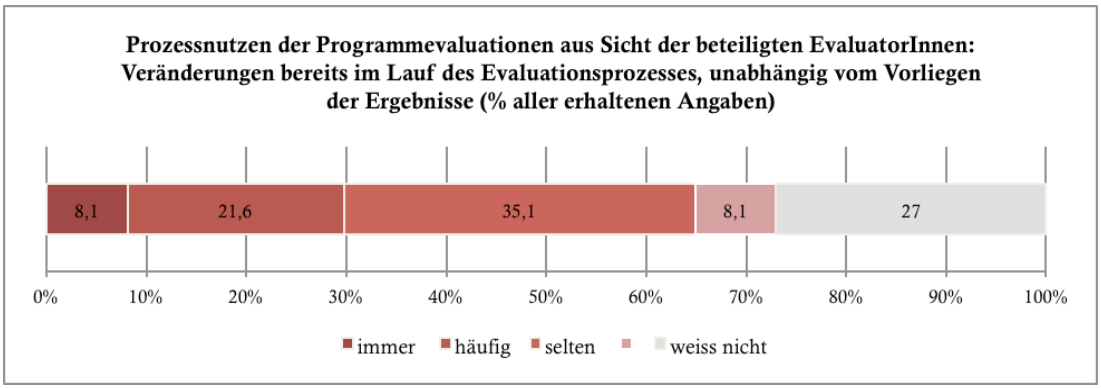
Prozessnutzen und organisatorische Anpassungen

Unter Prozessnutzen wird in der Evaluationsforschung und Evaluationstheorie verstanden, dass bereits im Lauf des Evaluationsprozesses Effekte bei AuftraggeberInnen oder anderen in die Evaluation einbezogenen Akteuren eintreten, schon vor Vorliegen der Ergebnisse bzw. unabhängig davon. Dabei kann es sich um kognitive, verhaltensförmige oder organisatorische Veränderungen handeln, die sodann auch Voraussetzungen für weitere Nutzungen im Sinne der bereits dargestellten Nutzungsweisen schaffen.

Die Interviews haben hier nur bedingt Hinweise erbracht, da die GesprächspartnerInnen vor allem auf andere Fragen des Interviewleitfadens unter starkem Bezug auf aktuelle Bedarfslagen eingingen. In einer Agentur äußerte man sich dahingehend, dass Veränderungen bereits während der Laufzeit von Programmevaluationen *„hin und wieder [eintreten, aber], eher die Ausnahme“* darstellen (A1). Zugleich wurde darauf hingewiesen, dass es gemäß den langjährigen Erfahrungen dafür doch gezielter Vorkehrungen bedarf. Die Frage nach unmittelbar eintretenden Effekten im Zuge einer Evaluationsdurchführung wurde in diesem Sinn beantwortet mit:*„Ja, aber nicht als Selbstläufer, der automatisch kommt.“* (A1)

Systematische Einschätzungen entlang der in der Evaluationsforschung gängigen Strukturierung des Phänomenbereichs liegen von den EvaluatorInnen aus der Befragung vor. Nur 8% der antwortenden EvaluatorInnen bezeichnen Veränderungen bereits während der Laufzeit der Programmevaluationen als einen Effekt, der in allen von ihnen durchgeführten Programmevaluationen eingetreten ist. Weitere 22% geben an, dass sie derartige Effekte häufig beobachten konnten. Damit ist Prozessnutzen die am seltensten eingetretene Form der Generierung von Evaluationsnutzen, doch liegt auch sie im Zuge des Aufbaus der Evaluationskultur durchaus vor.

Abbildung 11: Prozessnutzen aus Sicht der EvaluatorInnen



Seit den 1980er-Jahren richtet sich die Aufmerksamkeit der internationalen Nutzungsforschung zur Evaluation verstärkt auch auf Anpassungen in den Institutionen, die den Umgang mit Evaluation erleichtern und unterstützen, sowie auf Beiträge des Evaluierens zu einem Organisationslernen. M.Q.Pattons Entwurf zu einer Nützlichkeits-fokussierten Evaluation (*utilization focused evaluation*) (Patton 1978), die mittlerweile in der vierten überarbeiteten Auflage vorliegt (Patton 2008), gilt als Verkörperung entsprechender evaluationstheoretischer Überlegungen. Auch diese Perspektive wurde in den Interviews mit den AuftraggeberInnen und in der Befragung der EvaluatorInnen verfolgt.

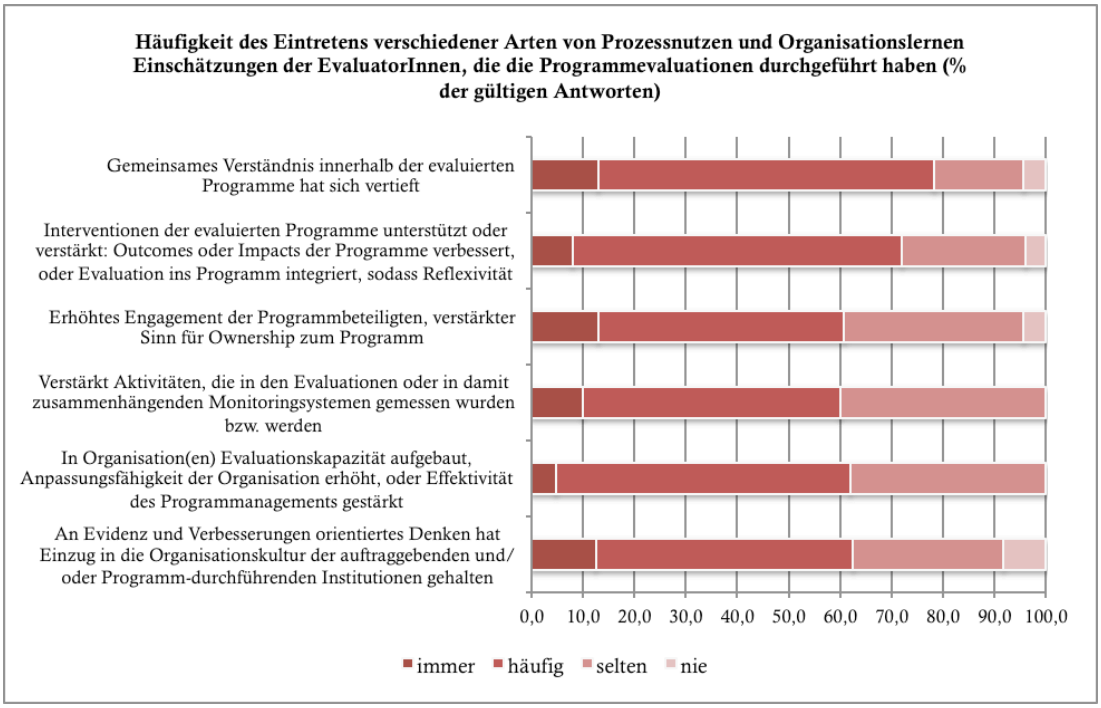
Das wesentlichste Ergebnis besteht hier zweifellos darin, dass sich zwei Agenturen mit Prozessen ausgestattet haben, die die Evaluationsfunktion innerhalb der Organisation klar verankern. In einem Fall wurde kürzlich ein spezifischer Prozess geschaffen, der die gezielte Auseinandersetzung mit Evaluationsergebnissen auf Strategieebene und die Umsetzung von als relevant erkannten Schritten auf operativer Ebene auch über das jeweils evaluierte Programm hinaus gewährleisten soll. Diese Verankerung der Evaluationsfunktion im Prozessmanagement stellt eine lernbasierte Reaktion auf den als unbefriedigend erkannten Vorzustand dar, dass Evalutionsergebnisse nur von Wenigen, oder nur einer einzigen Person auf Detailebene zur Kenntnis genommen wurden, und dass kaum für deren Weitervermittlung Vorsorge getragen wurde. Im zweiten Fall ist die Evaluationsfunktion mit allen Organisationsteilen einschließlich der Entscheidungsgremien verknüpft. Hier wird allerdings darauf hingewiesen, dass die Wahrnehmung der Evaluationsfunktion dadurch auch ein Stück weit von den handelnden Personen abhängig bleibt, sodass nicht vollkommen sichergestellt ist, dass sie nicht in

Zukunft im Zuge von größeren Veränderungen in Präferenzsystemen unter Umständen wieder geschwächt werden könnte.

Für die anderen Institutionen ist eine derartige strukturelle Verankerung des Umgangs mit Evaluationen, von der Planung bis zur Auseinandersetzung mit den Ergebnissen, nicht zu beobachten. Während in allen Ressorts ExponentInnen der FTI-Evaluation ihre Häuser in der *Plattform fteval* vertreten, sind sie doch nicht EvaluationsspezialistInnen in dem Sinne, dass sie sich ausschließlich dieser Materie widmen würden. Programmevaluationen stellen, von der Planung bis hin zum Umgang mit den Ergebnissen, eine Nebentätigkeit im Rahmen von fachlichen Zuständigkeiten für Programme dar. Damit kann die strukturelle Situation dahingehend beschrieben werden, dass die Evaluationsfunktion innerhalb der Institutionen verteilt ist, anstatt zentral angelegt zu sein, und nur in unklarer bzw. unsicherer Weise so zusammenläuft, dass es zu übergreifenden Nutzungen in der gesamten Institution kommen kann. Zur Frage der organisatorischen Anpassung an die Handhabung von Evaluation und ihren Ergebnissen wird so etwa festgestellt: *„Es gibt Luft nach oben, das gibt es sicher. Aber es ist sehr viel passiert in den letzten zehn Jahren." (M1)*

Den EvaluatorInnen wurde im Weiteren eine Frage gestellt, die Komponenten des Prozessnutzens bzw. organisatorischen Wirkungen auf und Anpassungen von Organisationen an Evaluation gemäß dem Konzept von M.Q.Patton aufschlüsselt (Patton 2007). Die EvaluatorInnen schätzen alle diese Nutzen-Aspekte dahingehend ein, dass sie in den verschiedenen Evaluationsprozessen „selten" bis „häufig" eingetreten sind.[3] Am ehesten haben die Programmevaluationen dazu beigetragen, dass sich unter den verschiedenen an ihrer Konzeption und Umsetzung beteiligten Akteuren ein gemeinsames Verständnis der evaluierten Programme vertieft hat. Dies findet eine Entsprechung in der folgenden Aussage aus den geführten Gesprächen zu Auswirkungen des Evaluierens auf die Umgangsweise mit Programmen im Rahmen der Principal-Agent-Beziehungen: *„Tendenziell trägt es zu einer Harmonisierung der Vorstellungen zu gewissen Dingen bei." (A1)* Es ist also berechtigter Weise anzunehmen, dass die Durchführung von Programmevaluationen immer wieder dazu beigetragen hat, dass Programmeigentümer und umsetzende Agenturen „am selben Strang ziehen".

Abbildung 12: Prozessnutzen und Organisationslernen aus Sicht der EvaluatorInnen



[3] Die erhaltenen Angaben der EvaluatorInnen beziehen sich sowohl auf längerfristig und übergreifende Wirkungen der Evaluationstätigkeit als auch auf unmittelbare Wirkungen einzelner Programmevaluationen. Im Vergleich mit der im Vorigen schon erläuterten Frage, inwiefern sie unmittelbare Auswirkungen schon während der Laufzeit der Evaluationen beobachtet haben, ergeben sich für die verschiedenen Items der hier behandelten Fragestellung positive Differenzen von 9% bis 21%.

Weitere positive Effekte, die in geringerem Ausmaß beobachtet werden, können der Abbildung 12 auf der vorigen Seite entnommen werden. Aufgegriffen werden soll hier noch speziell der Gesichtspunkt, inwiefern im Rahmen der bisherigen Evaluationspraxis ein evaluatives, also an Evidenz und Verbesserungen orientiertes Denken in die Organisationskultur der mit den Programmen befassten Institutionen Einzug gehalten gehalten hat. Nur 6% der EvaluatorInnen sind der Ansicht, dass dies nie erreicht wurde. Andererseits sind jedoch auch nur 9% der Ansicht, dass dies stets erreicht werden konnte. Das kann als Hinweis darauf gelten, dass die Einbettung der Evaluationsfunktion in die Institutionen noch verbessert werden kann, um die Orientierung an und Nutzung von kritischer und verbesserungsorientierter Evidenz noch tiefer und breiter zu verankern.

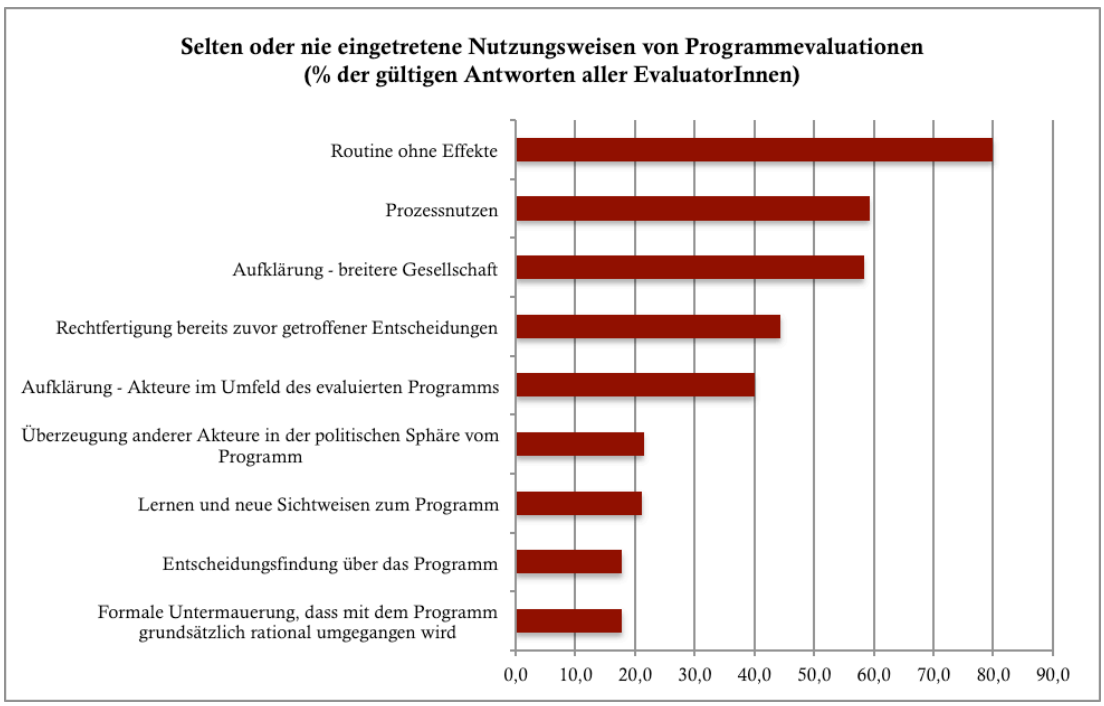
Lernen aus den Erfahrungen mit früheren Evaluationsprozessen liegt auch eindeutig bei der Identifikation von Herausforderungen an Programmevaluationen im Sinne ihrer Nützlichkeit vor, auf die in Kapitel 5 eingegangen wird. Andererseits spricht aus vielen Aussagen und Stellungnahmen, dass der Ertrag der Programmevaluationen aus Sicht der AuftraggeberInnen auch höher sein hätte könnte. Ein derartiger Ermüdungseffekt muss mit den in der Vergangenheit angewendeten Analysekonzepten, den auf deren Grundlage erbrachten Ergebnissen, und den resultierenden Möglichkeiten, substanzielle und zugleich gut gesicherte Schlussfolgerungen und Empfehlungen abzuleiten, in ursächlichem Zusammenhang gesehen werden. Entsprechende Gesichtpunkte werden in Kapitel 4 anhand der DeGeval-Standards analysiert.

Schlussbetrachtung

Die beschriebenen Nutzungsweisen erscheinen in erheblichem Maß als Schnittmengen von Nutzungsformen, die die Nutzungsforschung zu analytischen Zwecken aufgliedert, sodass es augenscheinlich zwar mehr oder weniger genutzte Evaluationen gibt, aber nicht eine klare Trennlinie zwischen „der nützlichen Evaluation“ und „der unnützen Evaluation“. Eine Gesamtdarstellung der Einschätzungen der EvaluatorInnen vermittelt nochmals die Relationen zwischen den Nutzungsweisen im Gesamtüberblick. Dabei wird die Perspektive gegenüber den vorangegangenen Darstellungen umgekehrt. Die folgende Abbildung zeigt, inwiefern bestimmte Arten des Nutzens aus der Sicht derjenigen EvaluatorInnen, die sich zu einer Einschätzung in der Lage sehen, nur selten oder nicht eingetreten sind.

In diese Gesamtbetrachtung ist auch die ebenfalls gestellte Kontrollfrage integriert, ob Programmevaluationen aus Sicht der EvaluatorInnen lediglich routinemäßig verarbeitet wurden, ohne dass eine Nutzung erkennbar geworden wäre. Aus Sicht der EvaluatorInnen ist dies nur in geringem Maß der Fall gewesen, wenn auch eine Quote an reinen „Plichtübungen“ bleibt, die auch durch Aussagen von AuftraggeberInnen bestätigt wird.

Abbildung 13: Selten oder nie eingetretene Nutzungsweisen von Programmevaluationen aus Sicht der EvalutorInnen



3. Einflussfaktoren auf die Nutzung der Programmevaluationen

Die Einschätzung von Einflussfaktoren auf die Evaluationsnutzung stützt sich wegen des Charakters der in den Erhebungsverfahren verfügbar gewordenen Daten in erster Linie auf Ergebnisse aus der EvaluatorInnen-Befragung, in der insgesamt 42 Faktoren erhoben wurden, die in der internationalen Nutzungsforschung als wesentlich oder potenziell bedeutsam gelten (Cousins/Leithwood 1986, Johnson 2009, Fleischer/Christie 2009). Die in der Umfrage erhaltenen Ergebnisse werden sodann auch mit Ergebnissen aus anderen Schritten der Metaevaluation im Beziehung gesetzt, anhand derer die Befragungsergebnisse interpretiert oder erhärtet werden können. Erhaltene Auskünfte von AuftraggeberInnen zu Evaluationseigenschaften, die mit einer Entstehung von Nutzen aus den Evaluationen in Verbindung gebracht werden, beziehen sich auf Defizite, die in Einzelfällen die Nutzung behindert haben oder pauschal als Hinderungsfaktoren für eine Nutzung begriffen werden. Insgesamt stand für die AuftraggeberInnen in der offenen Gesprächsführung anhand des Interviewleitfadens die strukturelle Situation für die Planung und Verwertung von Programmevaluationen im Vordergrund, auf die in Kapitel 5 eingegangen wird. In der folgenden Darstellung bilden die Ergebnisse der EvaluatorInnen-Befragung das Gerüst der Ergebnispräsentation, wobei Befragungsdaten da, wo auch Auftraggeber-Aussagen vorliegen, mit diesen in Beziehung gesetzt werden.

Die EvaluatorInnen wurden in zwei Fragen um Einschätzungen gebeten, welchen Einfluss verschiedene Faktoren ihren Erfahrungen nach darauf gehabt haben, dass von ihnen durchgeführte Evaluationen in größerem oder geringeren Maß Nutzen generiert haben. Eine Frage zielte auf Faktoren, die in dieser Nutzungsforschung und in evaluationstheoretischen Arbeiten als dem Verantwortungsbereich der EvaluatorInnen zugehörig betrachtet werden. Eine zweite Frage zielte auf Kontextfaktoren, für die davon auszugehen ist, dass EvaluatorInnen sie durch ihre Vorgehensweisen nicht beeinflussen können. Beide Fragen wurden mit einigen Faktoren angereichert, die Spezifika von Evaluationen im FTI-Bereich berücksichtigen und abbilden können.

Neben der Einstufung des Einflusses der Faktoren wurde den EvaluatorInnen auch die Möglichkeit geboten, den jeweiligen Faktor als unbekannt oder unzutreffend einzustufen. Bei der Einschätzung der Faktoren kamen zum Teil beträchtliche Quoten an Antwortenthaltungen zustande. Die Nichtantworten bewegen sich für die verschiedenen Faktoren zwischen 6,3% und 34,4%. Dabei sind es nicht nur Kontextfaktoren, zu denen die EvaluatorInnen in größerem Ausmaß mit „weiß nicht/trifft nicht zu“ antworten. Auch unter denjenigen Faktoren, die im Bereich der Evaluationsplanung, -durchführung und –präsentation angesiedelt sind, enthält sich stellenweise ein Drittel der EvaluatorInnen einer inhaltlichen Angabe.

Die im Folgenden dargestellten Ergebnisse stützen sich auf gültige Antworten, bei denen die Faktoren hinsichtlich ihres Einflusses tatsächlich eingestuft wurden. Es werden diejenigen unter allen abgefragten Faktoren dargestellt, die sich aus Sicht der EvaluatorInnen auf die österreichischen FTI-Evaluationspraxis als die auschlaggebendsten erwiesen haben.[4] Zunächst werden die den Evaluationen inhärenten Einflussfaktoren dargestellt. Im Anschluss wird auf den Einfluss von Kontextfaktoren eingegangen. In einem dritten Schritt werden die beiden Faktorengruppen in einem Gesamtbild miteinander in Beziehung gesetzt.

3.1 Faktoren im direkten Einzugsbereich einer Evaluation

Von vorrangigem Einfluss für das Ausmaß, in dem Evaluationen auch Nutzen erzeugen, ist nach Einschätzung der EvaluatorInnen ihre Glaubwürdigkeit. Die Glaubwürdigkeit bei den AuftraggeberInnen steht an erster Stelle, 63% der antwortenden EvaluatorInnen betrachten sie als sehr einflussreich, und weitere 30% als eher einflussreich. Aber auch die Glaubwürdigkeit der EvaluatorInnen gegenüber anderen einbezogenen Stakeholdern, wie der programmdurchführenden Agentur oder Zielgruppen des untersuchten Programms, hat hohen Stellenwert. Sie steht an dritter

[4] Zur Bestimmung der Rangreihenfolge der Faktoren wurden Mittelwerte herangezogen. Dadurch gehen auch Einstufungen als wenig oder nicht einflussreiche Faktoren in die Bestimmung des Ranges ein. Um die Lesbarkeit zu erleichtern, werden nicht alle Prozentwerte in der textlichen Darstellung der Ergebnisse wiedergegeben. Die genauen Werte, auch für weniger relevante Einflussfaktoren, die in der Konzentration auf die wesentlichsten Ergebnisse hier nicht zur Darstellung gelangen, können der Umfrageauswertung im Anhang 8 entnommen werden.

Stelle unter allen Faktoren (45% „sehr einflussreich“ und 48% „eher einflussreich“). Dies spiegelt unmittelbar eine Situation wieder, in der einige spezialisierte Evaluationsinstitute über Jahre, wenn nicht Jahrzehnte hinweg wiederholt tätig waren und - sofern es sich um österreichische EvaluatorInnen handelt - mit den AuftraggeberInnen in der *Plattform fteval* in ständigem Austausch stehen. Stellt man dem erhaltene Auskünfte der AuftraggeberInnen gegenüber, so relativiert sich die enorme Rolle der Glaubwürdigkeit etwas: Die AuftraggeberInnen signalisieren grundsätzliches hohes Vertrauen gegenüber den EvaluatorInnen, das mit der Notwendigkeit assoziiert ist, FachspezialistInnen heranzuziehen, die Datenlagen und FTI-Segmente genau kennen und damit auch in den Vergabeverfahren mit Terms of Reference konstruktiv umgehen können. Sie weisen aber auch darauf hin, dass Erwartungen an Evaluationsaufträge nicht vollkommen erfüllt wurden, und mehrere GesprächspartnerInnen haben der Ansicht Ausdruck verliehen, dass Ihnen für einen qualitätsvollen und mit Überzeugungskraft ausgestatteten „Blick von außen“ auf die Programme ausländische EvaluatorInnen heute unverzichtbar erscheinen.

Als zweitwichtigster unter allen Einflussfaktoren, so wie die EvaluatorInnen sie einstufen, steht die Klarheit der Berichterstattung. Dies wird ergänzt durch einen hohen Stellenwert der Präsentationsweise der Evaluationsergebnisse bzw. des Stil der Berichterstattung, den – an sechster Stelle der Skala - immerhin noch 39% als sehr einflussreich erachten. Aus der Perspektive der Evaluationsstandards entspricht die Aufmerksamkeit für diesen Gesichtspunkt einer guten Erfüllung des Standards N6 „Vollständigkeit und Klarheit der Berichterstattung“. AuftraggeberInnen haben angemerkt, dass Evaluationsberichte teilweise in ihrer Berichterstattung über Datenlagen nicht klare Aussagen getroffen haben, und unterstreichen die Rolle eines aussagekräftigen Executive Summary in der mit knappen Zeitressourcen konfrontierten administrativ-politischen Akteursarena.

In etwa gleichauf mit der Klarheit und Präsentationsweise der Berichterstattung rangiert der Zuschnitt der Evaluation auf Informationsbedürfnisse der Entscheidungsträger über das untersuchte Programm. Insgesamt 89% der antwortenden EvaluatorInnen bezeichnen dies als sehr oder zumindest eher einflussreich auf die Entstehung von Nutzen. Wie andere Ergebnisse der Metaevaluation zeigen, ist eine klare Orientierung von Evaluationsfragen an konkreten Informationsbedürfnisse immer wieder nur bedingt gegeben, da die Evaluationszwecke und Evaluationsfragestellungen zwischen unterschiedlichen Bedürfnissen der Akteure an den verschiedenen Systemstellen und zwischen den Evaluationszwecken des Lernens und der Rechenschaftslegung „aufgespannt“ sind (vgl. Kapitel 1 und Kapitel 5).

Eine zeitgerechte Übermittlung der Evaluationsergebnisse im Verhältnis zu konkreten Entscheidungsfindungsprozessen wird von 38% der antwortenden EvaluatorInnen als sehr einflussreicher Aspekt erachtet. Hingegen meinen nur 19%, dass dies in gleicher Weise für eine zeitgerechte Übermittlung der Evaluationsergebnisse gemäß den vertraglichen Vereinbarungen gilt. Es tritt damit eine Differenzierung zwischen Subaspekten der Zeitgerechtheit zu Tage, die auch in Aussagen der InterviewpartnerInnen bestätigt wird: *„Rechtzeitigkeit der Berichtsabgabe hat nichts mit der Relevanz für die Weiterentwicklung des Programms zu tun.“ (A1)* Der Erkenntniswert der Studien wird hier über formale Kriterien gestellt, und die im Auftraggeberbereich eingenommene Perspektive auf Evaluation ist eindeutig diejenige der Gewinnung guter Entscheidungsgrundlagen. Behinderungen einer Evaluationsnutzung, die sich aus einem Mangel an Zeitgerechtheit ergeben hätten, wurden von keiner/m InterviewpartnerInnen erwähnt.

Nicht beantwortet ist mit den vorliegenden Daten allerdings die Frage, wie oft Programmevaluationen zur direkten Informierung bestimmter Entscheidungsfindungsprozesse konzipiert und durchgeführt wurden. Die Informationen aus den Interviews deuten darauf hin, dass es hier am ehesten um die Verfügbarkeit von Evaluationsberichten zu vorprogrammierten Zeitpunkten gehen dürfte, die sich im Rahmen der Vereinbarungen von Fachressorts mit dem Finanzministerium ergeben, wobei sich in diesen Prozessen die Entscheidungsfindung auf etwaige Anpassungen der Programmformulierungen bezieht. Eine Auslösung von Programmevaluationen zur direkten Informierung der FTI-Politik angesichts aktuell wahrgenommener Informationsbedürfnissen oder „windows of opportunity“ stellt hingegen in der bisherigen FTI-Evakuationspraxis eine Seltenheit dar. In aller Regel geht es eher darum, die Gelegenheiten der vorprogrammierten Programmevaluationen auch dafür zu nutzen, aktuelle FTI-politische Fragestellungen „mitzunehmen“.

Von einigem Einfluss erscheint den EvaluatorInnen des Weiteren Objektivität (an 7. Stelle) und eine ausgewogene Darstellung von Stärken und Schwächen des untersuchten Programms (an 8. Stelle). Auf diese wesentlichen Merkmale einer qualitätsvollen Evaluation wird in der Evaluationspraxis des

FTI-Bereichs offensichtlich über weite Strecken geachtet. Der Einfluss dieser Gesichtspunkte auf das Zustandekommen von Nutzungen von Evaluationsergebnissen weist gegenüber den bereits genannten, an der Spitze der Rangreihenfolge platzierten Faktoren allerdings deutliche Niveauunterschiede auf. Gegenüber 63% der EvaluatorInnen, die die Glaubwürdigkeit bei den AuftraggeberInnen als sehr einflussreich bezeichnen, und 58%, die die Klarheit der Berichterstattung als sehr einflussreich einstufen, sind es hinsichtlich der Objektivität 42%, und hinsichtlich der Ausgewogenheit der Darstellung von Stärken und Schwächen des untersuchten Programms 22%. Die unterschiedliche Einschätzung des Einflusses der beiden Gesichtspunkte deutet darauf hin, dass „Objektivität" nicht immer mit jener Ausgewogenheit in der Darstellung von Stärken und Schwächen gleichgesetzt wird, die mit dem Fairnessstandard F3 gemeint ist. Von AuftraggerInnen wurde eine Schwäche mancher Evaluationen darin verortet, dass Sichtweisen bestimmter Akteursgruppen von den EvaluatorInnen unmittelbar übernommen wurden, was auch in der im Rahmen der vorliegenden Studie durchgeführten Berichtsanalyse wiederholt festzustellen war.

Ebenfalls unter den zehn aus Sicht der EvaluatorInnen einflussreichsten Faktoren, wenn auch etwas nachrangiger, befinden sich schließlich die Angemessenheit der Evaluationskriterien und die Art des Evaluationsansatzes. Mit diesen Gesichtspunkten werden wesentliche Weichenstellungen für die Gesamtvorgehensweise einer Evaluation angesprochen. Diese aus evaluationstheoretischer Sicht übergeordneten Gestaltungsaspekte rangieren in den Einschätzungen der EvaluatorInnen noch vor Methodenaspekten im engeren Sinn. Dies stellt im Sinne der Evaluationsstandards ein positives Ergebnis dar. Nicht ausgedrückt ist mit den genannten Daten, ob die Evaluationsätze und –kriterien, die zum Einsatz kamen, auch die besten denkbaren Lösungen darstellten. Es sollte davon ausgegangen werden, dass es hier um Konsensbildungen zwischen AuftraggeberInnen und EvaluatorInnen über die Anlage der Programmevaluationen geht, für die in einem konkreten Evaluationsprojekt die Ausschreibungsunterlagen (Terms of Reference) die Grundlage bilden, und in die auch Vorkommunikationen in der *Plattform fteval* und in Vorgängerprojekten mit einfließen. Was in den Daten in erster Linie zum Ausdruck kommt, ist, dass dieser Konsens zu den wesentlichen Erfolgsbedingungen einer Programmevaluation zählt, wobei hier ebenso vermeintliche Selbstverständlichkeiten zum tragen kommen können, wie Grenzen für geschätzte bzw. akzeptierte Vorgehensweisen existieren können.

Einfluss auf den Nutzen, den eine Programmevaluation entfaltet, sprechen die EvaluatorInnen Planungen hinsichtlich der Nutzung der Evaluation zu, die schon in einer Frühphase des Evaluationsprojekts vorgenommen werden. Dies entspricht zentralen Forderungen der Nützlichkeitsstandards. Dieser Faktor rangiert allerdings erst im Mittelfeld aller verschiedenen Einflussfaktoren (auf Platz 12). Es ist zugleich zu bemerken, dass derartige Planungen den untersuchten Evaluationsberichten, abgesehen von groben Umrissen einer lernorientierten Vorgehensweise (wie etwa „Überprüfen, ob das Programm angepasst werden sollte"), nicht zu entnehmen sind (vgl. Kapitel 4, Standards N2 „Klärung der Evaluationszwecke" und N8 „Nutzung und Nutzen der Evaluation"). Das kann unter Umständen darauf zurückzuführen sein, dass derartige Vorkehrungen nicht für berichtenswert gehalten wurden. Allerdings wurden frühzeitige Planungen von Nutzungsprozessen auf einem detaillierteren Niveau auch in den Gesprächen mit den AuftraggeberInnen und HauptadressatInnen der Programmevaluationen kaum sichtbar.

Interessant erscheint in diesem Zusammenhang, dass ein weiterer abgefragter Aspekt, der mit der Schaffung von Voraussetzungen für Nutzungen vor allem in breiteren Sphären, die über den engen Kreis der unmittelbar für das evaluierte Programm Zuständigen hinausreichen, als unwesentlichster Faktor unter allen eingestuft wird. Die Breite der Dissemination von Ergebnissen wird von keiner/m einzigen der antwortenden EvaluatorInnen als „sehr einflussreich" erachtet, um Nutzen zu generieren. Dies deutet darauf hin, dass Evaluationsnutzung bislang primär im engen Rahmen einer weitgehend geschlossenen Sphäre von unmittelbar mit dem untersuchten Programm Befassten gedacht und verstanden wird. In den Interviews mit den AuftraggeberInnen bestätigt sich diese Diagnose, indem von ihnen beschrieben wird, dass breitere Disseminationsprozesse in der Vergangenheit kaum vorgelegen haben (vgl. Kapitel 1).

Mehrere Faktoren, die den Methodeneinsatz zur Gewinnung und Analyse von Daten betreffen, rangieren ebenfalls im Mittelfeld der Skala. Methodenvielfalt wird dabei als vergleichsweise am Einflussreichsten erachtet, und dem Einsatz von in FTI-Evaluationen üblichen Methoden als

vergleichsweise wenig Gewicht zugemessen.[5] Dies spricht für eine Flexibilität der Evaluationskonzeptionen, die nicht auf vorgefasste Herangehensweisen festgelegt sind und nach dem für den Einzelfall bestmöglichen Methoden suchen. Im Sample der analysierten Berichte, die immerhin rund die Hälfte aller seit 2003 publizierten Programmevaluationen ausmachen, finden sich allerdings nur zwei Programmevaluationen, die ungewöhnliche bzw. innovative Methoden zur Anwendung gebracht haben.

Unter den als am Wenigsten einflussreichen Faktoren wird des Weiteren die Finesse der Methodenanwendung eingestuft. Immerhin 27% der antwortenden EvaluatorInnen bezeichnen sie als „gar nicht einflussreich", und weitere 32% als „eher nicht einflussreich". Es wird erkennbar, dass die oft angemerkte grundsätzliche Methodenzentriertheit der FTI-Evaluation nicht nur hinsichtlich des Gesamtstellenwerts von mehreren Methoden-relevanten Faktoren in einem breiteren Rahmen auch andersgelagerter Faktoren Grenzen findet, wenn es um die Nutzung der erbrachten Ergebnisse geht, sondern auch hinsichtlich ihrer fachlich-wissenschaftlichen Präzision und Verfeinerung. Dieses Ergebnis steht im Gegensatz zu solchen, die im US-amerikanischen Raum übergreifend für Evaluationen in verschiedenen Politikbereichen erbracht (Cooksy/Caracelli 2005). Manche AuftraggeberInnen haben auf Schwächen im Methodischen hingewiesen, indem sie z.B. die „Sauberkeit" einer Evaluation eingemahnt oder auf Schwächen im Umgang mit qualitativen Daten hingewiesen haben.

Am untersten Ende der Skala stehen schließlich zwei Faktoren, die mit der Tragfähigkeit und Aussagekraft der Untersuchungen und ihrer Schlussfolgerungen zu tun haben. Dies ist einerseits die Beschränkung der Analyse auf Aspekte, für die ausreichend gesicherte Daten vorlagen bzw. im Rahmen der gegebenen Ressourcen erhoben werden konnten. Der zweite, gegensinnige Faktor ist die Durchführung einer möglichst umfassenden Analyse, auch wenn im Rahmen der gegebenen Ressourcen nicht für alle behandelten Aspekte Daten vorlagen bzw. erhoben werden konnten. Während eine Beschränkung auf gesicherte Datenlagen noch von 16% der antwortenden EvaluatorInnen als „sehr einflussreich" erachtet wird, sind es bei der Durchführung einer nicht in allen Hinsichten gut gestützten Analyse lediglich 4%. Diese Beobachtung weist darauf hin, dass Programmevaluationen in der Vergangenheit offenbar nur bedingt im Bewusstsein genutzt – und unter Umständen auch angestellt - wurden, dass es sich um Beweisführungen wissenschaftlicher Bauart handeln sollte, die die in den Evaluationsberichten getätigten Aussagen auf erhärtete Fakten gründet. Es ist auf Basis einer vergleichenden Betrachtung der Ergebnisse umso mehr naheliegend, anzunehmen, dass sich das Verständnis von Objektivität in der Vergangenheit über Strecken vor allem darauf bezogen hat, dass überhaupt objektive Daten analysiert und dargestellt werden. In der Berichtsanalyse zeigt sich, dass Programmevaluationen auch Züge von Expertengutachten tragen, in denen persönliche Wissensstände und Sichtweisen zur Geltung gebracht werden. Ähnliche Ergebnisse wurden freilich auch für die Evaluationspraxis in bestimmten Politikbereichen der Schweiz erbracht (Lehmann/Balthasar 2004).

3.2 Kontextfaktoren

Im Folgenden werden die 10 wesentlichsten Faktoren, die durch die Gestaltung einer einzelnen Programmevaluation nicht beeinflusst werden können, erläutert. Als einflussreichsten Faktor erachten die antwortenden EvaluatorInnen hier die Erwartung der Auftraggeberinnen, dass die Evaluation ihnen und ihren Vorhaben von Nutzen sein wird. 62% halten diese Erwartungen für „sehr einflussreich", damit es zu Nutzungen der Evaluation und ihrer Ergebnisse kommt, und weitere 35% für „eher einflussreich". Damit wird Licht auf den Umstand geworfen, dass nicht alle Programmevaluationen in gleicher Weise von ihren jeweiligen AuftraggeberInnen mit hohen Nutzenerwartungen verbunden werden müssen. Erhaltene Aussagen in den Interviews deuten in der Tat darauf hin, dass hier doch zum Teil gewichtige Unterschiede vorlagen (vgl. dazu auch Kapitel 2).

An zweiter Stelle, mit bereits deutlichem Unterschied in der von den EvaluatorInnen eingeschätzten Bedeutung, steht ein direkter Konnex der Evaluationen mit einem aktuellen Entscheidungsbedarf oder

[5] Hinsichtlich von Methodenaspekten wurden insgesamt fünf Faktoren erhoben. Dies sind in absteigender Reihenfolge des Einflusses auf eine Evaluationsnutzung aus Sicht der EvaluatorInnen: Methodenvielfalt, Genauigkeit in der Methodenanwendung und Datenanalyse, Beleuchtung bestimmter Fakten durch mehrere parallel eingesetzte Methoden, Einsatz von in FTI-Evaluationen gebräuchlichen Methoden, Finesse der Methodenanwendung.

Problemdruck. Die Wahrscheinlichkeit, dass eine Evaluation auch genutzt wird, hängt – wenig überraschend – stark davon ab, ob sie innerhalb vorgegebener Planungen eher routinehaft abläuft, oder mit aktuellen Problemwahrnehmungen und Herausforderungen an das Handeln der FTI-politischen Akteure verknüpft ist. Auch dies wird in dem Sinn zu interpretieren sein, dass ein Eingehen auf aktuell wahrgenommene Herausforderungen für die an Evaluationsplanungen beteiligten Akteure an den verschiedenen Systemstellen bisweilen nur bedingt möglich war.

Bereits an dritter Stelle rangieren persönliche Sichtweisen und Denkstile des/r direkten Auftraggeberin. Immerhin 37% erachten sie aus ihren Erfahrungen heraus als „sehr einflussreich", und weitere 48% als „eher einflussreich". Des Weiteren schätzen es 33% der antwortenden EvaluatorInnen als „sehr einflussreich" ein, ob die vorgelegten Evaluationsergebnisse mit Sichtweisen und Erwartungen seitens der Auftraggeberinnen konsistent sind. Weitere 52% sprechen hier von einem „eher einflussreichen" Faktor.

Es sind darüber hinaus auch Erwartungen von Stakeholdern des Programms, dass die Evaluation ihnen und ihren Vorhaben von Nutzen sein wird, die die Wahrscheinlichkeit des Eintretens von Nutzen aus den Programmevaluationen deutlich beeinflussen. Geht man davon aus, dass bei der Beantwortung dieser Frage in erster Linie an VertreterInnen der mit den Programmumsetzungen betrauten Agenturen bzw. im Fall der autonomen Agentur FWF an VertreterInnen des Wissenschaftsressorts gedacht wurde, so ist diese Einschätzung konsistent mit dem herausragenden Stellenwert der Erwartungen der direkten AuftraggeberInnen. Die doch vorhandene Differenz zwischen den beiden Gesichtspunkten (33% „sehr einflussreich" für die Erwartungen der Stakeholder gegenüber 62% „sehr einflussreich" für die Erwartungen der AuftraggeberInnen) machen zugleich ersichtlich, dass es in den Principal-Agent-Beziehungen in der Regel einen tonangebenden Teil gibt und die jeweils an Evaluationen gerichteten Erwartungen nicht vollkommen identisch sind. Es ist darüber hinaus durchaus denkbar, dass EvaluatorInnen bei der Beantwortung der Frage auch an Stakeholder aus der institutionellen Umgebung des Programms oder im Bereich von Zielgruppen der evaluierten Programme (z.B. Fachverbänden, wissenschaftliche Einrichtungen) gedacht haben, die z.B. in einer Steuerungsgruppe für eine Programmevaluation einbezogen waren. Dies entspricht den Prinzipien des Standards N1 „Identifizierung der Beteiligten und Betroffenen" und verweist auf die Frage, wer in einem Evaluationsprojekt in der Planungsphase wie gut eingebunden wird, um Informationsbedürfnisse zu klären und realistische Erwartungen an die Evaluation zu erzeugen.

An sechster Stelle der Skala stehen die Ressourcen und organisatorische Anpassungen, die in den Auftraggeber-Organisationen und den in den Programmevaluationen miteinbezogenen Organisationen für die Verarbeitung von Evaluationsergebnissen vorhanden sind. Nur 20% der antwortenden EvaluatorInnen sind der Ansicht, dass diesen Faktoren der institutionellen Einbettung der Evaluationsfunktion eher keine oder gar keine Bedeutung zukommt, wenn es darum geht, ob und wie sehr Programmevaluationen genutzt werden. Ebenfalls unter den zehn einflussreichsten Faktoren, wenn auch mit vergleichsweise etwas geringerer Bedeutung, befinden sich die Erfahrung der auftraggebenden und einbezogenen Organisation mit Evaluation, sowie die Erfahrung der AuftraggeberIn als Person mit Evaluation. Diese Einschätzungen erschienen vor allem in restrospektiver Hinsicht interessant, da heute in allen relevanten Institutionen umfangreiche Erfahrungen vorliegen, die über die letzten beiden Jahrzehnte aufgebaut wurden. Mit dieser Expansion des „Unternehmens Evaluation" wurden wertvolle Kompetenzen aufgebaut, die gemäß den Erfahrungen der EvaluatorInnen nicht unwesentlich dazu beitragen, dass es zu Nutzungen der Programmevaluationen kommt.

Schließlich messen die antwortenden EvaluatorInnen auch der Wichtigkeit bzw. Tragweite der mit der Evaluation verbundenen Entscheidung einen substanziellen Stellenwert zu, der von immerhin rund drei Viertel als zumindest „eher einflussreich" betrachtet wird. Dies stellt einen zusätzlichen Hinweis zu der bereits dargestellten Beobachtung dar, dass Evaluationen umso eher genutzt werden, als sie sich mit aktuellen Herausforderungen für die entscheidungsverantwortlichen Akteure verbinden. In den Interviews wurde darauf hingewiesen, dass Programmevaluationen dann erhöhte politische Aufmerksamkeit finden, wenn es um „große", übergreifende Themen geht (wie z.B. das Thema Fachhochschulen, das sich durch mehrere Programme durchzieht) oder große Summen im Spiel sind.

An zehnter Stelle steht in den Einschätzungen der EvaluatorInnen die Rolle der direkten Auftraggeberin als Person in ihrer Organisation. Insgesamt wird deutlich, dass innerorganisatorische Strukturen und Befindlichkeiten als zentrale Bedingungen für die Nutzung von

Programmevaluationen im politisch-administrativen FTI-Bereich zu erachten sind, wobei auch der „Human Factor“ – durchaus in Übereinstimmung mit andernorts erbrachten Ergebnissen der Nutzungsforschung zur Evaluation – eine nicht unbeträchtliche Rolle spielt. Erst nach diesen Faktoren rangiert der Reifegrad eines Programms hinsichtlich seiner Evaluierbarkeit.

Wie schon bei den Evaluations-inhärenten Faktoren soll auch hier abschließend der Blick auf das untere Ende der Skale der eingeschätzten Einflussfaktoren auf Evaluationsnutzung geworfen werden. Hier zeigt sich, dass die Nutzungswahrscheinlichkeit gemäß den Erfahrungen der EvaluatorInnen nur in recht untergeordneter Hinsicht durch eine grundsätzlich geringe Neigungen von Entscheidungsträgerinnen, sich auf Evaluationsergebnisse zu stützen, determiniert wird. Es gibt nichtsdestoweniger eine kleine Gruppe von 13% der EvaluatorInnen, die eine grundsätzlich geringe Orientierung von Entscheidungsträgerinnen an Evaluationsergebnissen als „sehr einflussreich“ bezeichnen. Ein grundsätzliches Bekenntnis zu einer Evaluationskultur wäre demnach noch nicht durchgängig an allen Stellen des FTI-politischen Governancesystems eingetreten. Überraschen mag die Tatsache, dass sich unter den am Wenigsten einflussreichen Faktoren auch eine Begleitung der Evaluationen durch Evaluationsmanagerinnen in den Auftraggeber-Organisationen findet (lediglich 4% „sehr einflussreich“). Hier scheint sich eine Herangehensweise an Programmevaluationen auszudrücken, in der zunächst Evaluationsaufträge erteilt und sodann Evaluationsberichte abgenommen werden, ohne während der Evaluationsdurchführung eingehendere Kommunikationen und Interaktionen zwischen EvaluatorInnen und AuftraggeberInnen besonderen Stellenwert zuzumessen, aber auch wenig Ressourcen für ein Evaluationsmanagement zur Verfügung stehen.

3.3 Gesamtbetrachtung

Abschließend werden die insgesamt einflussreichsten Evaluations- und Kontextfaktoren gemeinsam dargestellt, um ihr Verhältnis zueinander einzuschätzen. Hier zeigt sich, dass Faktoren, die innerhalb eines Evaluationsprojekts beeinflusst werden können, und Faktoren, für die dies nicht der Fall ist, einander die Waage halten. Es sind jeweils zehn Faktoren aus den beiden Gruppen, die das Gesamtbild der 20 einflussreichsten Faktoren ausmachen. Diese 20 Faktoren sind in der Abbildung 14 auf der folgenden Seite dargestellt.

Einige genuin evaluationsmethodische Gesichtspunkte wie eine ausgewogene Darstellung von Stärken und Schwächen des untersuchten Programms, die Angemessenheit der Evaluationskriterien und die Art des Evaluationsansatzes fallen im Gesamtbild hinter bedeutendere Einflussfaktoren, die durch die Vorgehensweise einer Evaluation nicht beeinflusst werden können, zurück. Nutzenerwartungen der AuftraggeberInnen und die Glaubwürdigkeit, die die herangezogenen EvaluatorInnen besitzen, dominieren das Bild. Es geht augenscheinlich um Informationen, die in diesem Rahmen im Wechselspiel zwischen dem Informationsbedarf von EntscheidungsträgerInnen und persönlichen Sichtweisen und Bedarfslagen der direkten AuftraggeberInnen Relevanz gewinnen und möglichst klar präsentiert werden bzw. werden sollten. Organisatorische Aspekte im Bereich der Institutionen, die Programmevaluationen in Auftrag geben, kommen bei der Entstehung von Evaluationsnutzen deutlich zum Tragen.

Von den InterviewpartnerInnen im Auftraggeberbereich wurden vor sowohl Kontextfaktoren für die Auslösung und Planung sowie Verwertung von Evaluationen als auch die Qualität von Evaluationsberichten ins Feld geführt, wobei diese frei formulierten Aussagen mit den in der EvaluatorInnen-Befragung vorstrukturierten Einflussfaktoren nicht immer unmittelbar zur Deckung gebracht werden können.[6] AuftraggeberInnen haben in den Interviews positive Gestaltungsmerkmale von Programmevaluationen, die sich in ihren Erfahrungen mit mehr entstandenem Nutzen verbinden, in dieser Form nicht benannt. Dieser Umstand wird auch im Zusammenhang mit dem in der Nutzungsforschung bekannten Phänomen zu sehen seien, dass komplexere, längerfristige und ineinandergreifende Nutzungsweisen von Evaluation von den Beteiligen nur schlecht im Nachhinein einzelnen Evaluationen mit ihren jeweiligen Details zugeordnet werden können. Mitspielen mag auch, dass Erwartungen an „Evaluationsqualität“ streckenweise implizit bleiben. Die Rolle, die

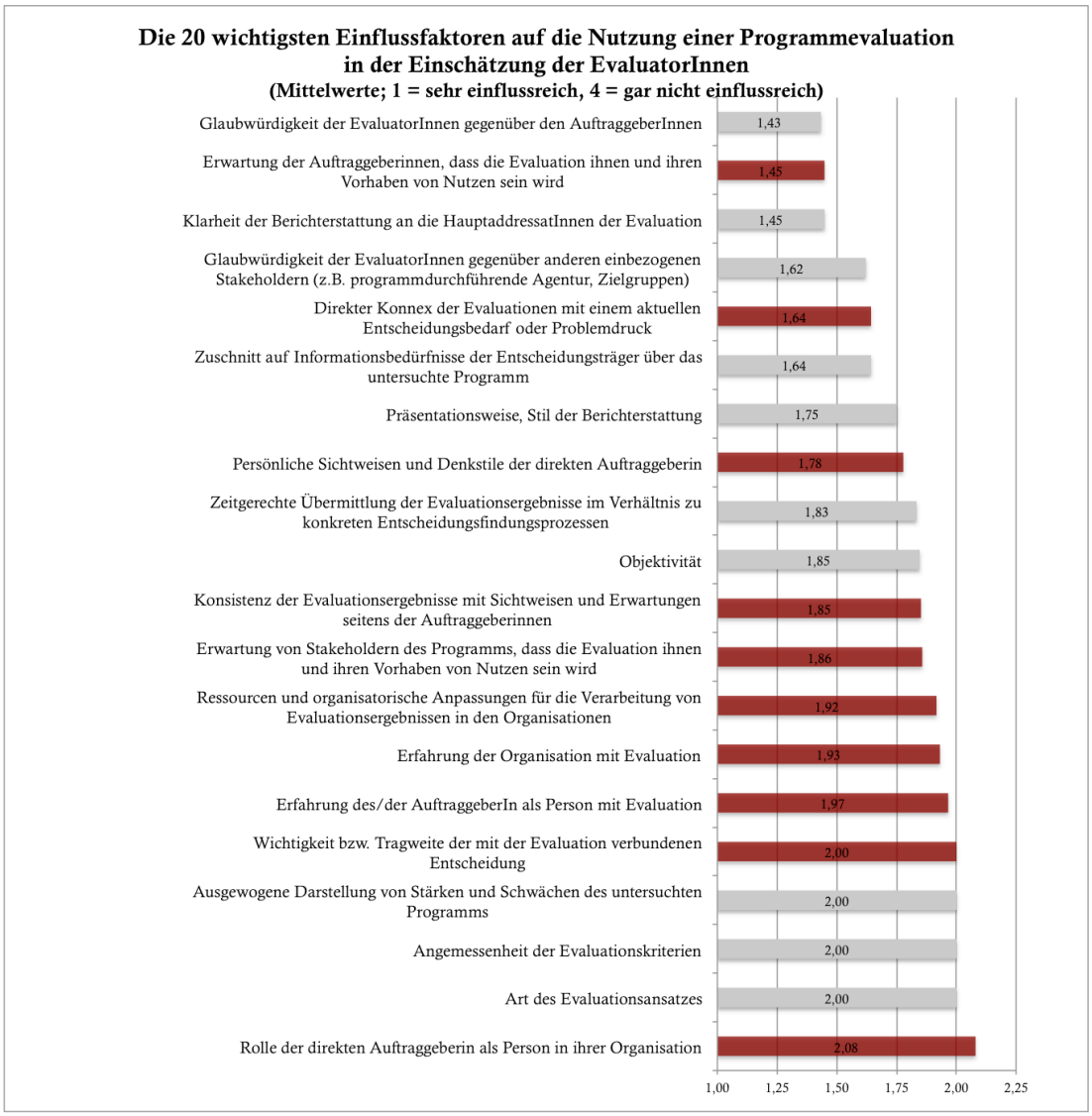
[6] Wenn z.B von teilweise trivialen Ergebnissen die Rede war oder von gelegentlich erhaltenen Empfehlungen, die in den Augen der AuftraggeberInnen die Spezifika des evaluierten Programms nicht gut getroffen haben, so kann dies ebenso auf Beeinträchtigungen der Vermittlung und Kenntnisnahme von Informationsbedürfnissen zurückzuführen sein wie auf den Evaluationsansatz oder Aspekte der Methodenwahl und – durchführung.

einzelnen Evaluationen inhärenten Qualitätsfaktoren aus der Sicht der AuftraggeberInnen zukommt, kann anhand zweier Zitate übergreifend charakterisiert werden: *„Man findet sehr saubere Evaluierungen, wo man auch zu Empfehlungen und Schlussfolgerungen kommt wo man sagt: das könnten wir eigentlich machen, macht Sinn. Das wird man auch bei sagen wir 80% aller Evaluierungen finden." (A2) „Also die Qualität von Evaluierungen kann man sehr wohl diskutieren, sie sind ganz unterschiedlich, es gibt auch wirklich Enttäuschungen, und am Schluss kommen oft Selbstverständlichkeiten." (M2)*

Von verschiedenen InterviewpartnerInnen wurden einige Eigenschaften von Programmevaluationen erwähnt, die sie in ihrer Erinnerung mit einer Beeinträchtigung der Nützlichkeit in Verbindung bringen, oder die sie heute pauschal als Anforderungen an gut nutzbare Evaluationen betrachten. Wenn es sich hier auch um unterschiedliche, partielle Sichtweisen handelt und Vorsicht bei der Verallgemeinerung geboten ist, so kann doch festgestellt werden, dass sich diese Hinweise auf nahezu alle Standards verteilen, die in der Berichtsanalyse herangezogen wurden. Einzelaspekte zur Qualität der durchgeführten Programmevaluationen aus Auftraggebersicht werden - ebenso wie solche aus EvaluatorInnensicht - in Kapitel 4 im Zusammenhang mit den jeweiligen Evaluationsstandards erläutert.

Abbildung 14: Wichtigste Einflussfaktoren auf die Evaluationsnutzung

In einem Evaluationsprojekt gestaltbare Faktoren sind in Grau dargestellt, durch die Gestaltung eines Evaluationsprojekts nicht beeinflussbare Kontextfaktoren in Rot.



4. Nützlichkeit der Programmevaluationen und Berichte im Licht der DeGEval- Standards

Im vorliegenden Kapitel werden die Ergebnisse der Analyse von 20 Evaluationsberichten anhand des herangezogenen Sets von 11 DeGEval-Standards dargestellt. Die durch Stichprobenziehung in einem theoretical sampling-Verfahren ermittelten Evaluationen sind in Anhang 1 aufgelistet.

Für jede Evaluation bzw. den Bericht über sie wurde ein Factsheet erstellt, das die Erfüllung jedes herangezogenen Standards durch eine Einstufung auf einer fünfstufigen Skala bezeichnet und durch einen qualitativen Kommentar näher darstellt. Die numerischen Einstufungen verstehen sich dabei als Erzeugung einer groben Übersicht, die auch in einem Gesamtbild Grundzüge und Entwicklungen leicht erkennen lässt. Den eigentlichen Kern der Berichtsanalyse bildet jedoch die qualitative Analyse, wie jede individuelle Evaluation bzw. der Bericht über sie in spezifischen Weise Empfehlungen und Forderungen der Standards besser oder schlechter entspricht. Diese qualitative Betrachtungsweise bildet die Basis für die Identifikation von Merkmalen, die die österreichiche Evalautionspraxis im FTI-Bereich im zwölfjährigen Beobachtungszeitraum gekennzeichnet haben und die sodann für Schlussfolgerungen und Empfehlungen genutzt wird.

Die Factsheets im Umfang von jeweils ca. 4 Seiten sind dem vorliegenden Bericht in Anhang 2 beigegeben. Die analysierten Evaluationen wurden dabei konzeptgemäß anonymisiert und sind mit aleatorisch vergebenen Nummern sowie dem Zeitabschnitt, in den sie fallen, bezeichnet. Im Sinne der Anonymisierung sowie der Vergleichbarkeit gemäß der Zielsetzung eines übergreifenden Bildes wurden auch die Kommentare zur Erfüllung der Standards in den einzelnen Programmevaluationen bzw. Berichten so gestaltet, dass die jeweils evaluierten Programme durchgehend als „das evaluierte Programm“ bezeichnet werden.

In den folgenden Abschnitten werden die 20 Evaluationen, deren Berichte analysiert wurden, zunächst in allgemeiner Weise charakterisiert. Anschließend wird das Gesamtergebnis der Berichtsanalyse für das gesamte Sample anhand aller herangezogenen Standards in Form der erreichten Einstufungen, wie gut oder weniger gut die Standards erfüllt wurden, dargestellt. Im Weiteren wird auf jeden Standard einzeln eingegangen, wobei eine Gesamtsymptomatik extrahiert wird, die sich bei aller Individualität der einzelnen Evaluationen bzw. Evaluationsberichte quer über die 20 unterschiedlichen Fälle identifizieren lässt. Dabei wird auch die Entwicklung der Erfüllung der Forderungen und Empfehlungen des jeweiligen Standards in der zeitlichen Entwicklung entlang der drei der Metaevaluation zugrunde gelegten Zeitabschnitte 2003-2006, 2007-2010 und 2011-2014 dargestellt.

Die qualitativen Ergebnisse der Berichtsanalyse werden hinsichtlich von Merkmalen der Evaluationsprozesse im FTI-politischen Bereich durch Ergebnisse der EvaluatorInnen-Befragung und der Auftraggeber-Interviews ergänzt. Die Umfrage von EvaluatorInnen wurde so konzipiert, dass gezielt Aspekte der Evaluationsprozesse erhoben wurden, für die damit zu rechnen war, dass sie in den analysierten Berichten nicht bzw. nicht ausreichend zur Darstellung gelangen, um eine gute Einschätzung vornehmen zu können. Befragungsergebnisse, die Komponenten von Evaluationsprozessen erschließen, im Anschluss an die Ergebnisse der Berichtsanalysen zum jeweiligen Einzelstandard dargestellt. Weiters werden Aussagen aus den Interviews im Auftraggeberbereich, die direkt dem Prinzip eines Standards zuzuordnen sind, punktuell herangezogen. Dadurch wird die Sichtweise sowohl der EvaluatorInnen als auch der AuftraggeberInnen bzw. HauptadressatInnen der Evaluationsberichte in ein Gesamtbild einbezogen. Die Ergebnisse aus den ergänzenden Erhebungsschritten der Online-Umfrage und der Interviews bilden die Erfahrungen und Sichtweisen der Akteure zur Evaluationspraxis des zwölfjährigen Untersuchungszeitraums gesamthaft ab und beziehen sich nicht auf das Berichtssample. Sie begreifen auch die Erfahrungen mit Programmevaluationen mit ein, die in der Berichtsanalyse nicht herangezogen werden konnten. Bei der Präsentation von Umfrageergebnissen in den folgenden Abschnitten werden bewusst keine Grafiken eingesetzt, um die Übersichtlichkeit zu erhöhen.

Die Analyse wurde konzeptgemäß unter dem Blickwinkel angestellt, Einschätzungen der herangezogenen Evaluationsberichte bzw. Evaluationen, über die jeweils berichtet wird, zu erbringen, die auf Verbesserungspotenzial der Evaluationspraxis hinweisen können. Wesentlich für die im Anschluss präsentierten Ergebnisse ist, dass die Standards in ihrer Verfasstheit und Intention eine Reflexionsgrundlage darstellen und auch in diesem Sinne genutzt wurden. Bei der angestellten Analyse handelt es sich jedoch nicht um eine Überprüfung, inwiefern die untersuchten Evaluationen

den ihnen zugrundeliegenden Evaluationsaufträgen gerecht wurden, wofür auch essentielle Grundlagen fehlen, da die Standards den Auftragsverhältnissen – im Unterschied etwa zu den SEVAL-Standards in der Evaluationspraxis von Schweizerischen Behörden – nicht zugrunde liegen. Es zeigte sich zudem in der Berichtsanalyse, dass die Voraussetzungen für eine derartige Überprüfung nur mit Einschränkungen gegeben wären, da die Evaluationsberichte die ihnen zugrunde liegenden Aufträge in aller Regel nicht oder nur in Andeutungen darstellen. Auch Angaben, wie mit der konkreten Vorgehensweise der Evaluation die jeweils zugrunde gelegten Informationsbedürfnisse und Erkenntnisinteressen verfolgt wurden, werden oft zu stark vernachlässigt, um eine konsequente Metaevaluation durchführen zu können (vgl. dazu im Folgenden die Ausführungen zum Standard G3 „Beschreibung von Zwecken und Vorgehen").

4.1 Allgemeine Charakterisierung der Programmevaluationen

Eine allgemeine Charakterisierung im ausgewählten Sample von analysierten Evaluationsberichten wird auf drei Ebenen vorgenommen, die international etablierten Herangehensweisen entsprechen. Grundsätzlich existieren verschiedene Möglichkeiten, Evaluationen in ihren Grundzügen zu charakterisieren, indem jeweils unterschiedliche Blickwinkel angelegt werden. In der vorliegenden Metaevaluation wurde der zeitliche Einsatzpunkt der Programmevaluationen, die dominante Evaluationsrolle sowie die Evaluationsschwerpunkte, auf die sich Konzept und Analysestrategie der Programmevaluationen beziehen, als Betrachtungswinkel gewählt. Nähere Ausführungen dazu finden sich in Kapitel 1, wo die konzeptuellen Grundlagen und die Vorgehensweise der Metaevalution erläutert werden.

Evaluationstyp in zeitlicher Hinsicht

Wie die nachfolgende Tabelle zeigt, handelt es sich bei den im Sample befindlichen Programmevaluationen in hohem Ausmaß um Interimsevaluationen. Dies entspricht Grundzügen der Evaluationspraxis im FTI-Bereich, wo Programmevaluationen mit den zu evaluierenden Programmen verkoppelt sind und in der Regel etwa nach zwei Jahren Programmlaufzeit eine Einschätzung erbringen sollen. Drei der im Sample befindlichen Berichte sind Bestandteil einer Reihe von mehreren Evaluationen zum selben Programm, sodass in diesen Fällen von einer begleitenden Evaluation gesprochen werden kann. Sie siedeln sich dabei in recht unterschiedlichen Stadien der Entfaltung und Entwicklung der jeweils evaluierten Programme an. In der rezenten Beobachtungsperiode findet sich eine ex post-Evaluation. Eine ex ante-Evaluation ist im Sample nicht enthalten, was gut der Tatsache korrespondiert, dass solche Evaluationen in der bisherigen Evaluationspraxis eher Seltenheitswert gehabt haben.

Tabelle 3: Analysierte Evaluationsberichte nach Evaluationstyp

	2003-2006 (n = 5)	2007-2010 (n = 5)	2011-2014 (n = 10)	Summe (n = 20)
ex ante-Evaluation	-	-	-	-
Interimsevaluation (Zwischenevaluation)	5	4	7	16
Begleitevaluation	-	1	2	3
ex post-Evaluation	-	-	1	1

Evaluationsrollen

Die älteste und gebräuchlichste Klassifikation von Programmevaluationen unterscheidet zwischen summativen Evaluationen, die zu einem Evaluationsgegenstand eine zusammenfassende Bilanz ziehen, um grundlegende Entscheidungen über den Evaluationsgegenstand zu ermöglichen, und formativen Evaluationen, die die Gestaltung des Evaluationsgegenstandes begleiten und vorrangig auf Verbesserungen zielen. Mit der jeweiligen Intention verbinden sich sodann grundsätzliche Anforderungen an und Möglichkeiten für die Anlage der Evaluationen. Dieses Verständnis hat sich inzwischen dahingehend erweitert, dass Evaluationen zugleich formativ und summativ sein können. Die Einordnung in die Kategorien erfolgt anhand von Berichtsangaben, aus denen die Intention und Stoßrichtung der jeweiligen Analyse zu erkennen ist.

Die im Sample befindlichen Programmevaluationen sind zu nahezu drei Viertel dem Mischtyp der formativ-summativen Evaluation zuzurechnen. Mit der formativ-summativen Orientierung wurden die Programmevaluationen tendenziell der heute weitgehend verankerten Sichtweise gerecht, dass

auch formative Evaluationen wenn möglich eine Orientierung an schon erkennbaren Ergebnissen des evaluierten Programms aufweisen sollten. Sie verkörpern zugleich Versuche, sowohl Gesamteinschätzungen zum Wert der evaluierten Programme als auch Erkenntnisse zu deren Entfaltung zu erbringen, wobei sie sich auf unterschiedliche Datentypen stützten, die von Input und Output bis zu ersten beobachtbaren Impacts reichten. Nur in sechs Fällen war die Rolle der Evaluation entweder auf eine formative oder eine summative Rolle eingegrenzt, indem sie entweder nur zeitnahe Aspekte der evaluierten Programme begleitend thematisierten oder nur Ergebnisdaten heranzogen, die eine abschließende Bilanz bilden lassen.

Tabelle 4: Analysierte Evaluationsberichte nach Evaluationsrollen				
	2003-2006 (n = 5)	2007-2010 (n = 5)	2011-2014 (n = 10)	Summe (n = 20)
formativ	1	1	-	2
summativ	-	-	4	4
formativ-summativ	4	4	6	14

Evaluationsschwerpunkte gemäß OECD DAC Standards

Anhand dieser international weithin gebräuchlichen Klassifikation lassen sich Programmevaluationen dahingehend einordnen, welche logisch und zeitlich gegliederten Ebenen einer Programmanlage und Programmentfaltung thematisiert werden. Daraus ergeben sich jeweils typische bzw. notwendigerweise anzulegende Herangehensweisen. Die OECD DAC-Klassifikation der Evaluationsschwerpunkte ist im Anhang 5 beigegeben.

Die in der folgenden Tabelle dargestellte Verteilung ist das Resultat einer analytischen Einordnung durch den Metaevaluator, in die Berichtsangaben über Evaluationszwecke, Vorgehensweisen der Evaluationen und ihren Methodeneinsatz eingeflossen sind. Die analysierten Evaluationsberichte setzen diese Terminologie nicht oder in unscharfer Weise ein, oft auch in Kombination mit andersartigen Bestimmungen von Evaluationsanlagen oder auch mit Angaben von Datentypen.

Bis auf eine Programmevaluation wurden mindestens zwei Evaluationsschwerpunkte verfolgt. Die häufigste Form stellt eine Erstreckung auf die drei Schwerpunkte von Relevanz, Effektivität und Impact dar. In einem Fall wurde eine umfassende Programmanalyse mit vier Scherpunkten durchgeführt. Der in der OECD DAC-Klassifikation ebenfalls enthaltene Schwerpunkt der Nachhaltigkeit von Programmeffekten kommt nicht vor, da alle im Sample enthaltenen Evaluationen für eine derartige Analyse zeitlich deutlich zu früh angelegt waren.

Tabelle 5: Analysierte Evaluationsberichte nach Evaluationsschwerpunkten gemäß OECD DAC Standards				
	2003-2006 (n = 5)	2007-2010 (n = 5)	2011-2014 (n = 10)	Summe (n = 20)
Effektivität	-	1	-	1
Relevanz, Effektivität	2		2	4
Effektivität, Impact	-	1	2	3
Relevanz, Effektivität, Impact	3	2	6	11
Relevanz, Effektivität, Effizienz, Impact	-	1	-	1

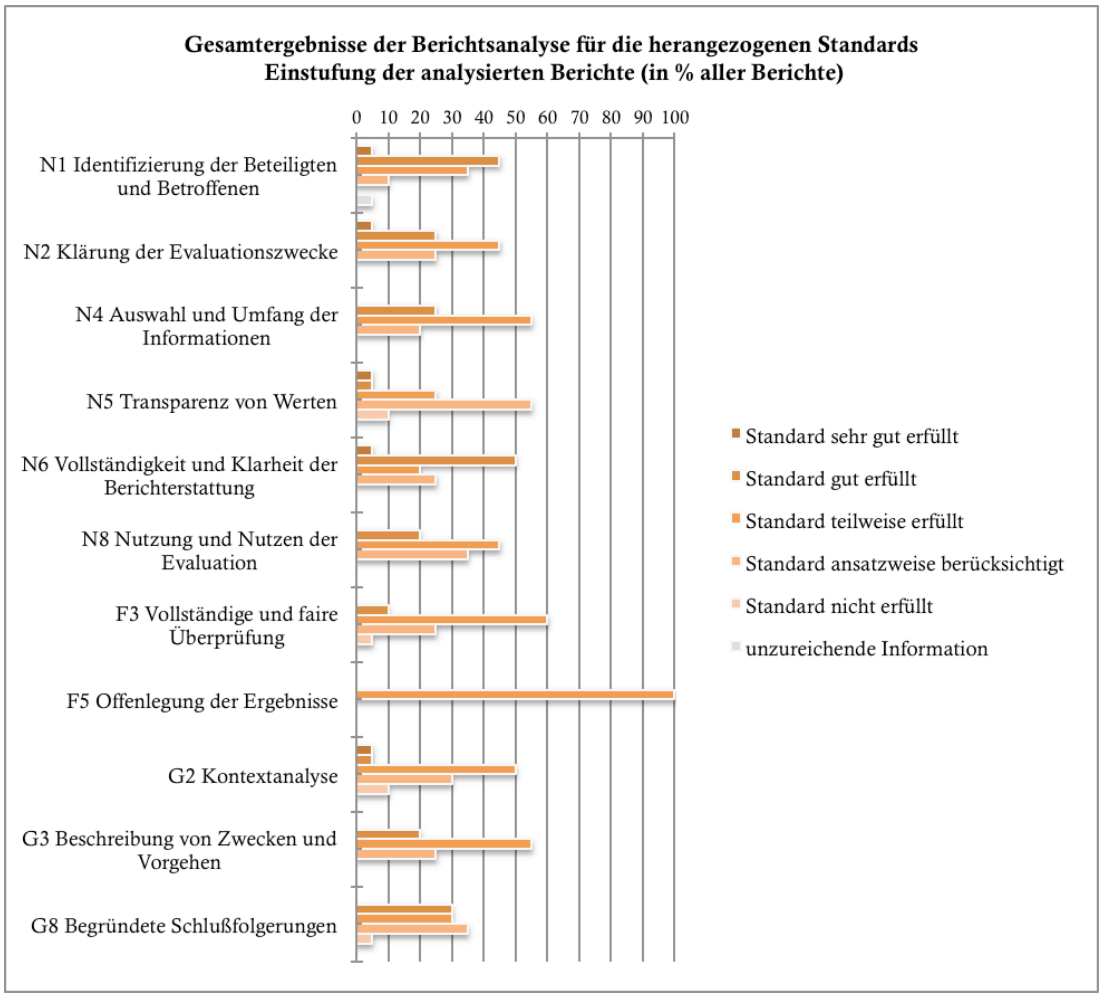
Für die untersuchten Programmevaluationen ist im Gesamtbild festzustellen, dass es sich in hohem Ausmaß um Mehrzweck-Evaluationen (*multi-purpose evaluations*) handelt, die als Begleitevaluationen eine ganze Reihe von Erkenntnissen unterschiedlichen Typs zu erbringen versuchten. Der Großteil der analysierten Evaluationen ist dadurch charakterisiert, dass relativ bald nach dem Programmstart im Sinn einer Konzepterprobung Einschätzungen dazu angestrebt wurden, ob und inwiefern die jeweiligen Programme im weiteren Verlauf angepasst werden sollten, wobei auch bereits Ergebnisdaten einbezogen werden sollten. Eine beträchtliche Quote dieser Programmevaluationen strebte zugleich an, bereits zu diesen frühen Zeitpunkten einzuschätzen, ob das Programm überhaupt weitergeführt oder aber eingestellt werden sollte. Charakteristika der Programmevaluationen, die sich in der folgenden Analyse auf Basis der Standards zeigen, müssen in diesem Zusammenhang gesehen werden.

4.2 Gesamtbild der Erfüllung der Standards

Ersichtlich wird in der Berichtsanalyse, dass zu jedem Standard eine nicht unerhebliche Schwankungsbreite bei der Erfüllung der jeweiligen Forderungen, Hinweise und Empfehlungen vorliegt. Jedem Evaluationsbericht bzw. jeder Programmevaluation, über die Bericht erstattet wird, kommt dabei ein individuelles Profil zu. Diese Perspektive, die auf Unterschiede anstatt auf Gemeinsamkeiten fokussiert, wurde angesichts der Zwecksetzung der Metaevaluation nicht verfolgt. Es steht eine übergreifende Charakterisierung der österreichischen Evaluationspraxis im Zentrum des Interesses, die sodann auch zu allgemein-strukturell orientierten Erkenntnissen und daraus ableitbaren Empfehlungen führen kann. Es steht für die Metaevaluation also nicht die Individualität der einzelnen Programmevaluationen bzw. Berichte im Vordergrund, sondern die Identifikation von gemeinsamen Zügen, die im Sinne einer Symptomatologie Stärken und Schwächen erkennen lassen.

Der Schwerpunkt der Verteilungen zu allen Standards liegt, wie die folgende Abbildung zeigt, im mittleren Bereich einer neutralen Einschätzung, dass Gesichtspunkte des Standards soweit erfüllt sind, dass weder eine klare Schwäche noch eine klare Stärke besteht. Die Einstufungen im mittleren Bereich kommen in etlichen Fällen auch dadurch zustande, dass feststellbare Stärken durch gleichzeitig feststellbare Schwächen aufgewogen werden. Dass essentielle Grundbedürfnisse einzelner Standards nicht erfüllt wurden, kommt über den gesamten zwölfjährigen Beobachtungszeitraum nur äußerst selten vor. Die wenigen betreffenden Fälle siedeln sich in den weiter zurückliegenden Zeitabschnitten des zwölfjährigen Analysezeitraums an, Nicht-Erfüllung eines Standards kommt in der rezenten Evaluationspraxis der Jahre 2011- 2014 nicht mehr vor. Ebenso selten ist jedoch auch eine Berichterstattung bzw. in der vorliegenden Berichterstattung erkennbare Vorgehensweise der Evaluationen, die als sehr gute Erfüllung der Standards vollumfänglich begrüßt werden kann. In einigen Fällen konnte die Einstufung „sehr gute Erfüllung“ trotz zahlreicher Stärken nicht vergeben werden, da gleichzeitig doch auch eine nicht übersehbare Schwäche vorlag.

Abbildung 15: Gesamtergebnisse der Berichtsanalyse für die herangezogenen Standards



Die Standards N1 „Identifizierung der Beteiligten und Betroffenen“ und F5 „Offenlegung der Ergebnisse“ können allein anhand der Berichte nur recht bedingt eingeschätzt werden. Die wesentlichen Informationsquellen für eine Einschätzung der bisherigen Evaluationspraxis bilden in diesen Fällen die Angaben von EvaluatorInnen, die die durchgeführte Umfrage beantwortet haben. Da es sich bei diesen antwortenden EvaluatorInnen hochgradig um solche handelt, die mehrere Programmevaluationen im österreichischen FTI-Bereich durchgeführt haben, kann davon ausgegangen werden, dass diese Auskünfte die wesentlichen Züge der Evaluationspraxis gut wiederspiegeln.

Nicht nur im Hinblick auf die beiden genannten Standards, sondern auch deutlich darüber hinaus war die Metaevaluation in der Vornahme von Einschätzungen behindert. Der bisherige Umgang mit den Anforderungen an einen Evaluationsbericht, den der Standard G3 „Beschreibung von Zwecken und Vorgehen“ formuliert, hat zum Ergebnis, dass verschiedene Qualitätsaspekte mit hoher Relevanz für die Nützlichkeit, die in dieser Metaevaluation anhand der Berichte beleuchtet werden sollten, nur bedingt in gut erhärteter Weise eingeschätzt werden konnten. In den Joint Committe-Standards, die hinter den DeGEVal-Standards stehen, wird klar darauf hingewiesen, dass eine gute Berichterstattung über alle Aspekte des Vorgehens einer Evaluation und alle Aspekte ihrer Methodik die Voraussetzung dafür bilde, dass eine Metaevaluation sinnvoll durchgeführt werden kann. Ein erstes, übergreifendes Ergebnis der Berichtsanalyse ist somit, dass die Evaluationsberichte in einer Konzentration auf Daten und Dateninterpretationen sowie Schlussfolgerungen, die aus diesen Faktenlagen gezogen und für Empfehlungen genutzt werden können, die Darstellung anderer Aspekte einer Evaluation, die auf Basis der Standards als ebenso wesentlich gelten müssen, und die Vermittlung von methodischen Hinweisen häufig zumindest ein Stück weit vernachlässigen.

Die Einschätzungen zum Standard N8 „Nutzung und Nutzen der Evaluation“ haben wegen dessen spezifischer Gestaltungsweise eher tentativen Charakter. Insofern hier spezifische Anforderungen erhoben werden, geben die analysierten Berichte nur ansatzweise relevante Auskünfte. Zugleich handelt es sich hier um eine übergreifende Sicht auf die Performance zu allen anderen Standards, auf deren Einschätzbarkeit wiederum die genannten Berichtsschwächen durchschlagen.

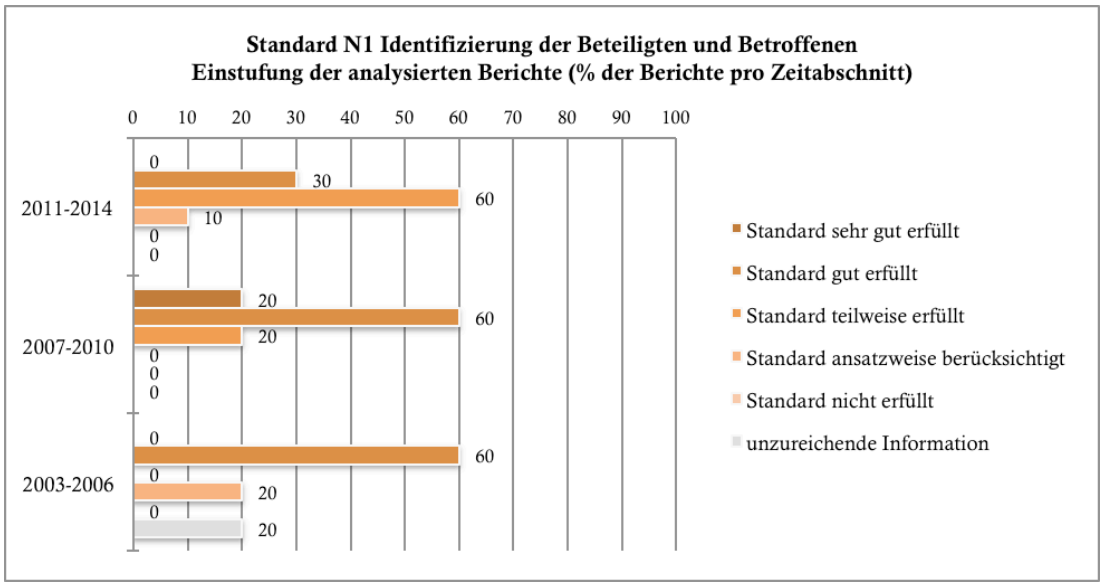
Wie die Übersicht über die erreichten Einstufungen zeigt, liegen die vergleichsweise schwächsten Erfüllungsniveaus im Bereich der Prinzipien und Hinweise von fünf der herangezogenen Standards vor. Es sind dies die Standards N4 „Auswahl und Umfang der Informationen“, N5 „Transparenz von Werten“, G2 „Kontextanalyse“, der bereits als eher problematisch genannte Standard G3 „Beschreibung von Zwecken und Vorgehen“, sowie tendenziell auch F3 „Vollständige und faire Überprüfung“. Hier zeichnen sich also spezifische Aufmerksamkeitspunkte für zukünftige Weiterentwicklungen der Evaluationspraxis und –kultur ab. Das bedeutet aber nicht, dass man sich nicht auch mit dem Spielraum im Bereich anderer Standards auseinandersetzen sollte, um höchste mögliche Evaluationsqualität zu erreichen. Freilich bemisst sich die Erfüllbarkeit jedes Standards auch an der Art der Evaluationsansätze und Evaluationsmodelle, die zum Einsatz gelangen, und es wird sich innerhalb einer bestimmten Herangehensweise unter gegebenen Rahmenbedingungen nicht alles erreichen lassen, was als Idealbild einer bestmöglichen Evaluation beschrieben wird. Auf identifizierbare Möglichkeiten, in Zukunft über das bereits erreichte Niveau hinaus noch Verbesserungen zu erzielen, wird im abschließenden Kapitel zu Schlussfolgerungen und Empfehlungen zurück gekommen.

Die Berichtseigenschaften und Gesichtspunkte der Standards, die jeweils relevant für die Einstufungen waren, und die die Aufmerksamkeitspunkte für eine künftige weitere Verbesserung der Evaluationspraxis bilden können, sind jeweils den Factsheets im Anhang zu entnehmen. Diese qualitative Ebene wird in den folgenden Abschnitten des vorliegenden Berichts in zusammenfassender und am Überblick orientierter Weise verfolgt. Dabei wird auch auf die Entwicklung der Performance zum jeweiligen Standard im Zeitablauf anhand der dem Berichtssampling zugrunde gelegten Perioden eingegangen. Zu fast allen Standards zeichnet sich dabei eine positive Entwicklung im Zeitverlauf ab.

4.3 Identifizierung der Beteiligten und Betroffenen

Mit der Identifikation und Einbeziehung der Beteiligten und Betroffenen thematisieren die DeGEval-Standards einen für die Nützlichkeit einer Evaluation relevanten Qualitätsaspekt,, dem die Evaluationsberichte im FTI-Bereich bislang keine Aufmerksamkeit geschenkt haben. Für die Einschätzung benötigte und relevante Angaben finden sich nahezu ausschließlich in Methodendarstellungen, die die Evaluationsberichte regelmäßig enthalten. Wegen der Vernachlässigung relevanter Information in der Berichterstattung kann die Einschätzung der Erfüllung des Standards nur mit Einschränkungen vorgenommen werden. Es ergibt sich dabei ein Bild einer Erfüllung des Standards auf im Großen und Ganzen mittlerem Niveau, wobei in rezenten Zeitabschnitten auch teilweise von guten Erfüllungen gesprochen werden kann, aber im Gesamtbild nur wenig Veränderungen eingetreten sind.

Abbildung 16: Ergebnisse der Berichtsanalyse für den Standard N1



Es gibt kaum eine Programmevaluation im untersuchten Sample, die nicht im Zuge ihrer Datenerhebungen auch Erfahrungen und Sichtweisen von Programmverantwortlichen und mit der Umsetzung des Programms Betrauten recherchiert und in ihre Analyse einbezogen hätte. Hinzu kommen oft Erfahrungen und Sichtweisen der Zielgruppen des evaluierten Programms oder von Ausschnitten dieser Zielgruppen, da in aller Regel auf Fördernehmer fokussiert wird. Diese wurden gelegentlich um nicht-erfolgreiche Antragsteller innerhalb des untersuchten Programms ergänzt. Allerdings wird in dieser Herangehensweise der Grundidee des Standards N1 nur recht bedingt Genüge getan, dass Akteure, für die eine Programmevaluation nützlich werden soll, auch bereits im Vorfeld in die Evaluationsplanung einbezogen werden sollten, sodass sie über die Programmevaluation informiert sind und dazu auch Stellung beziehen können. Hierfür kommen folgende Akteursgruppen in Frage:

- Für die Evaluation zuständige AnsprechpartnerInnen in den Auftraggeber-Organisationen;
- Personen oder Personengruppen, die abgesehen von dem/der direkten AuftraggeberIn oder neben ihm/ihr für die Konzeption und Gestaltung des Programms verantwortlich waren bzw. über die Zukunft des untersuchten Programms zu entscheiden hatten;
- Personen oder Personengruppen, die mit der Umsetzung des Programms befasst waren;
- Personen oder Personengruppen, die durch das Programm erreicht werden sollten;
- Personen oder Personengruppen, die durch das Programm oder durch Veränderungen des Programms Nachteile erleiden hätten können (z.B. im Fall einer Veränderung von Zielgruppen-Definitionen oder Antragsbedingungen; in einem weiten Verständnis auch gesellschaftliche Gruppen, die von sozio-ökonomischen Auswirkungen von FTI-Politiken betroffen sein können);

- Personen oder Personengruppen, die ähnliche Programme planten bzw. in ihren FTI-politischen Rollen Interesse am untersuchten Programm hatten bzw. haben konnten.

In den meisten der analysierten Berichte liegt kein Hinweis vor, dass abgesehen von den direkten AuftraggeberInnen auch andere Beteiligte und Betroffene des untersuchten Programms in die Planung der Evaluation einbezogen worden wären, um ihre Informationsbedürfnisse zum Evaluationsgegenstand in Erfahrung zu bringen. Oft ist von „den Programmverantwortlichen" die Rede, wobei die genaue Extension dieses Begriffs unscharf bleibt und z.B. bei von zwei Ministerien verantworteten oder von zwei Agenturen umgesetzten Programmen nicht klar wird, welche Akteure genau einbezogen waren. Inwieweit die Einbeziehung verschiedener Akteursgruppen in eine Programmevaluation angebracht erscheint, hängt anerkannter Maßen auch vom gewählten Evaluationsansatz, von der Art des evaluierten Programms und weiteren Umständen im gesellschaftlich-politischen Evaluationskontext ab. Bei der Einschätzung des Standards wurde daher mit den FTI-Programmevaluationen, denen oft ein generell stark datenorientierter, objektivistischer Zug nachgesagt wird, benevolent umgegangen. Gesagt werden kann jedoch jedenfalls, dass die Programmevaluationen sich nicht als partizipative Evaluationen (*participatory evaluation*) verstehen lassen, und dass daher in der Evaluationspraxis bislang auch Chancen, die sich aus einem solchen Evaluationsansatz ergeben können, nicht wahrgenommen wurden.[7] Nur in einem Bruchteil der Berichte liegt ein als gesichert erachtbarer Hinweis vor, dass in die Evaluationsplanung auch FTI-politische Akteure einbezogen wurden, die zwar nicht unmittelbarer mit dem Programm befasst waren, aber doch an der Evaluation und ihren Ergebnissen ein Interesse haben konnten.

Selten ist aber auch anhand der Berichte nachvollziehbar, dass eine gezielte intensivere Interaktion mit den AuftraggeberInnen nicht nur am Anfang des Evaluationsprozesses (Auftragsvergabeverfahren und Kick-Off), sondern auch im Weiteren während der Durchführung der Programmevaluationen stattgefunden hat, durch die auf ihre Informationsbedürfnisse und Sichtweisen auf den Evaluationsgegenstand genauer eingegangen werden konnte und zugleich auch Lernprozesse unter ihnen gefördert werden konnten (z.B. ein Workshop zur Logic Chart bzw. zur Programmlogik oder ein Workshop zur Diskussion von Zwischenergebnissen).

Die EvaluatorInnen-Befragung bestätigt, dass es am Ehesten die für die Evaluation unmittelbar zuständigen AnsprechpartnerInnen in den auftraggebenden Institutionen waren, die in der Planungsphase der Evaluationen einbezogen wurden. Allerdings geben nur 69% der EvaluatorInnen an, dass dies bei den von ihnen durchgeführten Evaluationen immer der Fall war. Die von einem Drittel der EvaluatorInnen berichtete Vernachlässigung einer näheren Einbeziehung der AuftraggeberInnen über die Beantwortung der Terms of Reference des Ausschreibungsverfahrens hinaus, zumindest in manchen ihrer Evaluationsprojekte, kann mit der Herangehensweise einer strikt objektivistischen Herangehensweise an Evaluation in Verbindung gebracht werden. In solchen Evaluationsansätzen wird die Frage nach spezifischen Informationsbedürfnissen der AuftraggeberInnen gegenüber der Sichtbarmachung von objektiven Wahrheiten zu den Programmen hintangestellt. Im Blickwinkel der Standards auf die Nützlichkeit von Evaluationen wird dies allerdings eher als nachteilig begriffen, und zahlreiche Aussagen von EvaluationstheoretikerInnen besagen, dass die Grenzen für eine Nutzenentfaltung hier eher eng gezogen sind (vgl. Kap. 1.2.5)

Gemäß den Angaben der EvaluatorInnen bestätigt sich im Wesentlichen das Bild aus den Berichtsangaben, dass neben den direkt Evaluationszuständigen (AuftraggeberInnen) unter verschiedenen Akteursgruppen, die dem Standard N1 gemäß als Beteiligte und Betroffene des untersuchten Programms zu begreifen sind, am ehesten Personen in die Evaluationsplanung einbezogen wurden, die mit der Umsetzung des untersuchten Programms befasst waren, sowie Personen, die für die Konzeption und Gestaltung des untersuchten Programms verantwortlich waren. Dies entspricht der Einbeziehung des jeweils gegenüberliegenden Akteurs innerhalb der Principal-Agent-Beziehungen, die den erhaltenen Angaben zufolge häufig, aber doch nicht durchgängig Usus war bzw. ist. Auch in den Gesprächen mit AuftraggeberInnen erwies sich, dass Abstimmungen innerhalb von Principal-Agent-Beziehungen nicht immer zur vollsten Zufriedenheit aller Seiten

[7] Unter *Participatory Evaluation* werden Evaluationsansätze verstanden, in denen Stakeholder bzw. Beteiligte und Betroffene stark eingebunden werden, um ihnen Mitsprache im Evaluationsprojekt zu geben und Lernen zu ermöglichen. Die Stakeholder können dabei auch Evaluationsaufgaben übernehmen oder an der Gestaltung von Erhebungsinstrumenten oder Dateninterpretationen mitwirken (vgl. z.B. Cousins/Whitmore 1998).

erfolgten, wobei am ehesten die beiden nicht völlig autonomen Agenturen signalisieren, dass sie manches Mal mit ihren spezifischen Informationsbedürfnissen „am kürzeren Ast saßen".

Es bestätigt sich in den Angaben der EvaluatorInnen ebenso, dass bislang nur selten andere FTI-politische Akteure, die am evaluierten Programm und seiner Einschätzung aus ihren jeweiligen Rollen heraus ein Interesse haben konnten, in die Planung von Programmevaluationen einbezogen wurden. Ebenso selten wurden Informationsbedürfnisse von potenziell durch das Programm bzw. dessen Veränderung Benachteiligte gezielt berücksichtigt. Am ehesten wurden noch Zielgruppen der evaluierten Programme in die Evaluationsplanung einbezogen. Dies stellt als solches einen positiven Hinweis dar, der auf die Aufbereitung eines guten Bodens für die Nützlichkeit der Programmevaluationen auch für diejenigen Akteure hinweist, deren Chancen oder Verhaltensweisen durch die evaluierten Maßnahme beeinflusst werden sollten. Allerdings wird doch von rund zwei Drittel der EvaluatorInnen angegeben, dass dies in den von ihnen durchgeführten Progammevaluationen nur selten oder nie der Fall war.

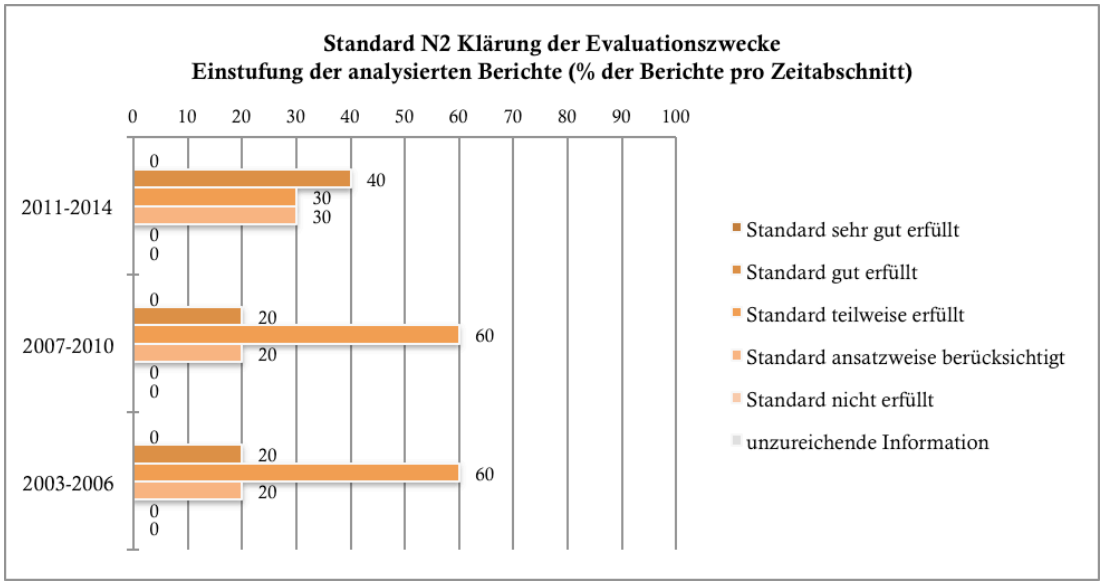
Was an der Frage nach der Einbeziehung in einen Evaluationsprozess freilich nicht sichtbar werden kann, sind eventuelle politische Rücksichtnahmen in der Formulierung eines Evaluationsprojekts und der Terms of Reference, die in den Interviews gelegentlich erwähnt wurden.

4.4 Klärung der Evaluationszwecke

Die untersuchten Evaluationsberichte machen zu den Evaluationszwecken, die die jeweilige Untersuchung angeleitet haben, lediglich umrisshafte und eher unscharfe Angaben. Es findet sich kein einziger Evaluationsbericht im Sample, der explizit unter dem Stichwort der Evaluationszwecke Angaben gezielt präsentieren würde. Die vorhandenen kursorischen Angaben zu den Evaluationszwecken haben oft eine Form, in der verschiedene, in sich berechtigte und alternativ mögliche Zugangsweisen zur allgemeinen Charakterisierung einer Evaluation additiv miteinander verbunden werden. Es werden auch keinerlei Aussagen über intendierte AdressatInnen der Evaluationen bzw. Evaluationsberichte gemacht, sodass offenbar weitgehend wie selbstverständlich davon ausgegangen wird, dass die direkten AuftraggeberInnen die NutzerInnen der Programmevaluationen sein werden und sich ein weiterer Kommentar erübrigt.

Auf Basis der in den untersuchten Evaluationsberichten vorfindlichen Angaben ergibt sich eine Einschätzung, die die Erfüllung der Forderung nach klaren Evaluationszwecken in einer Bandbreite zwischen eher guten und nur ansatzweisen Entsprechungen zeigt, die sich im Zeitablauf leicht zum Positiven entwickelt hat.

Abbildung 17: Ergebnisse der Berichtsanalyse für den Standard N2



Ersichtlich wird anhand der Berichtsangaben, dass zumindest drei Viertel der analysierten Evaluationen einen lernorientierten Ansatz verfolgten, um Erkenntnisse zu erbringen, auf Grund derer das evaluierte Programm angepasst werden kann oder andere zukünftige Programme besser ausgerichtet werden können. In 15% der Evaluationsberichte wurde (auch) die Fragestellung aufgeworfen, ob das evaluierte Programm eingestellt oder fortgesetzt werden sollte. 10% der Evaluationsberichte stellten einen formalen Bezug zu Evaluationsvorgaben in Programmrichtlinien her, ohne jedoch auf die Zwecksetzung, die sich aus diesen Evaluationsvorgaben ergibt, einzugehen. Ein erheblicher Teil der Evaluationsberichte benennt die jeweilige Zwecksetzung allerdings lediglich in einer recht oberflächlichen Weise, indem etwa von einer „Reflexion des Programmverlaufs", einer „Zusammenfassung von Erfahrungen" oder einer „kritische Würdigung des Programms" gesprochen wird oder lediglich gesagt wird, dass eine Bewertung des Programms vorgenommen werden soll. Derartige Angaben können nur schlecht als Definition von Evaluationszwecken im Sinn des Standards N2 erachtet werden. Eingeführte Begrifflichkeiten der Evaluationsmethodologie wie „Konsistenz und Kohärenz" eines Programms, „Implementation" der ursprünglich geplanten Programmanlage oder „Implementierungstreue" der tatsächlichen Programmumsetzung gegenüber dem ihr vorangegangen Programmkonzept kommen in den Darstellungen der Evaluationszwecke und den Beschreibungsweisen des Vorgehens der Evaluationen nicht vor.

In den Interviews mit den AuftraggeberInnen tritt ein Zweck der Rechenschaftslegung, der von den Evaluationsberichten nicht explizit gemacht wird, deutlich hervor. Zugleich wird erkennbar, dass die Evaluationsfunktion der Rechenschaftslegung, die im Rahmen der institutionell-rechtlichen Verankerung der Programmevaluationen immer schon mit einprogrammiert ist, auch zu Einschränkungen für die Lernfunktion der Programmevaluationen und für die Erzielung von instrumentellem und konzeptuellem Nutzen sowie von über das unmittelbar betrachtete Programm hinaus reichendem Wissenzuwachs („Aufklärung") führt. Stellvertretend für mehrere ähnliche Aussagen kann hier das folgende Zitat stehen: *„In der Praxis des Alltags evaluiert man die eigenen Programme, weil es die Praxis ist, weil es vorgesehen ist., weil man es machen muss – man arbeitet das sozusagen auch ab, natürlich ist es auch Pflicht und ein Stück weit Pflichtübung, und hat nicht immer großen Neuigkeitswert." (M1)* Es wird beschrieben, dass durch Evaluationsvorgaben und die frühzeitige Verankerung von Evaluationsfragestellungen aktueller Erkenntnisbedarf nur bedingt befriedigt werden kann. *„Wenn ich etwas [aus einem Evaluationsbericht] bekomme, dann freue ich mich. (...) Vielleicht gibt es da ein paar interessante Momente, die man so gar nicht gesehen hätte. Mich würde einfach immer auch interessieren, welche Fragen ich auf Grund meiner aktuellen beruflichen Herausforderungen gern beantwortet hätte. Das ist dann mehr oder weniger der Fall." (M1)*

Der Standard N2 formuliert das zentrale Anliegen, dass die EvaluatorInnen durch klare Definitionen der übergreifenden Evaluationszwecke, die der späterhin anzustellenden Analyse vorausgesetzt sind, mit einem klaren Arbeitsauftrag ausgestattet sein sollen. Anhand der Berichte lässt sich dies nur bedingt feststellen. Dazu wären nähere Angaben, die die Aufgabenstellung über grobe Intentionen hinaus spezifizieren, notwendig. Zu einem geringen Prozentsatz findet sich eine vollständige Angabe der mit den AuftraggeberInnen vereinbarten Evaluationsfragestellungen, oder zumindest Hinweise, dass innerhalb von breit formulierten Evaluationszwecken Hauptfragestellungen definiert wurden. Ein Großteil der analysierten Evaluationsberichte verzichtet freilich darauf, die vereinbarten Evaluationsfragestellungen, die der Bericht zu beantworten versucht, aufzulisten. Während grundsätzlich davon ausgegangen werden kann, dass im Rahmen der üblichen Vergabeverfahren allen Evaluationen ein in Terms of Reference niedergelegtes Set von Evaluationsfragestellungen zugrunde gelegt wurde, kann dieser Wesenszug der FTI-Evaluationspraxis in den Evaluationsberichten nur stark eingeschränkt nachvollzogen werden.

In der Umfrage unter EvaluatorInnen geben 89% an, dass in den von ihnen durchgeführten FTI-Programmevaluationen zumindest häufig Evaluationszwecke soweit verankert waren, dass das Evaluationsteam einen klaren Arbeitsauftrag hatte. Lediglich 39% meinen, dass dieses Grunderfordernis für ein gutes Gelingen einer Evaluation in allen durchgeführten Programmevaluationen stets erfüllt war. Zugleich geben 11% an, dass eine Klarheit der Evaluationszwecke und ein klarer Arbeitsauftrag in den von ihnen durchgeführten FTI-Programmevaluationen nur selten gegeben war.

Wesentliches Anliegen des Standards ist, dass eine Mehrzahl von Evaluationszecken sich mit einiger Wahrscheinlichkeit nachteilig auf die anzustellende Untersuchung und ihre Ergebnisse auswirken kann, und dass daher verschiedene Evaluationszwecke soweit wie möglich strukturiert und im Verhältnis zueinander priorisiert werden sollten. Es wird auf Differenzen zwischen verschiedenen

Arten von Erkenntnissen hingewiesen, die unterschiedlichen Nutzungen der Evaluation dienen. Diesbezüglich geben lediglich 28% der EvaluatorInnen an, dass bei den von ihnen durchgeführten Programmevaluationen immer ein Hauptzweck der Evaluation klar im Vordergrund stand, bzw. dass die Evaluationszwecke gemeinsam mit den Auftraggeberinnen mit klaren Prioritäten ausgestattet wurden. Weitere 61%, berichten, dass häufig ein Evaluationszweck priorisiert wurde. Immerhin 11% sagen freilich, dass eine Priorisierung eines Hauptzwecks nur selten vorlag bzw. im Dialog mit den AuftraggeberInnen erreicht wurde.

Der Standard erweist sich anhand dieser Angaben als in der bisherigen FTI-Evaluationspraxis als nicht durchgängig erfüllt. Die Wahrnehmungen der EvaluatorInnen über Zweckklärungen bzw. -priorisierungen finden nur bedingt eine Korrespondenz in den analysierten Berichten, sodass hinsichtlich der in der Berichtsanalyse festgestellten Unschärfe offenbar in erster Linie von Berichtsschwächen zu sprechen ist. Die Angaben der EvaluatorInnen scheinen darüber hinaus aber auch das Spannungsfeld zwischen einem Zweck des Lernens und einem Zweck der eher routineartigen Rechenschaftslegung, das in den Gesprächen mit AuftraggeberInnen sichtbar wurde, nur recht bedingt widerzuspiegeln.

Ein weiterer Hinweis des Standards N2 zur Gestaltung möglichst zielführender und nützlicher Evaluationen besagt, dass verschiedene Hauptzwecke, die für eine Evaluation angedacht sind, in zeitlich getrennten Phasen oder arbeitsteilig durch unterschiedliche Evaluationsteams bearbeitet werden sollten. Eine derartige Auftrennung und Verteilung von Evaluationszwecken auf mehrere Evaluationen desselben Programms ist in der bisherigen Evaluationspraxis kaum erfolgt, wenn man von einigen wenigen Begleitevaluationen absieht, bei denen zu verschiedenen Evaluationszeitpunkten unterschiedliche Fragestellungen fokussiert wurden. Hinzu kommen einige wenige Fälle, wo ex ante-Evaluationen oder ex post-Evaluationen mit den Begleitevaluationen kombiniert wurden, die weitaus am Verbreitetsten sind und das Gros der Programmevaluationen ausmachen. 60% der EvaluatorInnen geben an, dass von ihnen evaluierte Programme nie mehrfach evaluiert wurden. Für die 40%, die angeben, dass eine Mehrfachevaluation zumindest ab und an („selten“, „häufig“ oder „immer“) stattfand, bleibt unklar, inwieweit sie bei dieser Aussage die bereits angesprochenen Begleitevaluationen im Auge hatten.

Eine im Konzept der Metaevaluation ursprünglich ins Auge gefasste Katalogisierung der in den Berichten angegebenen Evaluationszwecke wurde nicht durchgeführt, da sich im Verlauf der Berichtsanalyse zeigte, dass die umrisshaften Angaben keinen Erkenntniswert beinhalten, der über die vorgenommenen Klassifikationen der Evaluationstypen und der Evaluationsschwerpunkte gemäß OECD DAC Minimum Standards (vgl. oben) hinausreicht.

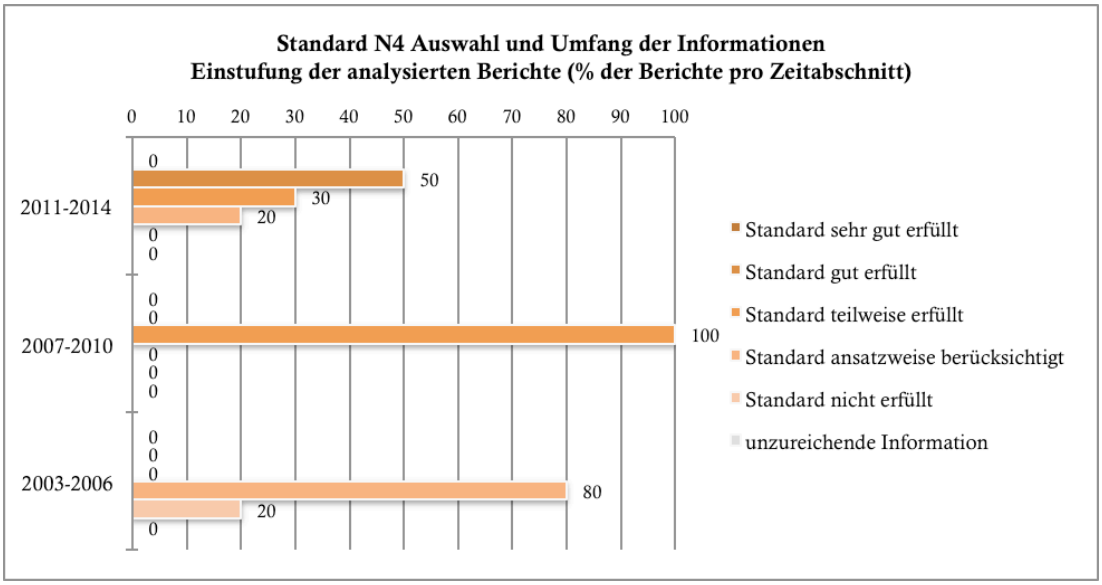
4.5 Auswahl und Umfang der Informationen

Die Erfüllung des Standards N4 zeigt über den zwölfjährigen Beobachtungszeitraum eine steigende und eindeutig positive Tendenz. In den letzten Jahren liegen Erfüllungsniveaus vor, die auf die positive Seite einer guten Erfüllung ausschlagen oder zumindest mit einer neutralen Einschätzung zu versehen sind. Schwächen, die im am weitesten zurückliegenden Zeitabschnitt noch zu beobachten waren, treten in deutlich geringerem Ausmaß auf, wenn auch in der jüngsten Vergangenheit noch ab und an Abstriche gemacht werden müssen.

Der Standard zielt darauf ab, die von einer Evaluation herangezogenen Informationen so zu bestimmen, dass alle vorgesehenen Fragestellungen gut behandelt werden können. Es geht hier also nicht primär um die eingesetzten Methoden im engen Sinn, sondern um einen konzeptiven Gesamtzusammenhang, innerhalb dessen auf Basis der Fragestellungen definiert wird, mithilfe welches Methodensets benötigte Informationen beschafft werden können. Die analysierten Programmevaluationen stützen sich durchwegs auf Daten und Informationen, die als sachdienlich zur Analyse der evaluierten Programme zu erachten sind und mithilfe von adäquaten Methoden erbracht werden. In einem beträchtlichen Teil wurden quantitativ-qualitative Methodenmixes zum Einsatz gebracht, ebenso wie fast immer auf Monitoringdaten der evaluierten Programme zugegriffen wurde, um den Analyseebenen gerecht zu werden, die den Schwerpunktsetzungen und Evaluationsfragestellungen entsprachen. Gemäß dem allgemeinen evaluatorischen Grundsatz, dass jede Evaluation ihr spezifisches Vorgehen nach den Umständen des Einzelfalls definieren soll, weisen auch die im Sample befindlichen Programmevaluationen eine ganze Bandbreite an breiteren oder schlankeren Informationsgewinnungsverfahren auf. Als wesentliche Bedingungen für die von den einzelnen Evaluationen eingeschlagenen Vorgehensweisen sind zugleich auch die für sie zur

Verfügung gestellten Ressourcen anzunehmen, die in der vorliegenden Untersuchung jedoch nicht in die Betrachtung einbezogen wurden. Die folgenden Ausführungen stützen sich somit nicht auf Überlegungen zu einer Kosten-Effektivität der Evaluationen, sondern auf den prinzipiellen Ansatzpunkt des Standards N4, dass Auswahl und Umfang der erfassten Informationen substanzielle und konsistente Erkenntnisse über den Evaluationsgegenstand ermöglichen sollen.

Abbildung 18: Ergebnisse der Berichtsanalyse für den Standard N4



Hinsichtlich der zentralen Fragestellung nach der für den konkreten Einzelfall passend gewählten bzw. recherchierten Information war die Berichtsanalyse mit einigen Schwierigkeiten konfrontiert. Art und Umfang der von einer Programmevaluation benötigten Information bemessen sich sowohl an der Konfiguration des zu evaluierenden Programms als auch an dem spezifischen Blickwinkel, den die Programmevaluation darauf einnimmt. Eine vollständige Listung der den Programmevaluationen aufgegebenen Evaluationsfragestellungen, durch die dieser evaluative Blickwinkel auf Detailebene definiert wird, ist nur in einer kleinen Anzahl der untersuchten Evaluationsberichte anzutreffen. Ein konsequenter Nachvollzug der Qualität, in der die Programmevaluationen Informationen für die Beantwortung ihrer Fragestellungen heranzogen, war in allen diesen Fällen nicht möglich. Für das Gros der untersuchten Berichte konnte sich die Metaevaluation nur auf die oft eher umrisshaften Angaben zu Zwecksetzungen und Untersuchungsschwerpunkten sowie auf die Angaben zum evaluierten Programm stützen, die die Berichte bereit stellen.[8] Für viele der in die Analyse einbezogenen Programmevaluationen zeigt sich, dass gerade dieser übergreifende konzeptive Zusmmenhang oft nur unzureichend zur Darstellung gelangt (vgl. dazu die Erläuterungen zum Standard G3 weiter unten).

[8] Einer guten Beurteilbarkeit der Erfüllung dieses Standards ist eine gute und konzise Darstellung der Evaluationszwecke und .fragestellungen sowie des untersuchten Programms vorausgesetzt, aus der sich die Sachgerechtheit und Produktivität der Mittel, die die Evaluation zu deren Behandlung bzw. Beantwortung einsetzt, erschließt. Da sich die meisten der analysierten Berichte auf eine allgemein-abstrakte Angabe von Evaluationszwecken beschränken und keine priorisierten Analyseschwerpunkte oder Evaluationsfragestellungen angeben, waren die Voraussetzungen für diese Einschätzung im Gesamtbild letztlich nur eingeschränkt gegeben. Die Analyse bezieht sich daher in erster Linie auf die grundsätzliche Eignung des Methodeneinsatzes. Sie bezieht sich des Weiteren auf Datendefizite oder konzeptive Lücken, die im jeweils berichteten Analyseverfahren augenfällig werden. Den Hintergrund der Einschätzungen bildet das Methodenwisssen des Metaevaluators. Zu berücksichtigen bleibt, dass die im Sinne des Gehalts des Standards und einer verbesserungsorientierten Denkweise, die neue Sichtweisen auf den Evaluationsgegenstand (also die methodischen Vorgehensweisen der Evaluationen) eröffnen kann, die Frage „Was hätte besser gemacht werden können?“ naheliegend ist, aber auch leicht ausufern kann, da sie auch mit methodischen Positionierungen und disziplinär verankerten Präferenzen einhergeht. Die Analyse war nach bestem Wissen und Gewissen bestrebt, sich auf die Bedarfslagen der konkreten Studie zu beschränken, so wie sie vorgegangen ist bzw. nach eigenen Angaben vorgehen wollte oder vom Sachlichen her sollte. Dies darf nicht mit einer „Kritik um der Kritik willen“ verwechselt werden.

Deutlich anwesend ist in zahlreichen Programmevaluationen ein Zug eines „data first approach“, der von verfügbaren Monitoringdaten aus den programmumsetzenden Agenturen seinen Ausgang nimmt und diese durch Interviews und Befragungen mit Programmverantwortlichen und Zielgruppen ergänzt. Dabei ist eine weitgehende Beschränkung auf Fördernehmer zu beobachten, die gelegentlich um nicht erfolgreiche Förderwerber (Antragsteller) ergänzt werden. Teile der Zielgruppen, die von der Maßnahme bislang nicht erreicht wurden, werden in aller Regel nicht untersucht. Manche Programmevaluationen haben Datenlücken oder Einschränkungen der Tragfähigkeit von Monitoringdaten thematisiert, doch wurden in solchen Fällen keineswegs durchgehend eigene Erhebungen angestellt, um diese Probleme auszugleichen. Durch diese Selbstbeschränkungen im Vorgehen der betreffenden Evaluationen können wesentliche Gesichtspunkte der Wirkungsweise von Programmen, die zu eine potenziell besseren Programmgestaltung genutzt werden könnten, nicht erkannt worden sein.

Die herangezogenen Informationen haben zweifellos Stellenwert in einer sachgerechten Analyse von verschiedenen Dimensionen der Relevanz, Effektivität und Wirksamkeit des untersuchten Programms. Allerdings werden zu diesen Untersuchungsdimensionen meistens jeweils nur einige Daten herangezogen. Es bleibt über Strecken unklar, ob aus den genutzten Datenquellen auch noch weitere Informationen zur Verfügung gestanden hätten, die jeweils eine Vertiefung bzw. Verfeinerung der Analyse ermöglicht hätten, oder ob solche Daten über die genutzten Datenquellen hinaus zusätzlich erhoben werden hätten müssen. Stellenweise wird von den EvaluatorInnen festgehalten, dass für eine zielführende Analyse benötigte Daten nicht verfügbar waren oder nicht erhoben werden konnten.

Immer wieder sind im Ansatz durchaus sinnvolle, aber letztlich nur bruchstückhafte oder nicht konsequent zu Ende geführte Analysen festzustellen. So werden manchmal Programmaspekte anhand von Auskünften der Fördernehmer des Programms untersucht, die mehr Ertrag bringen hätten können, wenn man sie auch Akteuren innerhalb der Zielgruppen gestellt hätte, die noch nicht vom evaluierten Programme erfasst worden waren. Andere Lücken liegen z.B. dann vor, wenn zwar angemerkt wird, dass Kooperationen mit bestimmten Akteursgruppen oder regionale Synergiebildungen integraler Bestandteil der beabsichtigten Wirkungsweise eines Programms waren, dem dann aber doch nicht anhand von adäquaten Daten nachgegangen wurde. Nicht-monetäre Programmbestandteile (Begleitmaßnahmen) wurden in aller Regel nicht in die Analyse einbezogen, und auch über geförderte Aktivitäten hinausreichende Merkmale von Fördernehmern wurden kaum beleuchtet (z.B. Verhaltensweisen oder Bewußtseinslagen, die mit der Verwirklichung der finanziell geförderten Aktivitäten in Zusammenhang stehen). Indirekte Zielgruppen, die in manchen Programmen die eigentlichen Begünstigten darstellten, werden des Öfteren wenig beachtet bzw. in ihrem Status und hinsichtlich der Art der Erreichung von Effekten nicht klar bestimmt. Häufig wurden auch keine Beziehungen zwischen der Ebene von Programmoutputs und Programmoutcomes hergestellt, und die dieser Wirkungsbeziehung vorausgesetzten Schritte der Programmumsetzung wurden nicht eingehend untersucht. Mit zunehmender Breite der Evaluationsanlagen erhöht sich zugleich die Tendenz, dass die Aufarbeitung von Inputs, Outputs, Outcomes und eventuellen bereits beobachtbaren Impacts zunehmend unvollständiger erscheint, indem zwar durchaus auf verschiedene Daten zu den genannten Untersuchungsebenen zugegriffen wird, aber ein gesamtlogischer Zusammenhang der schrittweisen Entfaltung der Programmaktivitäten und ihrer erfolgreicheren oder weniger erfolgreichen Schritte hin zur Erreichung der intendierten Ziele nicht gesamthaft und vollständig verfolgt wird.

Einige Berichte kommen mit einer nur abrisshaften bzw. punktuellen Bezugnahme auf die Programmkonzepte und -anlagen aus. Für die weiter zurückliegende Vergangenheit kann angenommen werden, dass unzureichende Programmbeschreibungen durch die Programmeigentümer die Ursache waren, zu denen insbesondere hinsichtlich der intendierten Wirkungen der Programm währen des Beabachtungszeitraums Verbesserungen moniert wurden, um gezieltere und tragfähigere Evaluationen durchführen zu können. Für die rezenten Abschnitte des Beobachtungszeitraums kann grundsätzlich von einer tendenziellen Verbesserung dieser Voraussetzungen ausgegangen werden. Es kann anhand der konkreten Berichterstattungen aber doch nicht gesagt werden, dass die Qualität der Bezugnahme auf die Programmkonzeptionen und –strukturen systematisch zugenommen hätte.

Entscheidend erscheint im vorliegenden Zusammenhang, wie Entscheidungen getroffen wurden, welche Untersuchungsdimensionen mithilfe welcher Daten im Verhältnis zur gesamten Programmanlage sinnvoll und ausreichend bearbeitet werden können und wie so jeweils zu

tragfähigen Aussagen gelangt werden kann. Eine explizite Gliederung von Outputs, Outcomes und längerfristigen Wirkungen bzw. von logisch-hierarchisch Wirkungspfaden und -voraussetzungen der evaluierten Programme wird bis auf wenige Ausnahmen, die sich in diese Analyserichtung bewegt haben, nicht vorgenommen. Es liegen einige wenige Programmevaluationen aus den beiden jüngeren Beobachtungsperioden vor, in denen es den EvaluatorInnen gelang, nicht nur Outputs und Outcomes der untersuchten Initiative darzustellen, sondern auch einige Mechanismen zu identifizieren, die für das Handeln der Zielgruppen im Innovationsprozess relevant sind, und damit auch für die genauere Gestaltung des Programms. Hinsichtlich der Programmumsetzung in den betrauten Agenturen wurden Prozessveränderungen greifbar, die durch die Initiative ausgelöst wurden bzw. als positive Wirkungen im Sinne der Zielsetzung der untersuchten Initiative gelten können.

Im Gros der analysierten Programmevaluationen wird allerdings nicht in erkennbarer Weise ein genuin evaluationsmethodisches Konzept angewendet, das gut geeignet ist, um verschiedene Informationen im Verhältnis zu einer strukturierten Programmlogik zu organisieren, womit dann auch bestimmt werden kann, was in der konkreten Vorgehensweise wie gut beleuchtet wird, und was durch zusätzliche Daten noch besser beleuchtet werden könnte bzw. sollte. Spezifische evaluationsmethodische Tools, die zur Auseinandersetzung mit der „Sinnhaftigkeit eines Programms", wie die JC-Standards es ausdrücken, entwickelt wurden, spielen in der FTI-Evaluationspraxis kaum eine Rolle. Logic Charts wurden in nahezu der Hälfte der untersuchten Evaluationsberichte eingesetzt, aber offenbar eher als Präsentationsmittel denn als Analysetools verstanden. Es wird zumindest in der Art der Berichterstattung nicht erkennbar, dass diese Aufarbeitungen der Programmlogiken auch dafür genutzt worden wären, die Ansatzpunkte der Analyse und die damit verbundenen Datenbedürfnisse und geeigneten Methoden zu bestimmen.[9]

Der DeGeval-Standard N4 basiert wesentlich auf dem Konzept, dass in einer Programmevaluation die Beantwortung von Kernfragestellungen im Mittelpunkt stehen sollte und die für die Evaluation vorhandenen Ressourcen zur Datenerhebung entsprechend dieser Prioritäten eingesetzt werden sollten. Die meisten Evaluationsberichte lassen einen solchen Zuschnitt auf Kernfragen nicht erkennen. Wo Evaluationsfragestellungen gelistet oder wenigstens erwähnt werden, ist eine große Zahl an Fragestellungen festzustellen – von 25 bis zu 44 -, die man weitgehend gleichgeordnet behandelt sollte bzw. zu behandeln versuchte. Wesentlicher erscheint noch, dass eine Anzahl von Berichten auch Fragestellungen behandelt, zu denen aus den Darstellungen zu den evaluierten Programmen und den Angaben über die Evaluationsschwerpunkte nicht ersichtlich wird, inwiefern sie dienliche Fragestellungen zur Einschätzung des evaluierten Programms darstellen können. Es handelt sich hier meistens um Regionalverteilungen der zustande gekommenen Projekte, manchmal auch um Geschlechterverteilungen, die deskriptiv behandelt werden, ohne in ein Analysekonzept eingebunden zu sein. Da hier oft auch gleichzeitig das eine oder andere Defizit bei der Heranziehung bestgeeigneter Information für ausgewiesene Untersuchungsschwerpunkte vorliegt, kann geschlossen werden, dass gemäß dem Grundsatz der Ressourcenkonzentration auf die wesentlichsten Fragestellungen besser, und in Bezug auf die Wirkungsweise der Programme erkenntnisreicher, vorgegangen werden hätte können.

Zwei der untersuchten Evaluationen haben ungewöhnliche, innovative Methoden eingesetzt. Insgesamt ist jedoch eine Verankerung in üblichen Methoden festzustellen. So wurde etwa Fachliteratur, die den Operationsbereich des Programms noch näher beleuchten hätte können, um Bedingungen und Einflussfaktoren zu erkennen, die die Erreichung von Zielen beeinflussen konnten, in keinem Evaluationsdesign herangezogen. Konzeptiv in sich geschlossene Analysen wie Quasi-Kotrollgruppendesigns in Bezug auf definierte Programmziele sind selten. Echte Additionalitätsmessungen finden sich nicht, ebenso wenig wie pre-post-Designs oder Bedarfsanalysen bzw. Überprüfungen, inwieweit die programmotivierenden Zielgruppenbedürfnisse oder Systemschwächen zum Evaluationszeitpunkt noch aufrecht waren.

Nur wenige Evaluationen haben in ihrer Thematisierung der Programmereignisse und –wirkungen wissenschaftliche Bezugspunkte aufgesucht, und nicht alle davon haben sie in der angestellten Analyse dann auch effektiv genutzt. Eine Kategorisierung der Evaluationsberichte entlang von wesentlichen, theoretisch begründeten Zugangsweisen zu Forschung und Innovation, wie sie etwa

[9] Sinn und Aufgabe eines „Logic Modelling", zu dem Logic Charts gemeinsam mit anderen Tools zählen, wird in den Ansätzen der sogenannten theoriebasierten Evaluation (*theory-based evaluation*) darin gesehen, ein Programm in der aktiven Auseinandersetzung mit Annahmen über seine Wirklogik analysierbar zu machen (vgl. Kapitel 6).

Barjak in der Untersuchung von innovationspolitischen Programmen in der Schweiz vorgenommen hat (Barjak 2013), wäre für die analysierten Evaluationsberichte gar nicht möglich. In diesem Zusammenhang erscheint auch bezeichnend, dass die in der Umfrage antwortenden EvaluatorInnen nur zu einem marginalen Prozentsatz angeben, dass den von ihnen durchgeführten Programmevaluationen eine Leitwissenschaft oder ein Leitparadigma zugrunde gelegen hat (nur 8% machen inhaltlich verwertbare Angaben, die der Fragestellung entsprechen).

Aus der EvaluatorInnen-Befragung wurden noch weiterführende Hinweise zur Realisierung der Anliegen des Standards N4 erhalten. Dass Auswahl und Umfang der Informationen so bestimmt werden konnten, dass alle vorgesehenen Fragestellungen gut behandelt werden konnten, war aus der Sicht von 73% der EvaluatorInnen in allen von ihnen durchgeführten Programmevaluationen, oder doch zumindest häufig, gegeben. Immer erfüllt war diese wesentliche Voraussetzung für hohe Evaluationsqualität nur aus Sicht von 23%. Ein Viertel bezeichnet die zentrale Forderung des Standards als selten oder nie erfüllt.

Alle antwortenden EvaluatorInnen sprechen davon, dass die Erwartungen der AuftraggeberInnen daran, was unter Berücksichtigung der verfügbaren Ressourcen und der vorgesehenen bzw. zum Einsatz gelangenden Methoden erbracht werden konnte, zumindest in manchen Fällen unrealistisch waren. Immerhin 23% meinen, dass dies in allen von ihnen durchgeführten Programmevaluationen der Fall gewesen sei. Nur 29% geben, dass in den Evaluationsplanungen stets auch gezielte Schritte unternommen wurden, um etwaige Verständnisunterschiede zwischen AuftraggeberInnen und EvaluatorInnen zu klären. In den Gesprächen mit den AuftraggeberInnen wurde deutlich, dass zunehmend erkannt wurde, dass in früheren Entwicklungsphasen der Evalautionspraxis gehegte Vorstellungen über die Leistungskraft von Programmevaluationen im Verhältnis zu verfügbaren Ressourcen oft unrealistisch waren.

Im JC-Standard, der den Interpretationshintergrund für den DeGEval-Standard bildet, wird des Weiteren auf einige Merkmale von Evaluationsprozessen eingegangen, von denen angenommen wird, dass sie gute Voraussetzungen für eine guten Informationsauswahl schaffen. Die wesentlichsten Ergebnisse in diesem Zusammenhang sind:

- Bei der Formulierung und Strukturierung von Evaluationsfragestellungen floss aus Sicht von 68% der EvaluatorInnen auch ihr eigenes Know How ein („immer“ oder „häufig“), sodass auch Fragestellungen Berücksichtigung fanden, die sie als wesentlich erachteten, auch wenn die AuftraggeberInnen zunächst an sie nicht gedacht hatten. In der Erfahrung eines Drittels der EvaluatorInnen war dies hingegen nur selten oder nie der Fall.
- Ein Drittel der EvaluatorInnen gibt an, dass ihre Programmevaluationen nur selten oder nie so angelegt waren, dass sie auch eine Auseinandersetzung mit nicht-intendierten Wirkungen oder unerwünschten Nebenwirkungen des evaluierten Programms ermöglichten. Dies gilt allerdings als integraler Bestandteil einer umfassenden Analyse aller wichtigen Dimensionen eines Programms, für die sowohl auf konzeptiver als auch auf Datenebene die Voraussetzungen geschaffen werden müssen.
- Der JC-Standard bezeichnet es als einen Fallstrick für eine qualitätsvolle Evaluation, wenn bei der Bestimmung des Evaluationsdesigns auf Methoden gesetzt wird, die die AuftraggeberInnen eingesetzt sehen möchten, oder für die die EvaluatorInnen Vorlieben haben. Aus den Angaben derjenigen EvaluatorInnen, die die entsprechenden Fragen beantwortet haben, folgert allerdings, dass solche Vorgehensweisen in der bisherigen FTI-Evaluationspraxis durchaus verbreitet waren bzw. sind. 62% geben an, dass das Evaluationsdesign immer oder zumindest häufig auf Methoden beruhte, die der/die AuftraggeberIn eingesetzt sehen wollte. 92% geben an, dass das Evaluationsdesign immer oder zumindest häufig auf Methoden beruhte, die die EvaluatorIn bzw. sein/ihr Institut regelmäßig einsetzen. Dadurch ist nicht gesagt, dass die eingesetzten Methoden in allen derartigen Fällen nicht auch die passenden bzw. inadäquat gewesen wären. Es sollte aber davon ausgegangen werden, dass ein Risiko vorhanden war und ist, in eingespielten Vorgehensweisen etwaige bessere Alternativen zu übersehen.
- Ein wesentlicher Stellenwert für das bestmögliche Gelingen einer Programmevaluation wird einer schrittweisen Ausgestaltung der Evaluation und einer von vornherein eingeplanten Flexibilität zugemessen, um während der Durchführung auf veränderte Umstände reagieren zu können (z.B. unerwartete Datenlage, überraschende Zwischenergebnisse, Veränderungen des untersuchten Programms, Veränderungen der Informationsbedürfnisse). Lediglich 18% der EvaluatorInnen attestieren allen von ihnen durchgeführten Evaluationen eine solche Flexibilität.

40% geben hingegen an, dass diese Voraussetzung für eine optimale Evaluationsdurchführung nur selten oder nie gegeben war.

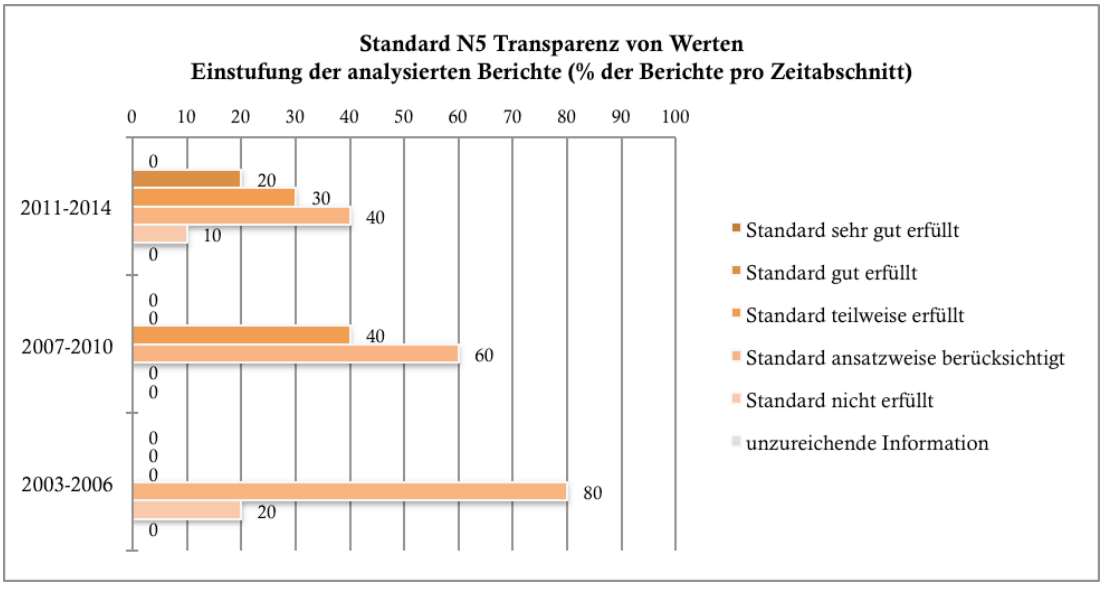
Auftraggeberseitig war zum Teil von *„Suchprozessen"* die Rede, wenn die Erfahrungen mit bisherigen Porgrammevaluationen zusammengefasst wurden. Bezeichnend erscheint auch die folgende Aussage: *„Da braucht es viele Operationalisierungsschritte im Programmdesign, die selten geleistet werden. Viele Evaluierungen evaluieren dann relativ vage vor einer Zielkulisse, die nicht eindeutig ist. Dann greifen sie zurück auf Plausibilitätsüberlegungen, Konsistenzfragen, Zufriedenheitsfragen, Wahrnehmungsfragen, Erwartungsfragen. Dann bekommt man ein Bild, das dann eher so ausschaut: ja, es macht irgendwie Sinn unterm Strich, die und die Aspekte scheinen gut zu funktionieren, andere weniger." (A1)* Ein/e GesprächspartnerIn hat darauf hingewiesen, dass in manchen Evaluationen sachdienliche Daten von den EvaluatorInnen nicht herangezogen wurden. Aufschlussreich erscheint auch die folgende Aussage zur Findung von Evaluationsdesigns: *„Da gibt es unterschiedliche Stile unter EvaluatorInnen. Da gibt es einen Stil, der versucht, die TORs recht ernst zu nehmen und brav abzuarbeiten. Dann gibt es den Stil der etwas Selbstbewussteren, die hin und wieder auch die Terms reflektieren und (...) durchaus versuchen, die TORs als Startpunkt zu nehmen, aber damit weiterarbeiten und versuchen, neue Kontexte herzustellen. Die Frage ist immer, wer setzt sich im Wettbewerb durch? Und das liegt dann wieder beim Auftraggeber - will er mehr mit den Selbstbewussteren mitgehen, oder bleibt er wirklich fest auf seinen TORs. Beides ist legitim, beides gibt es in Österreich am Markt." (A2)*

Insgesamt ist zu diagnostizieren, dass der Umgang mit Aspekten der Evaluationsqualität, die gemäß Standard N4 eine wesentliche Rolle für die spätere Nützlichkeit der Evaluation spielen, in der bisherigen Evaluationspraxis im FTI-Bereich unter einem optimalen Niveau liegt. Evaluationsanlagen erscheinen oft kompromisshaft zwischen Herangehensweisen einer umfassenden Evaluation (*comprehensive evaluation*), die aus der allgemeinen Charakterisierung der Evaluationsprojekte und ihren Benennungen der Zwecksetzungen nahe gelegt ist, und einer gezielt auf eingegrenzte Programmaspekte zugeschnittenen Evaluation (*tailored evaluation*), was zwar nicht umfassend konzipierten und umgesetzten Analysen entspräche, zu der sich die Evaluationsberichte jedoch nicht explizit bekennen. Verbesserungspotenzial im Verhältnis zu einem internationalen, Politkfeld-übergreifenden *State-of-the-Art* existiert und sollte künftig nach Möglichkeit genutzt werden.

4.6 Transparenz von Werten

Eine gute oder zumindest zufriedenstellende Erfüllung des Standards N5 erweist sich als eine Herausforderung für die im Sample enthaltenen Programmevaluationen, die trotz einer tendenziell positiven Entwicklung über den Beobachtungszeitraum hinweg im Wesentlichen bis zur Gegenwart erhalten bleibt.

Abbildung 19: Ergebnisse der Berichtsanalyse für den Standard N5



Ausgangspunkt des Standards ist ein Verständnis von Evaluation, das diese als aus den drei gleichwertigen Hauptkomponenten der Methodik, der Bewertung und der Nutzung aufgebaut sieht (vgl. Kapitel 1). Im Haupttext der JC-Standards heisst es: „Das Bewerten - die Einschätzung oder Klassifizierung einer Sache nach ihrer Nützlichkeit und ihrem allgemeinen Wert - ist die grundlegende Aufgabe jeder Evaluation. Im Mittelpunkt dieser Aufgabe steht das Erfordernis, die bei einer Evaluation gewonnenen Informationen zu interpretieren. Solche Informationen - ob quantitativ oder qualitativ, prozess- oder produktbezogen, formativ oder summativ - sind nur von geringem Interesse oder Nutzen, wenn sie nicht anhand einer geeigneten und vertretbaren Idee dazu, was Wert hat und was nicht, interpretiert werden." In den analysierten Evaluationsberichten ist eine derartige Grundperspektive freilich kaum zu bemerken.

In den meisten Evaluationsberichten werden einzelne Programmaspekte isoliert für sich bewertet, wobei allgemein eingespielte und scheinbar selbstverständliche Sichtweisen tonangebend sind (z.B. Zufriedenheit der Fördernehmer mit der Beratung und Förderabwicklung, Alleinstellungsmerkmal eines Programms von der Anlage her). Wie sich diese Bewertungen zu einem konzisen Gesamtbild fügen, und welcher Betrachtungswinkel letzendes ausschlaggebend wird, um dem Programm größeren oder geringeren Wert zuzusprechen, verdankt sich dabei meist nicht erkennbaren übergreifenden Systematiken, wie sie etwa bei einer konsequenten Messung von Zielerreichungen vorliegen. Des Öfteren ist die vorfindliche Ergebnisinterpretation dadurch gekennzeichnet, dass sie quer über verschiedene Berichtsabschnitte, oder auch im Umgang mit einzelnen Datenlagen, unentschieden zwischen mehreren angelegten Bewertungsmaßstäben bzw. der Einordnung eines Faktums ist. Bisweilen stimmen die Bewertungsmasstäbe, die in der Dateninterpretation angelegt werden, mit denen, die in den Schlussfolgerungen tragend werden, nicht überein.

Einige Evaluationen wurden von Überprüfungen von Zielerreichungen getragen, soweit das zum Evaluationszeitpunkt überhaupt möglich war. Dabei wurden allerdings bis auf eine Ausnahme keine konkreten Schwellenwerte bestimmt, um die späterhin gemachten Beobachtungen gezielt einzuordnen. Im Umgang mit anderen Aspekten der Programmentfaltung, wie etwa der Zufriedenheit von Fördernehmern mit der Betreuung durch die Agentur, wurden vorab festgelegte Kriterien nicht eingesetzt, oder zumindest nicht berichtet.

Die Frage des Bewertungsmaßstabs siedelt sich des Öfteren auf der Ebene von Abwägungen an, welcher Maßstab geeignet sein könnte. Entscheidungen, die diesbezüglich de facto bei der Auswahl von Daten oder in der Ergebnisinterpretation getroffen werden, werden jedoch nicht speziell begründet. Derartige Fälle liegen etwa vor, wenn es angesichts eines hohen Stellenwerts von Kooperationen im evaluierten Programm entweder um die Quantität von Kooperationen oder um die Qualität von eingegangenen Kooperationen geht, oder wenn die Fördermotivationen mit der Bewältigung von Risiko verbunden wird, dann aber parallele Überlegungen zur Auslösung einer bestimmten Innovationsaktivität schlagend werden.

In der Befragung der EvaluatorInnen wurden mehrere Fragen gestellt, die den Impetus des Standards anhand von Formulierungen des JC-Standards näher beleuchten. Die folgenden Ergebnisse wurden dabei erhalten:

- Die FTI-Programmevaluationen werden in hohem Ausmaß als wertfrei betrachtet. Eine derartige Herangehensweise ignoriert freilich die Bewertungsproblematik, und die Standards warnen vor ihr. 84% der EvaluatorInnen, die diese Frage beantworten, betrachten die von ihnen durchgeführten Evaluationen als immer oder zumindest häufig als wertfrei. Lediglich 4% geben an, dass die durchgeführten Evaluationen nie wertfrei waren, was der konzeptuellen Herangehensweise des Standards direkt entspricht.
- Die JC-Standards besagen: „Der Kernpunkt dieses Standards ist, dass Evaluatorinnen und ihre Auftraggeber zusammen mit den verschiedenen Beteiligten und Betroffenen sorgfältig festlegen sollten, welcher Ansatz zugrunde gelegt werden soll, um den gewonnenen Informationen Wert zuzuweisen. Den gewählten Ansatz sollten sie dann offenlegen und begründen." Dass in einer Frühphase der Evaluation in Abstimmung mit den Auftraggeberinnen festgelegt wurde, welche Wertmaßstäbe später herangezogen werden sollten, um die Ergebnisse zu interpretieren, war 42% der antwortenden EvaluatorInnen zufolge in den durchgeführten FTI-Programmevaluationen immer oder zumindest häufig verwirklicht. Allerdings attestieren nur 4%, dass ein solcher wesentlicher Klärungsschritt immer erfolgte. Hingegen gibt fast ein Drittel (29%) an, dass dies nie der Fall war. Viel eher wurden in der bisherigen Evaluationspraxis die EvaluatorInnen als alleine dafür zuständig betrachtet, Wertmaßstäbe an die untersuchten Programme zu finden und die aufgearbeiteten Fakten anhand korrespondierender Kriterien zu

bewerten. 80% aller auf die Frage antwortenden EvaluatorInnen geben an, dass dies immer oder zumindest häufig der Fall war. Lediglich 4% berichten, dass eine völlige Überantwortung der Bewertungsfrage an die EvaluatorInnen in den von ihnen durchgeführten FTI-Programmevaluationen nie vorkam.

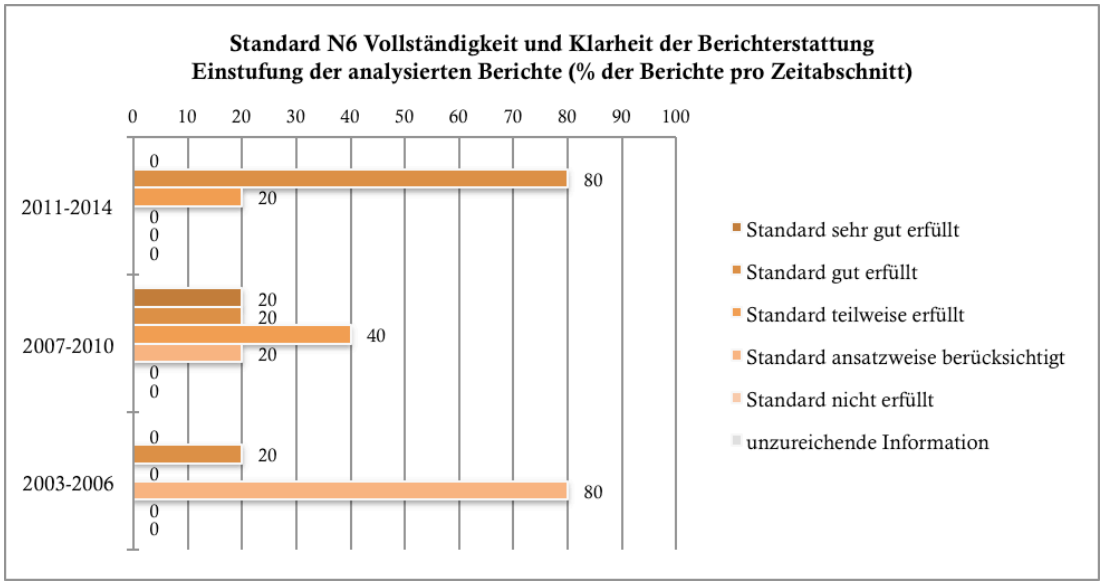
- In den Bewertungsverfahren der FTI-Programmevaluationen wurden über weite Strecken Maßstäbe und Kriterien angewendet, die die AuftraggeberInnen vorgaben bzw. die aus Sicht der EvaluatorInnen den Maßstäben und Kriterien ihrer jeweiligen AuftraggeberInnen entsprachen. 54% geben an, dass so immer oder zumindest häufig vorgegangen wurde. Ein Viertel (25%) gibt zugleich an, dass dies nie der Fall gewesen ist. 44% der antwortenden EvaluatorInnen berichten hingegen, dass Bewertungsmaßstäbe, die bei den Auftraggeberinnen oder bei anderen Stakeholdern des Programms möglicherweise existierten, für die konkrete Vorgehensweise der durchgeführten FTI-Programmevaluationen nicht wesentlich waren. Dies entspricht einem objektivistischen Evaluationsansatz, der nach Auffassung wesentlicher Evaluations-theoretikerInnen aber auch Gefahr läuft, mit der Entfernung von Sichtweisen der Programm-beteiligten von diesen weniger leicht akzeptiert zu werden. Über die Systematik der eingeschlagenen Vorgangsweisen ist damit noch nichts gesagt.
- Die JC-Standards enthalten die Überlegung, dass Bewertungen nicht zwingend von den EvaluatorInnen vorgenommen werden müssen. Ebenso können Verfahren gewählt werden, in denen die von den EvaluatorInnen vorgelegten Diagnosen auf Faktenebene einem getrennten Bewertungsverfahren gemeinsam mit den AuftraggeberInnen zugeführt werden, in die potenziell auch Stkeholder einbezogen werden könnten. Gleichfalls erwähnt werden Verfahren, in denen Steuerungsgruppen der Evaluationen oder eigens eingerichtete Jurys die Bewertung vornehmen. Gemäß den Angaben der EvaluatorInnen, die auf die korrespondierende Frage antworten, kommen solche Vorgehensweisen in der FTI-Evaluationspraxis gelegentlich vor, stellen aber doch im Gesamtbild die Ausnahme dar. Ein gutes Drittel (38%) gibt an, dass derartige Verfahren bei den von ihnen durchgeführten Programmevaluationen nie angewendet wurden. Hinsichtlich positiver Antworten muss dahingestellt bleiben, ob die antwortenden EvaluatorInnen nicht teilweise auch Vorgänge im Rahmen der üblichen Abstimmungsverfahren bei der Abnahme der Endberichte unter den in der Frageformulierung angesprochenen Tatbestand subsumiert haben, sodass sich ihre Antworten nicht ausschließlich auf von Anfang an verfahrenstechnisch vorgesehene und von der Evaluationserarbeitung klar getrennte Schritte beziehen.

Seitens der AuftraggeberInnen wird die Frage der gezielten Wahl von Bewertungsmaßstäben offenbar unterschiedlich betrachtet. Die Wahl der genauen Evaluationskriterien wird offenbar oft der Formulierung von Evalautionsfragestellungen zugeschlagen, und die EvaluatorInnen-Angaben, dass die Vornahme von Bewertungen als deren alleinige Aufgabe betrachtet wird, wurde in merheren Gesprächen bestätigt. Teilweise wurde die Frage von Bewertungsmaßstäben aber auch ausdrücklich angesprochen, wie es im folgenden Zitat zum Ausdruck kommt. *„Wenn ich die Frage habe: die Innovationskraft der Unternehmen als Zielgröße der FTI-Strategie, dann muss ich (...) definieren: was ist Innovationskraft.“ (M1)* Ein/e GesprächspartnerIn beschreibt einen hohen Pragmatismus der evaluativen Vorgehensweisen, der offenbar retrospektiv auch mit einem gewissen Ausmaß an Frustration einhergeht: *„Grundsätzlich müssen ja in einer Ausschreibung einer Evaluierung Fragestellungen drinstehen. Die versucht man dann zu beantworten. (...) Die Aussage, die der Evaluator dann trifft, ist sicher irgendwo eine Bewertung. Da wird er hoffentlich ein paar Evidenzen gefunden haben, aus denen man schlüssig eine Antwort konstruiert. (...) Da ist es schon so, dass die Wahl der Evidenzen stark beeinflusst wird davon, was mit vernünftigem Aufwand greifbar ist. Da unterstelle ich jetzt sonst keinen gezielten Zugang.“ (A2)*

4.7 Vollständigkeit und Klarheit der Berichterstattung

Für die Erfüllung des Standards N6 zeigt sich ein kontinuierlich steigendes Niveau. Es sind jedoch immer wieder auch gewisse Schwächen der Präsentationen auszumachen, sodass für den jüngsten Zeitabschnitt die bestmögliche Einstufung in keinem Fall vorgenommen werden konnte.

Abbildung 20: Ergebnisse der Berichtsanalyse für den Standard N6



Die im Sample enthaltenen Berichte sind in aller Regel in einer gut verständlichen Sprache abgefasst, die die BerichtsadressatInnen nicht mit Technizismen belastet. Sie präsentieren wesentliche Informationen in Tabellen und Grafiken, die oft, aber nicht in allen Fällen anhand von Tabellen- und Grafikverzeichnissen leicht auffindbar sind. In seltenen Fällen veranschaulichen Grafiken auch konzeptive Inhalte, und bisweilen werden Grafiken auch zur Vorstellung der Vorgehensweise der Evaluation eingesetzt. Textgestaltungen im Layout, die auch in den Fließtexten eine weitere Gliederung herbeiführen und zentrale Botschaften herausheben, sind gelegentlich anzutreffen, bilden aber insgesamt die Ausnahme.

Die Lesbarkeit der Evaluationsberichte ist im Allgemeinen als gut zu bezeichnen. Manchmal leidet sie allerdings an einer schlechten Gliederung der Beobachtungen und Argumente, die zu Dateninterpretationen und Schlussfolgerungen herangezogenen werden. In zwei Ausnahmefällen wird in anspruchsvollem wissenschaftlichem Duktus berichtet, wodurch Gruppen von BerichtsadressatInnen überfordert werden könnten, die nicht laufend mit wissenschaftlichen Diskussionen umgehen und nur wenig Zeit in die Berichtslektüre investieren können.

Zwei Drittel der analysierten Evaluationsberichte ist ein Executive Summary beigegeben. In zwei Fällen wurde es auch in englischer Sprache verfügbar gemacht, sodass die Evaluation zumindest hinsichtlich ihrer zentralsten Eckpunkte über den deutschen Sprachraum hinaus bekannt werden und aufgegriffen werden kann. In ihrem Inhalt und bezüglich der Qualität der Vermittlung des Berichtsinhalts variieren die Executive Summaries deutlich. In manchen Fällen werden das evaluierte Programm und zentrale Ergebnisse in den Vordergrund gestellt, in anderen Fällen die Schlussfolgerungen und Empfehlungen. Eine Strukturierung nach Evaluationsfragestellungen stellt den seltenen Ausnahmefall dar. Es besteht auch einige Varianz hinsichtlich einer gesamthaften Vermittlung dessen, was als Gesamtgehalt der komplexen Untersuchungen erachtet werden kann. Dies kann in all jenen Fällen nachteilig sein, wo die Zeitressourcen der BerichtsadressatInnen knapp sind und es nicht zu einem Zugriff auf die ausführlichen Berichte kommt. In einigen wenigen Fällen wurden ausführliche Executive Summaries erstellt, die eher schon Kurzberichte darstellen. Während diese heute oft erhobenen Ansprüchen nach extrem verknappter und sehr rasch auffassbarer Information nicht vollständig Genüge tun, kommt es doch gerade durch diese Vorgehensweise zu einer kompakten Kurzdarstellung, die dem Gesamtinhalt der jeweiligen Evaluation besser gerecht wird.

In den Berichten selbst ist eine Konzentration auf die Vermittlung der wesentlichen Inhalte und Ergebnisse in unterschiedlichem Ausmaß vorzufinden. In einigen Berichten aus allen drei Beobachtungsperioden finden sich zumindest einzelne Abschnitte, wo der Forderung nach fokussierten und möglichst nicht weitschweifigen Berichten nicht gut entsprochen wurde. In einem Bericht neueren Datums werden die LeserInnen in eine umfängliche Präsentation von wenig aussagekräftigen Daten hineingezogen, und eine deutliche Raffung auf den wesentlichsten Gehalt wäre hier vorzuziehen gewesen.

Auch die Darstellung der evaluierten Programme variiert deutlich. Einige Berichte geben klare und eingehende Darstellungen des evaluierten Programms in seiner Motivation und seinem Entstehungskontext, die es auch bislang nicht mit dem Programm Vertrauten ermöglicht, den Evaluationsgegenstand zu verstehen und der Berichterstattung über die Evaluation zu folgen. In anderen Berichten erfolgt die Darstellung des evaluierten Programms abrisshaft, teils auch in Bezug auf die Programmziele. In einer geringen Anzahl von Berichten wurde auf eine Programmdarstellung fast vollständig verzichtet. Oft wird Information über das evaluierte Programm bruchstückhaft und verteilt über verschiedene Berichtsabschnitte vermittelt, und nicht immer sind solche verteilten Darstellungen dann auch vollkommen konsistent. Eine nicht unbeträchtliche Anzahl von Berichten ist in einer Weise gestaltet, als ob die BerichtsaddressatInnen das untersuchte Programm, Kontextprogramme oder auch im FTI-System vorhandene Informationen bereits kennen. Eine solche Berichterstattung kann für AuftraggeberInnen und HauptadressatInnen akzeptiert werden, die über die entsprechenden Informationen tatsächlich verfügen. Die Verständlichkeit und Nachvollziehberkeit dieser Programmevaluationen ist jedoch für Nicht-InsiderInnen deutlich eingeschränkt.

Eine Logic Chart erfüllt in knapp der Hälfte der Berichte die Funktion eines Präsentationsmittels, das wesentliche Information über die Programmkonfiguration übersichtlich versammelt. Vergleicht man diese Präsentationen mit anderen, textlichen Angaben über das Programm, so ergibt sich freilich des Öfteren, dass doch keine vollständige oder wirklich konzis strukturierte Programmlogik vermittelt wird (vgl. hierzu auch die Erläuterung zum Standard N4). In Fällen, wo das evaluierte Programm während seiner Laufzeit bis zum Evaluationszeitpunkt Veränderungen erfahren hat, wird teils auf diese Veränderungen eingegangen, teils wird die genaue Verfasstheit des Programms in seinen unterschiedlichen Entwicklungsstadien nicht greifbar.

In einigen besseren Berichten entspricht die Kapitelstruktur einem logisch gegliederten und stringenten Aufbau der Untersuchung bzw. Ergebnispräsentation, etwa nach Programmelementen bzw. Segmenten der Programmlogik. In schlechteren Berichten ist die Darstellungsweise von Untersuchungsschritten trotz einer auf den ersten Blick gut anmutenden Strukturierung nicht kompakt, sodass BerichtsaddressatInnen Informationen, Argumente und Hinweise doch quer über unterschiedliche Berichtsabschnitte hinweg vorgelegt bekommen und zu einer vollständigen Auffassung von Analyseschritten zwischen verschiedenen Berichtsteilen hin und her geblättert werden muss. Eine konsequente Strukturierung nach den Untersuchungsebenen von Inputs, Outputs, Outomes und allenfalls bereits beobachtbaren ersten Impacts stellt eher die Ausnahme als die Regel dar. Eine ganze Anzahl von Berichten hat es vorgezogen, Ergebnisse in einer Gliederung nach den eingesetzten Methoden zu präsentieren, wodurch die Auffassbarkeit der Ergebnisse hinsichtlich ihres systematischen Bezugs auf die Programmlogiken erschwert wird.

Wiederholt festzustellen ist, dass sich die EvaluatorInnen an manchen Stellen, und teils auch in Passagen mit schlussfolgerndem Charakter, nicht so ausdrücken, dass jedenfalls zweifelsfrei von allen AdressatInnen verstanden wird, welchen Stellenwert sie bestimmten Ergebnissen zumessen und zu welchen Ansichten sie gelangt sind. In einer deskriptiven Herangehensweise werden bisweilen Ergebnisse in detaillierter Weise präsentiert, ohne aber eine Interpretation anzubieten und so den BerichtsaddressatInnen zu vermitteln, wie diese Daten aus Sicht der EvaluatorInnen zu verstehen sind. Manche Berichte halten sich passagenweise in in den Raum gestellten Hinweisen auf, bei denen es den AdressatInnen überlassen bleibt, welche Bedeutung sie ihnen beimessen und welche Schlüsse sie daraus ziehen. Gelegentlich werden in Berichtstexten Ergebnisse mit verschiedenen Überlegungen verbunden, die in ihrem Vortrag den Status dieser Ergebnisse in einem analytischen und bewertenden Konzept (vgl. N4, N5) und ihre Lesart durch die EvalutorInnen nicht leicht nachvollziehbar machen bzw. zu keinem klaren, den AdressatInnen unmissverständlich angebotenen Resultat führen. Dass wiederholt auch implizit verbleibende Fragestellungen verhandelt werden, verleiht einigen Berichten auch neueren Datums ein Stück weit den Charakter eines Insider-Berichts an die AuftraggeberInnen, der nur von diesen vollständig aufgefasst werden kann.

In zahlreichen Berichten finden sich neben den explizit ausgewiesenen Schlussfolgerungen auch in verschiedene Textpassagen eingestreute Schlussfolgerungen und Empfehlungen. Hier erleichtert zwar der unmittelbare Konnex zur vorangegangenen Argumentation die Auffassbarkeit, doch setzt die Kenntnisnahme aller Schlussfolgerungen und Empfehlungen hier auch die genaue Lektüre des

gesamten Berichts voraus (zur Qualität von Schlussfolgerungen vgl. auch Ausführungen zum Standard G8 weiter unten.)

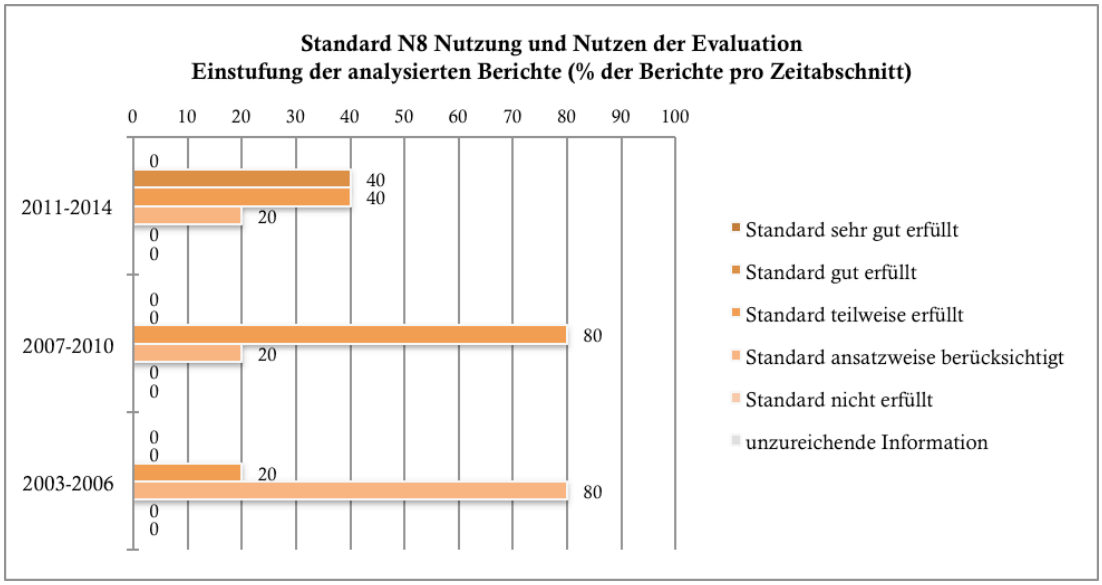
Von Seiten der interviewten AuftraggeberInnen wurde wiederholt darauf hingewiesen, dass Berichte in ihren Aussagen nicht klar waren. Dies betrifft auch Schlussfolgerungen und Empfehlungen, die in mehreren Gesprächen als gelegentlich aussagearm charakterisiert wurden.

Die Forderung nach Klarheit und Vollständigkeit der Berichterstattung betrifft schließlich auch die Methodendarstellung, die klar und verständlich sein soll, wobei sich die Auftrennung in nicht-technisch formulierte Berichtsinhalte und technische Anhänge anbietet. Hierzu kann festgestellt werden, dass die untersuchten Berichte zwar in aller Regel Anhänge enthalten, in denen Zusatzinformationen zu den eingesetzten Methoden angeboten werden (z.B. Fragebögen, Listen von InterviewpartnerInnen), aber eine systematische Auslagerung eher technischer Informationen in dieser Form nicht vorliegt. Da die HauptadressatInnen dieser Berichte über Jahre hinweg regelmäßig mit Evaluationsberichten umgehen und in der *Plattform fteval* mit Methodendiskussionen befasst sind, kann davon ausgegangen werden, dass die Unterbringung eines Gr0ßteils der methodisch-technischen Informationen im Bericht selbst, die regelmäßig anzutreffen ist, grundsätzlich keine Beeinträchtigung der Verständlichkeit der Berichte für diesen primären Adressatenkreis darstellt. Es wurden auch im Rahmen der Interviews keinerlei Hinweise erhalten, dass auf dieser Ebene Probleme wahrgenommen würden.

4.8 Nutzung und Nutzen der Evaluation

Für den zentralen Standard im Sinne der in der vorliegenden Metaevaluation thematisierten Nützlichkeit ergibt sich eine Einschätzung auf einem befriedigenden und auch in den letzten Jahren steigenden Niveau. Es muss zugleich angemerkt werden, dass die Beurteilbarkeit der Erfüllung des Standards auf Basis der Evaluationsberichte nur eingeschränkt gegeben war, da hier stark Merkmale des jeweiligen Evaluationsprozesses zum Tragen kommen, zu denen sich die untersuchten Evaluationsberichte nicht äußern. Insbesondere aus den JC-Standards ist zu entnehmen, dass sich die Intention des Standards stark darauf richtet, dass von der Evaluationsplanung weg eine gute Kommunikation zwischen den EvaluatorInnen und den AuftraggeberInnen sowie den weiteren Beteiligten und Betroffenen des untersuchten Programms im Hinblick auf den intendierten Nutzen gewährleistet wird.

Abbildung 21: Ergebnisse der Berichtsanalyse zum Standard N8



Aus der Formulierung des Standards ergibt sich, dass die Nützlichkeit, die einer Evaluation zugemessen werden kann, als Resultat der Erfüllung aller anderen Nützlichkeitsstandards sowie auch von Standards aus anderen Gruppen aufzufassen ist. Ein holistisches Qualitätsverständnis kommt zum Ausdruck, dem zufolge letztlich kein Gesichtspunkt einer umsichtigen Planung, Durchführung und Präsentation einer Evalaution vernachlässigt werden kann.

Der Standard wurde so angewendet, dass er sich auf jene Informationen stützt, die aus den Berichten heraus zugänglich werden:

- Inwieweit anhand der Berichtsinhalte davon ausgegangen werden kann, dass Stakeholder (Beteiligte und Betroffene) des Programms überhaupt soweit einbezogen wurden, dass eine Auslösung von Interesse an der Evaluation und ihren Ergebnissen sinnvoll angenommen werden kann, um eine Basis für eine mögliche Nutzung zu bilden;
- Funktion aus den Ergebnissen zu allen herangezogenen Standards.

Beide Komponenten sind, wie bereits festgestellt, auf Grund der konkret vorliegenden Berichtsinhalte in ihren Informationsgrundlagen eingeschränkt.

In den in der EvaluatorInnen-Befragung erhaltenen Angaben zeigt sich, dass gezielte Schritte hin auf eine Nutzung bislang stark auf die direkten AuftraggeberInnen der Evaluation konzentriert waren. Verschiedene Optionen, die eine Kommunikation mit den AuftraggeberInnen während der Evaluationsdurchführung betreffen, werden von jeweils mindestens einem Drittel der EvaluatorInnen als in allen von ihnen durchgeführten Programmevaluationen gesetzte Schritte bezeichnet, wobei auch Werte von bis zu 70% für einzelne Optionen erreicht werden. Die entsprechenden Werte für weitere Stakeholder der Evaluationen bewegen sich hingegen zwischen 3% und 20%. Kehrt man die Perspektive um, so ergibt sich, dass neben den AuftraggeberInnen bestehende Stakeholder der Programmevaluationen in der bisherigen Evaluationspraxis nicht völlig vernachlässigt wurden. Dass die verschiedenen Optionen zur Interessenförderung überhaupt nicht genutzt wurden, wird zu maximal 30% angegeben. In der Kommunikation mit den AuftraggeberInnen kam die am seltensten genutzte Kommunikation lediglich 13% der EvaluatorInnen zufolge in den von ihnen durchgeführten Programmevaluationen nie zum Einsatz.

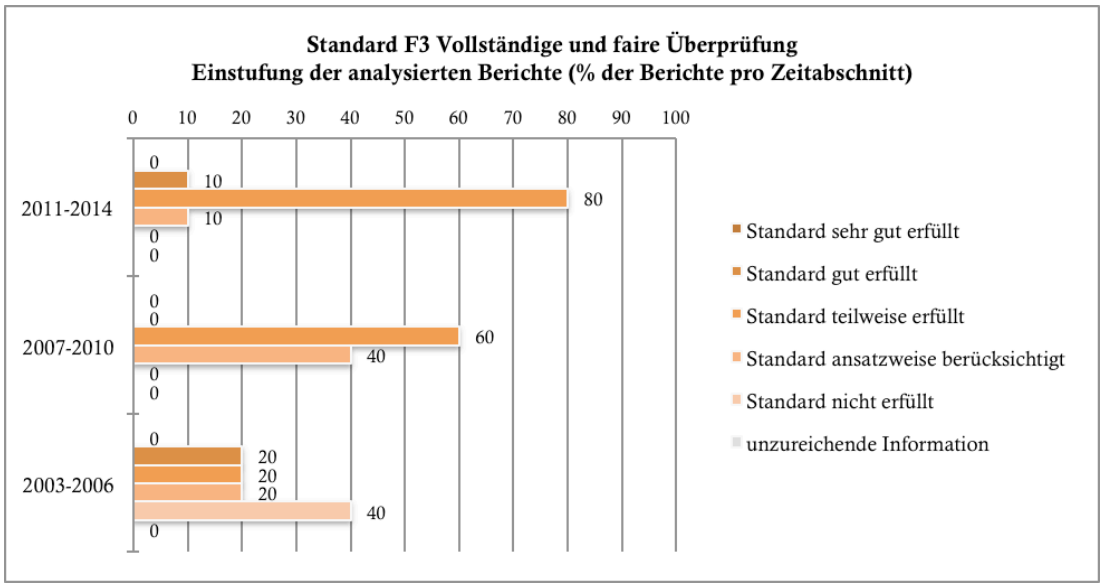
Im Gesamtbild bewegt sich die Kommunikation mit den AuftraggeberInnen auf einem positiven Niveau, das allerdings augenscheinlich auch noch einigen Spielraum für weitere Verbesserungen offen lässt, sodass als zielführend zu erachtende Interaktionen im Resultat durchgehend systematisch zum Einsatz gelangen. Hinsichtlich der Stakeholder der Evaluationen besteht deutliches Verbesserungspotenzial. Verstärkte Einbindungen könnten zu einer effektiveren Verbreitung der Evaluationsergebnisse, zur Anregung von Reflexion und Diskussion im Sinne der Nutzenform der Aufklärung, aber auch zur Erreichung von Zielsetzungen der evaluierten Programme selbst beitragen, indem Programmintentionen und der Beitrag der evaluierten Programme zu den beabsichtigten Veränderungen gemeinsam mit systemreferentiellen und selbstreflexiven Akteuren des FTI-Sytems thematisiert werden.

Die interviewten AuftraggeberInnen erachten die bisher üblichen Kommunikationen während der Durchführung einer Evaluation und anlässlich ihrer Präsentation im Wesentlichen als ausreichend. Die typische Abwicklungsweise und Interaktionsdichte mit den EvaluatorInnen wird etwa von einem/r InterviewpartnerInn so beschrieben: *„Dass man Zwischenschritte vereinbart, dass man am Anfang möglichst klar feststellt was sinnvoll ist, was machbar ist, wie ist die Datenlage - die Evaluatoren schlagen dann meistens die geeigneten Methoden vor -, man muss sich über die Ziele verständigen, wie soll der Bericht aussehen, wie soll das Evaluierungsdesign aussehen, und dann einmal dazwischen anlässlich des Zwischenberichts, und dann am Ende nochmals, damit nicht eine Seite überrascht ist oder Erwartungen enttäuscht werden." (M1)* Aus mehreren Gesprächen konnte entnommen werden, dass im Rahmen solcher eher standardisierten Vorgänge frühzeitige Planungen bzw. Vorkehrungen für die spätere Evaluationsnutzung in der bisherigen Evaluationspraxis nicht Usus waren oder sind. Das Interesse an Evaluationsnutzung ist auf der Ebene unter den GesprächspartnerInnen hoch, wie auch die Auswertungen zu Nutzungsweisen gezeigt haben (vgl. Kapitel 3), doch selten ist diese Evaluationsnutzung auch *„in sich ein Plan" (A1).* In einem Interview wurde freilich herausgestrichen, dass Evaluationsplanung auch als „Erwartungsmanagement" betrachtet wird. *„Man artikuliert in TORs ja auch Erwartungen. Das ist unser wichtigster Punkt, um zu umreißen, was wir eigentlich meinen, was wir wissen wollen, und gleichzeitig, wie wir es aufbereitet haben wollen." (A3)* Hier werden für die Nützlichkeit im Sinne der Standards relevante Gesichtspunkte genannt, die über die bloße Definition eines Evaluationsdesigns als Methodenset klar hinaus gehen. Es wird zugleich auch darauf verwiesen, dass innerhalb der vielfältigen Agenden der befassten Abteilungen und der Agenturen kaum weiterer Spielraum besteht, um die Kommunikationsintensität über das bisher übliche Ausmaß hinaus zu erhöhen.

4.9 Vollständige und faire Überprüfung

Der Standard wurde mit fortschreitender Entwicklung der Evaluationspraxis zunehmend besser erfüllt. Das Niveau, das im Großen und Ganzen vorherrscht, ist als ein zufriedenstellendes zu bezeichnen. Gröbere Schwächen, die in der ersten Untersuchungsperiode gelegentlich vorlagen, treten späterhin nicht mehr auf, doch wird auch eine sehr gute Erfüllung in keinem der analysierten Berichte erreicht.

Abbildung 22: Ergebnisse der Berichtsanalyse für den Standard F3



Die Evaluationsberichte arbeiten durchwegs Stärken und Schwächen der untersuchten Programme heraus. Aus der insgesamt prägenden, lernorientierten Anlage der Programmevaluationen (vgl. N2) ergibt sich gerechtfertigter Weise, dass eine Identifikation von Schwächen der evaluierten Programme, die sodann zu einer zielführenden Adjustierung führen können, öfters im Vordergrund steht. Solche Berichte verfahren kritisch-überprüfend, ohne dass sie deshalb als einseitige „Schwächen-Bericht" bezeichnet werden könnten. Berichte, in denen Ergebnisse oder Schlussfolgerungen entlang von Stärken und Schwächen strukturiert sind, liegen vor, stellen jedoch die Ausnahme dar. Eher wird eine Diktion gepflegt, die Stärken und Schwächen nicht ausdrücklich explizit macht. Es liegen auch einige Berichte vor, die sich in ihren verschiedenen Abschnitten einerseits auf Stärken und an anderen Stellen auf Schwächen konzentrieren, sodass die Herangehens- bzw. Darstellungsweisen zu verschiedenen Evaluationsschwerpunkten oder Programmaspekten unterschiedliche Züge tragen.

Meist sind die Berichte in einer neutralen und strikt faktenorientierten Sprache gehalten. Die EvaluatorInnen stellen in besseren Berichten ihre Beobachtungen neutral dar und teilen in klar getrennten Passagen mit, zu welchen Ansichten über den Evaluationsgegenstand sie auf Grund dieser Beobachtungen kommen und welche Schlussfolgerungen sie daraus ziehen. Der Optimalfall, dass im gesamten Bericht alle Daten bzw. Dateninterpretationen von bewertenden oder schlussfolgernden Formulierungsweisen frei sind, liegt jedoch nur selten vor. Einzelne Berichte haben jedoch auch bei der Vermittlung der Faktenlagen oder bei Darstellungen von Eigenschaften des untersuchten Programms unmittelbar wertende Sprache eingesetzt.

Sowohl Sichtweisen der Programmverantwortlichen als auch Sichtweisen von Fördernehmern werden in aller Regel herangezogen und dargestellt. Völlig objektivistische Untersuchungen, die sich ausschließlich auf objektive Daten stützen und keine Sichtweisen von Programmbeteiligten oder Zielgruppen erhoben haben, kommen so gut wie nicht vor. In einigen Evaluationen wurde über das „Standardset" an Sichtweisen auf den Evaluationsgegenstand hinaus auch weiteren Stakeholdern die Gelegenheit gegeben, sich im Rahmen von Datenerhebungen zum Programm zu äußern. Ein Konzept der wechselseitigen Spiegelung, in dem Kongruenzen wie Inkongruenzen verhandelt werden, ist dabei selten vorhanden. Viel eher anzutreffen ist eine Vorgehensweise, in der ein Teil der Programmlogik

oder ein Evaluationsschwerpunkt anhand der Sichtweisen einer Akteursgruppe analysiert wird, und andere Teile anhand der Sichtweisen einer anderen Gruppe. Es ist denkbar, dass solche Versäumnisse zumindest in Teilen Auswirkungen einer pragmatisch verknappten Datenerhebung sind, innerhalb derer nicht alle relevanten Untersuchungsaspekte für alle untersuchten Akteursgruppen geleichermaßen erhoben werden konnten. Beobachtbar ist auch, dass die EvaluatorInnen verschiedene Sichtweisen zwar darstellen, sie dann aber in unterschiedlichem Maß in ihre Argumentation aufnehmen oder ihnen in der Interpretation zu unterschiedlichen Untersuchungsaspekten unterschiedliches Gewicht verleihen, wobei nicht immer transparente Begründungen mitgeliefert werden. Abwägungen über gleichzeitig vorhandene unterschiedliche Sichtweisen auf die evaluierten Programme von verschiedenen Seiten sind selten anzutreffen. Im Resultat werden die evaluierten Programm nur selten bzw. nur hinsichtlich von Einzelaspekten auch im Spannungsfeld von unterschiedlichen Sichtweisen begriffen, so wie es der Standard empfiehlt.

Ausweise von möglichem Bias oder gezielte Kontrollen dazu finden sich so gut wie nie. In der Analyse von qualitativen Daten wird auch da, wo Zitate ausgewiesen werden, doch oft nicht vermittelt, welcher Stellenwert den berichteten Aussagen in der Gesamtheit aller erhaltenen Aussagen zukommt. Sind somit für manche der analysierten Evaluationsberichte aus der Art der Berichtsformulierung relativ hohe Freiheitsgrade in der Interpretation qualitativer Daten nicht auszuschließen, so werden in manchen schlechteren Berichten auch keine Angaben über Gesamtverteilungen von quantitativen Daten gemacht, sondern lediglich herausgegriffene Einzelergebnisse berichtet, in Ausnahmefällen auch ohne klare Bezifferung. Gelegentlich ist auch das Phänomen anzutreffen, dass EvaluatorInnen textlich nicht klar differenzieren, inwiefern es sich bei Hinweisen auf Stärken oder Schwächen des untersuchten Programms um gesicherte faktengestützte Erkenntnisse handelt, und inwiefern um Hintergrundwissen oder Annahmen der EvaluatorInnen.

Im Rahmen stark deskriptiv angelegter Untersuchungen ist es des Öfteren dazu gekommen, dass Fakten ausgebreitet werden, ohne dass die EvaluatorInnen zu erkennen geben, was sie über diese Fakten denken. Dies kann auch als eine Konsequenz eines Mangels an anderen Daten begriffen werden, die den EvaluatorInnen Hinweise auf Interpretationsmöglichkeiten liefern hätten können, zumal dann, wenn sich die EvaluatorInnen freier Annahmen enthalten. Manchmal werden Ergebnisse lediglich deskriptiv dargestellt und es wird nicht klar, welche Gründe die EvaluatorInnen haben, aus ihnen keine Schlüsse im Rahmen des angewendeten Analysekonzepts zu ziehen.

Grundsätzlich wird im Standard F3 die Perspektive eingenommen, dass sich die tatsächlichen Stärken und Schwächen eines Programms aus allen ihren beabsichtigten und unbeabsichtigten Effekten zusammensetzen. Allerdings meinen doch 36% der EvaluatorInnen in der Befragung, dass die in den von ihnen durchgeführten Programmevaluationen eingesetzten Evaluationsdesigns nur selten oder nie dazu geeignet waren, auch unbeabsichtigte Wirkungen der Programme zu erkennen. Eine explizite Benennung eines unbeabsichtigten Programmeffekts wurde in keinem der analysierten Berichte vorgefunden.

Der hinter dem DeGEval-Standard F3 stehenden Joint Committe-Standard unterhält auch einen Bezug zur Methodendarstellung, da die Methodik einer Evaluation ebenfalls als relevant für die Erzielung von Fairness und einer ausgewogenen Identifikation von Stärken und Schwächen des Evaluationsgegenstands erachtet wird. Da dieser Qualitätsaspekt einer Evaluation in den DeGEval-Standards dem Standard G3 zugeordnet ist, werden Gesichtspunkte der Methodendarstellung auch im Interesse der Übersichtlichkeit in der vorliegenden Metaevalution unter dem Standard G3 behandelt.

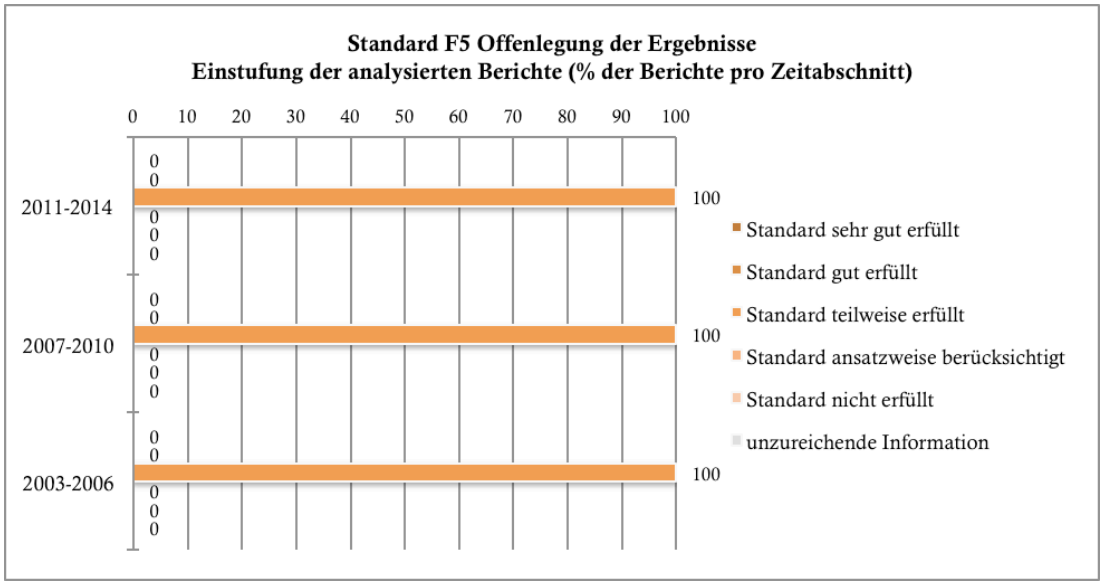
4.10 Offenlegung der Ergebnisse

Im Zentrum des Standards F5 steht die Forderung, dass die Evaluationsergebnisse neben den direkten AuftraggeberInnen auch allen Stakeholdern einer Programmevaluation bzw. eines evaluierten Programms zugänglich gemacht werden sollten. Alle analysierten Berichte enthalten sich jedweder Angabe, ob und inwiefern Schritte unternommen oder geplant wurden, um Evaluationsergebnisse an AdressatInnen auch jenseits der direkten AuftraggeberInnen heranzutragen oder diese AdressatInnen durch speziell zugeschnittene Vorgehensweisen zu unterstützen, die Evaluationsergebnisse aufzufassen und im Weiteren mit ihnen umzugehen. Die einzige Information, die einer Einschätzung im Rahmen der Berichtsanalyse zugrunde gelegt werden kann, ist der Umstand, dass die Berichte auf der Homepage der *Plattform fteval* publiziert wurden und dort abrufbar sind. Hinsichtlich dieses Wegs der Offenlegung muss allerdings auch gesehen werden, dass die Verfügbarkeit der Evaluationsergebnisse für AdressatInnen auf der Homepage der *Plattform fteval* sowohl Wissen um das Vorliegen

des jeweiligen Evaluationsberichts als auch die Kenntnis des Publikationsortes voraussetzt, damit Betroffene und InteressentInnen die Berichte aufrufen und einsehen können. Insofern trifft eine öffentliche Publikation nur bedingt den eigentlichen Gehalt des Standards, der sich auf eine gezielte Informationsvermittlung an Stakeholder richtet, wozu es nicht zwingend einer allgemein zugänglichen Publikation bedarf.

Da einerseits nur unzureichende Informationen vorliegen, und andererseits die genannten Überlegungen gelten, wurde die Erfüllung dieses Standards bei allen analysierten Berichten zunächst pauschal mit der neutralen Einstufung 3 versehen.

Abbildung 23: Ergebnisse der Berichtsanalyse für den Standard F5



Wesentliche ergänzende Hinweise zur Verfügbarmachung der Evaluationsergebnisse ergeben sich aus der EvaluatorInnen-Befragung. Demnach stellen vollständige Publikationen der Evaluationsergebnisse keine durchgängige Praxis dar. Nur 14% der EvaluatorInnen zufolge wurde der Endbericht der von ihnen durchgeführten Programmevaluationen immer in vollständiger Fassung einschließlich aller Anhänge ohne irgendeine Abänderung publiziert. 75% berichten, dass zumindest ab und an Abstriche von einer vollumfänglichen Publikation gemacht wurden, und immerhin 11% geben an, dass eine solche bei keiner der von ihnen durchgeführten Programmevaluationen stattgefunden hat. Ebenso war es in der bisherigen Evaluationspraxis nicht allzu selten so, dass ein eigens auftraggeberseitig produzierter Bericht statt des Original-Endberichts publiziert wurde. Lediglich 62% der EvaluatorInnen schließen dies für die von ihnen durchgeführten Evaluationen ausdrücklich aus („nie“ vorgekommen). Immerhin 12% geben an, dass bei den von ihnen durchgeführten Programmevaluationen „immer“ oder „häufig“ so vorgegangen wurde. Generell unterliegt die Möglichkeit, Informationen aus Progammevaluationen freizugeben oder zu veröffentlichen, sehr stark der Kontrolle der Auftraggeber-Institutionen. 76% der EvaluatorInnen geben an, dass die Kontrolle über die Informationsweitergabe bei allen Progammevaluationen, an denen sie mitgearbeitet haben, ausschließlich bei den Auftraggebern lag. Lediglich ein Viertel meint, dass zumindest in einzelnen Fällen eine Informationsweitergabe prinzipiell möglich war, ohne die Zustimmung der Auftraggeber-Insitutionen einzuholen.

Es wird aus der EvaluatorInnen-Befragung aber auch ersichtlich, dass abgesehen von einer Publikation der Evaluationsberichte durchaus weitere Schritte gesetzt wurden, um Stakeholder jenseits der direkten AuftraggeberInnen direkt zu informieren und ihnen den Umgang mit den Evaluationsergebnissen zu erleichtern. Knapp mehr als die Hälfte der EvaluatorInnen (54%) gibt an, dass Stakeholdern immer oder zumindest häufig der Endbericht direkt übermittelt wurde, oder eine eigens auf sie zugeschnittene Kurzfassung. Ebenfalls mehr als die Hälfte (54%) gibt an, dass gezielte Präsentationen, Diskussionsrunden oder Workshops für Stakeholder immer oder zumindest häufig veranstaltet wurden. 30% sprechen von immer oder häufig stattgefundenen Nachbereitungen, um Unterstützung bei der Interpretation und Anwendung der Ergebnisse und Empfehlungen zu geben.

Anhand in den Interviews mit AuftraggeberInnen der erhaltenen Auskünfte muss dieses Bild allerdings relativiert werden. In mehreren erhaltenen Aussagen drückt sich ein Verständnis aus, dem gemäß die AuftraggeberInnen der Programmevaluationen sich nicht für eine Weitergabe von Evaluationsergebnissen über die Principal-Agent-Verhältnisse hinaus zuständig fühlen, in denen sich die evaluierten Programme ansiedeln. In der jüngeren Vergangenheit haben auch Präsentationen und Diskussionen über diesen Kreis hinaus in einzelnen Ressorts stattgefunden. Während von allen Seiten davon ausgegangen wird, dass die Kommunikation von Evaluationsergebnissen und über deren mögliche Konsequenzen innerhalb der Principal-Agent-Beziehungen im Allgemeinen gut ist, verläuft offensichtlich eine Trennlinie gegenüber Akteursgruppen, die in ihrer Eigenschaft als Zielgruppen von Programmen außerhalb der essentiellen Kommunikationszone zur Programmgestaltung wahrgenommen werden. Bei der Beantwortung der Umfrage durch die EvaluatorInnen dürfte diese eingespielte Auffassung ebenfalls ihre Rolle gespielt haben. Es kommt offenbar auf das genaue Verständnis des Begriffs der „Stakeholder" an, und die relativ hohen Quoten an Stakeholder-Kommunikationen in den Evaluatorinnen-Angaben dürften sich vor allem auf das Gegenüber des jeweiligen Auftraggebers in den Principal-Agent-Beziehungen sowie die ressort-internen Präsentationen beziehen.

Die beschriebene Umgrenzung der unmittelbaren Kommunikationssphäre über Evaluationsergebnisse schließt freilich nicht vollkommen aus, dass Stakeholder die Evaluationsergebnisse zur Kenntnis nehmen und sich damit auseinandersetzen. Von einem Ressort wird berichtet, dass immer wieder Fachverbände, VertreterInnen der Zielgruppen und weitere FTI-Akteure im Einzugsbereich von evaluierten Programmen auf neue Evaluationsberichte hin in Interaktion mit den Programmeigentümern treten. *Grosso modo* zeigt sich allerdings klar, dass ein aktives Herantragen von Erkenntnissen aus Programmevaluationen an die Zielgruppen, die das jeweilige Programm unterstützen soll oder deren Verhaltensweisen es beeinflussen soll, kaum angedacht wurde, wohl nicht zuletzt da es auch Bedenken hinsichtlich einer potenziellen Vermischung mit Interaktionsformen gibt, die dem Lobbyismus zuzurechnen sind.

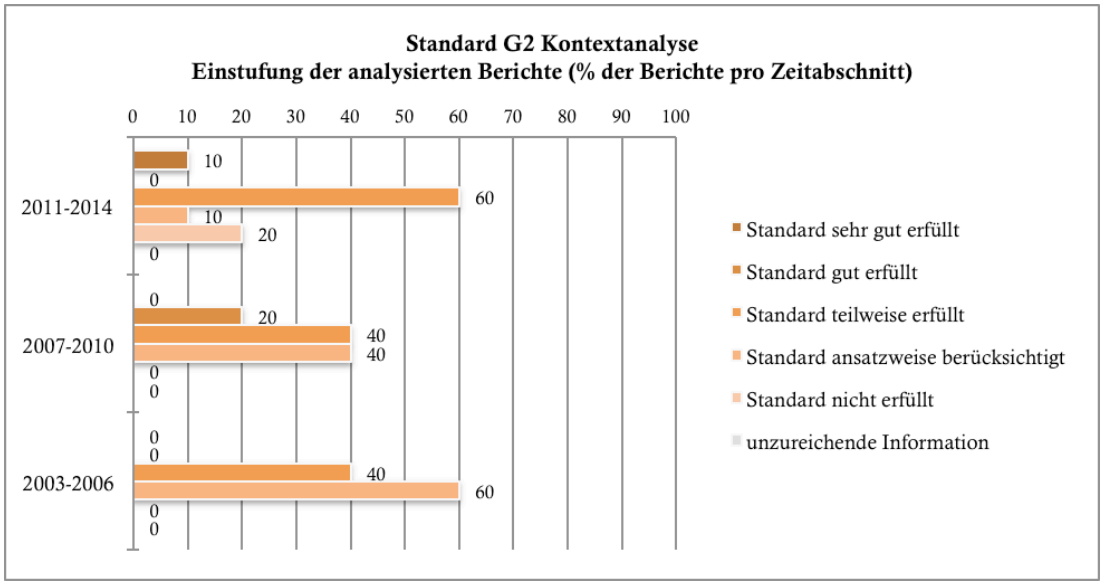
Auf den Umgang mit Evaluationsergebnissen innerhalb der Principal-Agent-Beziehungen und in der aus mehreren Ressorts und Agenturen zusammengesetzten Sphäre der FTI-Politik wird im Kapitel 5 unter dem Stichwort der Zirkulation von evaluativer Information im FTI-politischen Governance-System nochmals eingegangen.

4.11 Kontextanalyse

Die Hinweise des Standards G2 zur Kontextanalyse wurden im am weitesten zurückliegenden Zeitabschnitt des Analysezeitraums eher nur ansatzweise erfüllt. Mit dem Voranschreiten hin zur Gegenwart wurden diese Fingerzeige zunehmend zumindest teilweise erfüllt, und es wurde eine Evaluation erstellt, bei der von einer sehr guten Behandlung nach Maßgabe des in einer einzelnen Programmevaluation Möglichen gesprochen werden kann. Allerdings liegen auch in der jüngsten Vergangenheit Evaluationsberichte bzw. Evaluationen vor, die gerade auch angesichts der von diesen Evaluationen selbst erhobenen Ansprüche eine noch bessere Auseinandersetzung mit dem Kontext der evaluierten Programme wünschenswert erscheinen hätten lassen. Es ist zu erkennen, dass eine ausreichende und fundierte Auseinandersetzung mit dem Kontext der FTI-Programme insgesamt eine permanente Herausforderung für die Programmevaluationen darstellt und künftig noch verstärkt Aufmerksamkeit verdient.

Dass eine Evaluation bzw. ein Evaluationsbericht auf den Kontext, in dem sich das Programm entfaltet, überhaupt nicht Bezug nimmt, kommt äußerst selten vor. Recht unterschiedlich ist allerdings die Intensität bzw. Qualität der erfolgten Auseinandersetzung. Die Programmevaluationen haben ihre Schwerpunkte bei der Analyse von Programmkontexten unterschiedlich gesetzt, und es ist festzustellen, dass mit diesen Fokussierungen auch jeweils Vernachlässigungen anderer Kontextaspekte einhergegangen sind. Eine nur eingeschränkte Verfügbarkeit bzw. Nicht-Verfügbarkeit heranziehbarer Datenbasen sowie eine Sparsamkeit beim Einsatz von zielführenden Methoden hat in vielen Fällen dazu geführt, dass die EvaluatorInnen insbesondere in länger zurückliegenden Evaluationen, aber auch noch bis herauf zur Gegenwart, nicht direkt untersuchte Kontextfaktoren und –bedingungen durch ihr eigenes Hintergrundwissen wett zu machen versucht haben.

Abbildung 24: Ergebnisse der Berichtsanalyse für den Standard G2 Kontextanalyse



Vor allem auf der Basis von qualitativen Untersuchungsstrategien wurden von manchen Evaluationen essentielle Randbedingungen greifbar gemacht, unter denen das jeweilige Programm in seinen Zielgruppen Wirkungen erreichen konnte bzw. daran gehindert war. Hier wurden einzelne Hinweise auf wirksame Mechanismen im Umfeld des Programms bzw. für das Handeln der Zielgruppen erbracht, die für die Erreichung von übergeordneten Programmzielen Bedeutung hatten. Eine wirklich systematische Analyse auf diesem Niveau wurde jedoch kaum erbracht, was auch im Zusammenhang mit einem oft nur ausschnitthaften Aufgreifen der Programmlogiken begriffen werden muss (vgl. dazu die Erläuterungen zum Standard N4). Etliche Kontextanalysen leiden doch darunter, dass zwar einige Faktoren untersucht und für skizzenhafte Bilder fruchtbar gemacht wurden, aber der systematische Stellenwert dieser untersuchten Faktoren unklar bleibt bzw. keinen expliziten Bezug zu einer strukturierten und gesamthaft verstandenen Wirklogik des jeweiligen Programms aufweist.

Eine etwaige innere Differenzierung von Zielgruppen der evaluierten Programme wurde kaum thematisiert. In einem beträchtlichen Teil der Programmevaluationen scheint vielmehr die Annahme vorgeherrscht zu haben, dass die Zielgruppen von Programmen in sich homogen waren bzw. sind. Strukturbedingungen für das Handeln von Zielgruppen werden somit eher nur ansatzweise reflektiert und teilweise unter Heranziehung von Plausibilitätsargumenten behandelt. Nur in Ausnahmefällen wurden die Programme so auf ihren Kontext bezogen, dass durch eine Fokussierung auf die ausgelösten Projekte, die in unterschiedlicher Form zustande kamen und in unterschiedlichen Akteurskonstellationen ihren Verlauf nahmen, eine Relation zu den Umwelten hergestellt wurde, in denen das evaluierte Programm operierte. Schwachpunkte zeigen sich unter anderem bei der Berücksichtigung von vorgesehenen Kooperationspartnern für die Erreichung von Programmerfolg.

Eine Wirkung von Programmen auf ihren Kontext wurde von einer Handvoll der untersuchten Evaluationen verfolgt, vor allem dann, wenn die evaluierten Programme entsprechende Zielsetzungen formuliert hatten. Im günstigsten Fall wurde hier eine Analysestrategie eingesetzt, die gezielt an dem Umstand ansetzte, dass das untersuchte Programm auf Veränderungen in einem Umfeld abzielte, das es nur bedingt und indirekt beeinflussen konnte. Von da aus wurde die passende Fragestellung formuliert, inwiefern das Steuerungsinstrument unter in diesem Umfeld anzutreffenden Handlungsbedingungen als zielführend eingeschätzt werden konnte, und zur Beantwortung Daten zu relevanten Merkmalen und Entwicklungen des Kontextes herangezogen. Dabei wurden auch Einsatzpunkte anderer Steuerungsinstrumente, durch die während der Programmlaufzeit synergetisch auf die Erreichung der Programmziele hingewirkt werden sollte, berücksichtigt. Trotz dieser adäquaten Herangehensweise ist die betreffende Analyse allerdings doch letztlich oberflächlich ausgefallen, da prinzipiell verfügbare weitere Daten, die den Kontext noch genauer ausleuchten hätten lassen, nicht herangezogen wurden.

Immer wieder waren Programmevaluationen bestrebt, Erkenntnisse über die Positionierung des evaluierten Programms in der Förderlandschaft zu erbringen, und gerade in neueren Programmevaluationen wurde diese systemorientierte Frage häufig aufgeworfen. Daneben steht eine Reihe von Programmevaluationen, die potenzielle Überlappungen mit anderen Programmen lediglich angemerkt, aber nicht untersucht haben. Wo Analysen zur Positionierung vorgenommen wurden, gestalteten sich die Herangehensweisen recht unterschiedlich. In manchen Fällen bezogen sie sich nur auf übergeordnete Programmziele und Eckdaten wie mobilisierte Fördersummen und gaben keinen Aufschluss darüber, auf welche Zielgruppen die Programme wie einzuwirken versuchten, und welche Überschneidungen oder Ergänzungen dabei vorliegen konnten. In anderen Fällen begnügten sich die Evaluationen mit einer vergleichenden Zusammenstellung der verschiedenen relevanten Programme von ihrer Anlage her, ohne sich auf eine eingehende Analyse des einzeln und insgesamt ausgelösten Fördergeschehens einzulassen und sich so in die Lage zu versetzen, den konkreten Beitrag des untersuchten Programms im Kontext des Portfolios zu erschließen. In einem Fall beschränkte sich die Analyse auf die Wahrnehmung der im Portfolio enthaltenen Programme durch die Zielgruppe des evaluierten Programms, ohne auch objektive Daten heranzuziehen.

In aller Regel beschränkte sich die Analyse auf Bundesprogramme, während potenzielle Synergien, Überschneidungen und Anknüpfungspunkte auf Regionalebene nicht analysiert wurden. Häufig beschränkte sich der Blickwinkel auf benachbarte Programme im Portfolio der mit der Umsetzung betrauten Agentur, und die Analyse wurde anhand von Förderdaten dieser Agentur beleuchtet. In einem Fall konnten hier Förderkarrieren von Zielgruppen-Segmenten und kumulative Förderungen sichtbar gemacht werden. Die durchgeführte Analysen reicht jedoch auch hier nicht soweit, dass Maßnahmen und Regelungen aus anderen Maßnahmen- und Steuerungsbereichen in ein Gesamtbild einbezogen worden wären. Nur äußerst seltenen wurde das evaluierte Programm auch mit einem thematisch verwandten Programm einer anderen Agentur verglichen, was sich dann aber auf eine Gegenüberstellung der Programmkonzeptionen beschränkte.

Unbefriedigende Herangehensweisen an die Positionierungsfrage lagen da vor, wo lediglich Daten über Fördernehmer oder Antragsteller herangezogen wurden, in denen sich die Programmanlagen und Antragserfordernisse unmittelbar widerspiegeln, oder in denen stark mit Alleinstellungsmerkmalen der evaluierten Programme argumentiert wurde, die von der jeweiligen Programmkonzeption her gegeben waren, aber nicht anhand des konkret beobachtbaren Operierens des Programms auf ihre Einlösung in der Programmwirklichkeit hin untersucht wurden.

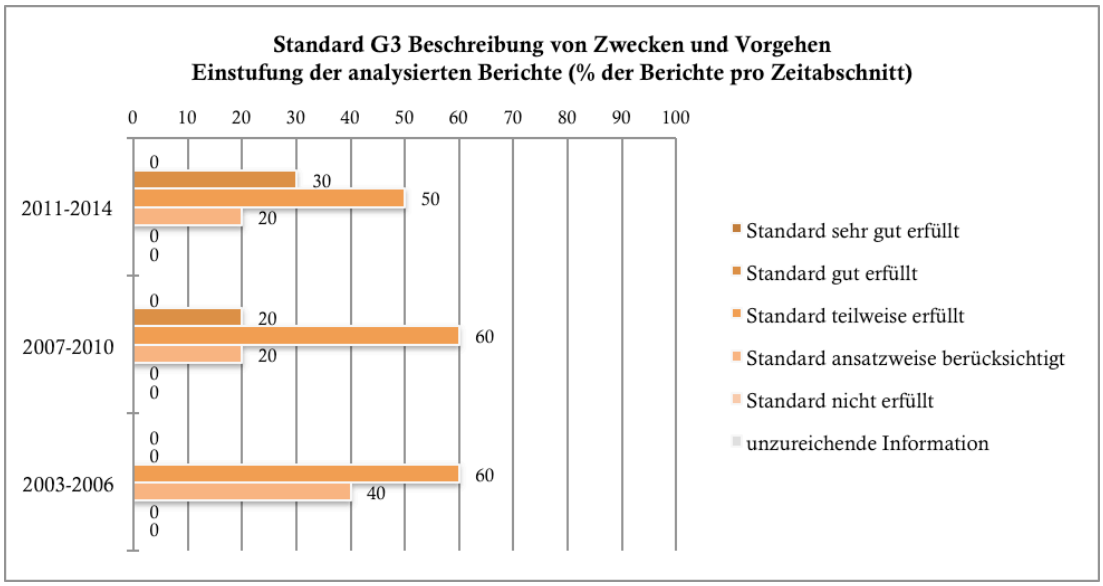
Wiederholt wurden Programmanlagen international kontextualisiert, indem einige Vergleichsprogramme mit ähnlichen thematischen Schwerpunktsetzungen beleuchtet werden. In der Regel handelt es sich dabei um kurze Dokumentenanalysen, echte Fallstudien umfangreicheren Stils zu den Vergleichsprogrammen wurden nicht durchgeführt. Die hier erbrachten Informationen reichen jeweils aus, um einen Vergleich auf der Ebene grundsätzlich verfolgter Strategien anzustellen, jedoch nicht, um die genaueren Wirklogiken der Vergleichsprogramme sowie die Rolle von Faktoren zu erfassen, die deren Erfolg in den jeweiligen Umwelten fördern oder behindern. Dazu hätte es in allen Fällen wesentlich umfangreicherer Daten und Analysen bedurft. In einem Fall wurden punktuelle Erkenntnisse aus unterschiedlichen nationalen Kontexten zu einem verallgemeinerten Bild zusammengezogen, das die Rolle von kontextuellen Randbedingungen für einen jeweiligen Programmerfolg zwar erwähnt, aber de facto vernachlässigt.

Blickt man quer über die 20 analysierten Berichte, so wurde im Großen und Ganzen in den Ansätzen zur Kontextanalyse nicht klar strukturiert und nur ansatzweise reflektiert, welche beobachteten Eigenschaften der evaluierten Programme in ihren realen Umwelten in welcher Weise zu besseren oder schlechteren - und in diesem Sinne dann verbesserungsfähigen - Entwicklungen hin auf intendierte Ziele beigetragen haben. Die vorliegenden Analysen machen in aller Regel vor einer umfassenden Einbeziehung relevanter anderer FTI- Maßnahmen und Steuerungen bald halt, und sie erreichen jedenfalls nicht die systemische Ebene, die sich übergreifend über alle Maßnahmen und Steuerungen ergibt (etwa im Bereich der Unternehmensinnovation auch in Form indirekter Förderung, im Wissenschaftsbereich in Form institutioneller Steuerung und Aushandlung). In jüngeren Evaluationen wurden solche Überschneidungslogiken zwar angetastet, aber nicht gut bewältigt. Zusammenfassend kann gesagt werden, dass ein essentieller Teil der Kausalitätsproblematik, welcher Stellenwert einzelnen Programmen im Zusammenhang aller Interventionen und Steuerungen – auch in Relation zu Strukturen und wirksamen Einflussfaktoren außerhalb ihrer jeweiligen unmittelbaren Reichweite – zukommt, in den bisher üblichen Herangehensweisen an Kontextanalysen nicht bewältigt werden konnte.

4.12 Beschreibung von Zwecken und Vorgehen

Die Erfüllung dieses Standards siedelt sich im Wesentlichen auf einem mittleren Niveau an, das von Negativtendenzen in der frühesten Beobachtungsperiode und zunehmenden positiven Tendenzen in der jüngeren Entwicklung begleitet ist. Von einer sehr guten Erfüllung dieses Standards, der für die Nutzung von Evaluationen durch die direkten AuftraggeberInnen und vor allem durch Außenstehende, die sich für die Evaluationsergebnisse interessieren könnten oder sollten oder die die Evaluation einzuschätzen versuchen, übergreifende Bedeutung hat, kann in keinem einzigen Fall gesprochen werden. Wie bereits dargestellt, hat eine durchwegs nicht optimale Erfüllung der in diesem Standard enthaltenen Hinweise und Empfehlungen auch die vorliegende Metaevaluation behindert.

Abbildung 25: Ergebnisse der Berichtsanalyse für den Standard G3



Alle Evaluationsberichte jüngeren Datums enthalten Methodendarstellungen. Über den gesamten Untersuchungszeitraum hinweg war dies keineswegs immer der Fall. Oft finden sich nur abrisshafte Angaben, die nicht ausreichen, um die Aussagekraft und Tragfähigkeit von erbrachten Ergebnissen einzuschätzen, z.B. wenn Rücklaufquoten aus Umfragen angegeben sind, Fragebögen aber nicht beigegeben wurden. Auch bei Programmevaluationen, denen einige Anhänge beigegeben sind, ist festzustellen, dass doch nicht alle eingesetzten Methoden gleichermaßen durch entsprechende Angaben einschätzbar gemacht werden.

Oft wurden Evaluationsberichte offenbar vor allem als Datenberichte verstanden, in denen in erster Linie den direkten AuftraggeberInnen, die über die Vorgehensweise der jeweiligen Evaluation bereits informiert waren, Datenergebnisse und -interpretationen bzw. Schlussfolgerungen vorgelegt wurden. Die Evaluationsprozesse mit ihren verschiedenen Entscheidungen, die die Vorgehensweise einer Evaluation prägen, wurden in aller Regel nicht als berichtenswert erachtet. Aber auch das Set der verfolgten Evaluationsfragestellungen, das im Rahmen der üblichen Ausschreibungs- und Vergabeverfahren jeder Programmevaluation vorgegeben ist, und das evaluierte Programm finden sich oft nur in Ansätzen oder Auszügen dargestellt. Eine umfassende Darstellung von allen wesentlichen konzeptiven Grundlagen und Entscheidungen einer Evaluation ist freilich erforderlich, um eine Evaluation insgesamt hinsichtlich verschiedener Gesichtspunkte, die in das holistische Qualitätsverständnis der Standards einfließen, gut einschätzen zu können. Die unvollständige Berichterstattung über die Gesamt-Vorgehensweise der Evaluationen hat nicht nur die Einschätzung verschiedener Aspekte der Evaluationen bzw. Evaluationsberichte in der vorliegende Metaevaluation behindert. Sie ist auch als entscheidendes Hindernis für eine spätere Nutzung der in den Evaluationsberichten enthaltenen Informationen zu erachten, da die Art der Fragestellungen, die hinter den präsentierten Daten standen, und konzeptuelle Status von Informationen (z.B. als Outcomes oder Outputs bestimmter Programmaktivitäten) oft nicht gut nachvollzogen werden kann.

Einen deutlichen Schwachpunkt bildet die Darstellung der organisatorischen Vorgehensweise der Evaluation, zu der nur wenige Berichte einige Hinweise enthalten. So wird in keinem einzigen Fall dargelegt, ob einzelne Schritte der Evaluation mit den AuftraggeberInnen abgestimmt wurden, wie z.B. im Methodeneinsatz bei der Auswahl von InterviewpartnerInnen oder von Cases bei einem Case Study-Ansatz, und wozu derartige Abstimmungen gegebenenfalls geführt haben. Interessieren würde hier z.B. auch, ob eingesetzte Fragebögen Pretests unterworfen wurden. Aussagen über etwaige Anpassung der weiteren Evaluationsdurchführung im Anschluss an einen Austausch über Zwischenergebnisse werden ebenso wenig gemacht. Interaktionen mit den Auftraggebern und gegebenenfalls mit den Stakeholdern, die aus Sicht der Standards wesentliche Komponenten des Vorgehens einer Programmevaluation sind, werden bestenfalls aus Methodenangaben erschließbar. Keiner der analysierten Evaluationsberichte gibt an, ob zur betreffenden Evaluation eine Steuerungsgruppe eingerichtet war, was in der Evaluationspraxis allerdings schon der Fall war. Es ist auch einige Variabilität dabei festzustellen, was als darstellungswürdige Methode aufgefasst wird und was nicht. So wird die Erstellung einer Logic Chart oft nicht als Methode geführt, was zugleich auch darauf hindeutet, dass sie weit eher als Präsentationsmittel denn als Analyseinstrument verstanden wurde. Wollte man einen Vergleich des Methodeneinsatzes in allen 20 von der Metaevaluation betrachteten Evaluationen anstellen, so bedürfte dies einiger Rekonstruktionsarbeit.

Trotz der in neueren Berichten immer besseren Methodendarstellungen wird die grundsätzlich geforderte konzeptiv-logische Bezugskette, die von Evaluationszwecken ausgehend einen Evalutionsansatz wählt, um von da zur Formulierung der Evaluationsfragen voranzuschreiten und schließlich das geeignete Evaluationsdesign (Datenerhebungs- und Analysemethoden) zu definieren, fast immer nicht transparent. Der Ausweis von Evaluationsfragestellungen erfolgt des Öfteren in einer Weise, als ob die im Bericht behandelten Fragestellungen die ursprünglich formulierten Fragestellungen wären. Damit wird nicht greifbar, welche eigentlich vorgesehenen Fragestellungen im Zuge der Evaluationsdurchführung wie gut verfolgt werden konnten, oder ob einzelne Fragestellungen angesichts verfügbarer Datenlagen, angesichts von Zwischenergebnissen, oder in Reaktion auf Programmveränderungen folgerichtig abgeändert wurden.

Von Seiten der AuftraggeberInnen wurde teilweise eine Sorge über die gute Einschätzbarkeit der in einer Programmevaluation durchgeführten Analyse deutlich. Das folgende Zitat bringt dies zum Ausdruck: *„Wie glaubhaft ist die Studie, wie glaubhaft sind die daraus gezogenen Empfehlungen, wie gut argumentiert ist das, wie belastbar ist es, wie gut kann ich damit arbeiten - wir wollen ja damit arbeiten, wir wollen Information daraus gewinnen, wir wollen etwas lernen für unser Tagesgeschäft." (A3)* Während andere AuftraggeberInnen ihre Herangehensweise an das Einschätzen der erhaltenen Evaluationsberichte weniger in dieser methodischen Form schilderten, kann doch davon ausgegangen werden, dass bereits für die Nutzung der Evaluationsberichte durch ihre unmittelbaren AuftraggeberInnen eine gute Gesamtdarstellung der Vorgehensweise einer Programmevaluation mit expliziten Hinweisen auf Tragfähigkeit und Aussagekraft der verschiedenen Analyseschritte und der insgesamt gegebenen Vor- und etwaigen Nachteile der Gesamtvorgehensweise einen wesentlichen Zugewinn für die Nützlichkeit der Berichte darstellen würde.

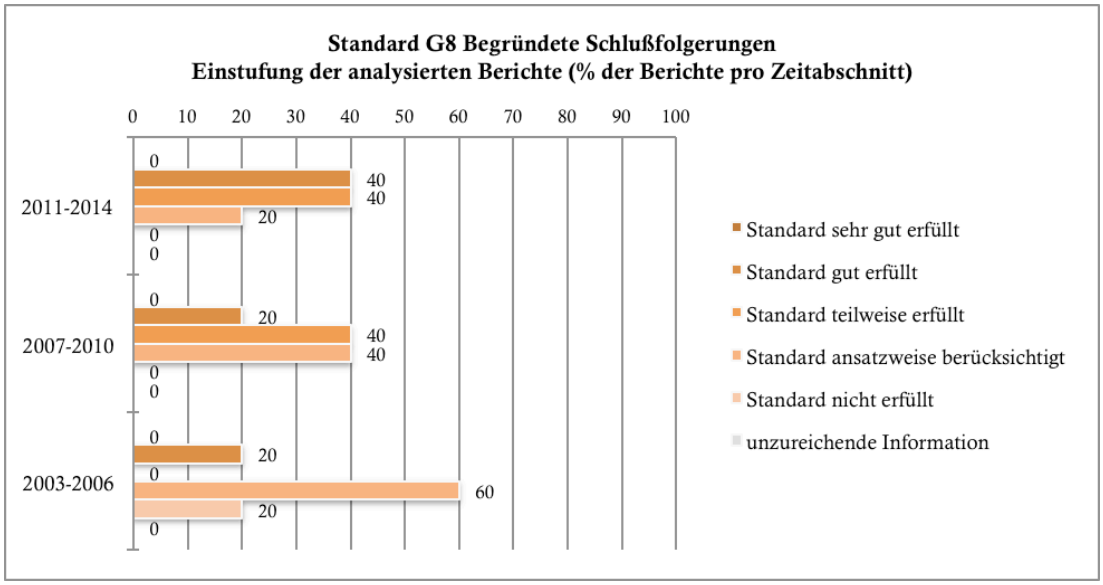
4.13 Begründete Schlussfolgerungen

Die Forderungen des Standards G8 nach einer guten Absicherung aller Schlussfolgerungen in ausgewiesenen Fakten und nachvollziehbaren Argumentationen wurde in den analysierten Evaluationsberichten nur bedingt erfüllt.[10] Es liegen in allen drei Untersuchungsabschnitten jeweils Berichte vor, bei denen von einer guten Erfüllung zu sprechen ist, wie Berichte, die den Standard nur ansatzweise erfüllt haben. Dass Schlussfolgerungen gar keinen Rückhalt in den aufgearbeiteten Fakten haben, kommt kein einziges Mal vor. Ebenso gibt es aber auch keinen Bericht im Sample, der den

[10] Die Schlussfolgerungen werden in der vorliegenden Metaevaluation nicht nach ihrem Inhalt oder dessen grundsätzlicher Plausibilität in einer Reflexion von FTI-Maßnahmen beurteilt, sondern gemäß der Textierung des Standards nach dem formalen Kriterium, ob die ausgesprochenen Schlussfolgerungen und Empfehlungen hinsichtlich der sie tragenden Daten, Interpretationen und Argumente transparent und unter Rückbezug auf die Berichterstattung über die verschiedenen Beobachtungen zum Evaluationsgegenstand gut nachvollziehbar sind.

Standard sehr gut erfüllt hätte, da doch immer wieder Unschärfen vorliegen und Freiheiten gegenüber dem Rahmen der – oft zur Ausfüllung der vorgesehenen Evaluationsschwerpunkte nicht vollkommen zureichenden – Daten genommen wurden (vgl. zur oft doch nur eingeschränkten Informationsgrundlage die Ausführungen zum Standard N4).

Abbildung 26: Ergebnisse der Berichtsanalyse für den Standard G8



Es sollte davon ausgegangen werden, dass gerade der Wohlbegründetheit der Schlussfolgerungen zentrale Bedeutung für die Gesamtqualität jeder Evaluation bzw. jedes Evaluationsberichts zukommt. So empfiehlt der eminente Evaluationsexperte D.L. Stufflebeam, dass bei der Anwendung der Standards zur Einschätzung der Evaluationsqualität eine Evaluation insgesamt als Misserfolg betrachtet werden sollte, sobald sie diesen Standard nicht erfüllt (Stufflebeam 1999). Im untersuchten Sample wären zwei Programmevaluationen älteren Datums davon betroffen.

Schlussfolgerungen und Empfehlungen finden sich in allen Berichten in Form eines von Ergebnispräsentationen getrennten Kapitels, das meist am Ende, manchmal auch Anfang des Berichtes steht. In einem Fall wird die Funktion eines Schlussfolgerungskapitels vom Executive Summary übernommen. Daneben finden sich in nahezu allen Berichten auch Schlussfolgerungen und Aussagen mit Empfehlungscharakter, die am Ende von Kapiteln zu Datenanalysen platziert oder im Fließtext eingeflochten sind.

Nur in seltenen Fällen werden Gesamt-Schlussfolgerungen so formuliert, dass sie direkt auf die der durchgeführten Analyse zugrunde liegenden Evaluationsfragen bezogen sind. Dies ist im Zusammenhang mit der allgemein geringen Neigung zu sehen, die Evaluationsfragen klar, übersichtlich und vollständig darzustellen. Hinsichtlich der Transparenz hin auf die aufgearbeiteten Fakten und zu deren Interpretation herangezogenen Argumentationen ist es nahezu der Regelfall, dass die Schlussfolgerungen nur einen teilweisen, aber nicht durchgängigen Bezug zu diesen Grundlagen unterhalten. Ein nahezu durchgängiges Merkmal ist darin zu erblicken, dass nochmals neue, bislang nicht berichtete Fakten herangezogen werden und auch Annahmen und Vermutungen der EvaluatorInnen zum Tragen kommen. Es kommen teilweise Plausibilitätsargumente zur Anwendung, zu denen die EvaluatorInnen nicht ausweisen, dass sie nicht direkt aus vorliegenden Daten erwachsen, und dass unter Umständen andere Konsequenzen aus der Datenlage ebenfalls möglich sein könnten. In einigen Fällen sind die Schlussfolgerungen ausschnitthaft, indem sie sich nur auf einzelne Programmbestandteile beziehen, ohne dass diese Selektivität explizit gemacht würde. Zu beobachten ist auch immer wieder, dass aus in Berichtsabschnitten identifizierten Faktenlagen auf der Ebene der Schlussfolgerungen keine Konsequenz gezogen wird, sodass Daten und Interpretationen, die für Reflexion und Lernen über das in realen Verhältnissen tatsächlich entfaltete Programm wesentlich sein könnten, im ausführlichen Berichtstext gleichsam versteckt bleiben. Nicht zuletzt wegen dieser Selektivität sind die Inhalte des ausgewiesenen Schlussfolgerungskapitels des Öfteren mit über den Bericht verteilten Schlussfolgerungen und Aussagen empfehlungsartigen Charakters nicht

ident. In mehreren Fällen werden Einzelbewertungen zu bestimmten Programmaspekten herausgestellt, während andere Programmaspekte mit den auf sie angewendeten, andersartigen Bewertungsmaßstäben in den Hintergrund rücken (vgl. die Ausführungen zum Standard N5). In Extremfällen kommt es in den verschiedenen Abschnitten zu gegenläufigen Aussagen, etwa über das Ausmaß der festgestellten Zielerreichung.

Die thematische Gestaltungsweise der Schlussfolgerungen mancher Programmvaluationen könnte damit zusammenhängen, dass Entscheidungen zur Zukunft des evaluierten Programms zum Zeitpunkt der Berichtslegung bereits getroffen waren und daher nur bestimmte Perspektiven für die AuftraggeberInnen von Interesse waren. Über derartige Umstände schweigen die Berichte jedoch. Relativen Freiheiten, die sich die Schlussfolgerungen vieler Programmevaluationen hinsichtlich der Transparenz gegenüber berichteten Datenlagen genommen haben, dürften auch im Zusammenhang damit stehen, dass EvaluatorInnen sich auch als IdeengeberInnen für die FTI-Politik verstehen und in dieser Rolle auch durchaus nachgefragt werden. Mit dem Blickwinkel der DeGEval- und JC-Standards auf die Qualität einer Programmevaluation ist dies allerdings schlecht zu vereinbaren. Das hier verankerte Verständnis von Rrogrammevaluation unterschiedet sie deutlich von einem Expertengutachten. Es mag schließlich auch ab und an eine Ressourcenproblematik mitgespielt haben, indem in den Evaluationsprozessen nicht in ausreichendem Umfang Ressourcen vorhanden waren bzw. von den EvaluatorInnen eingesetzt wurden, um die Gesamtbedeutung aller Faktenlagen und logisch-argumentativen Schritte und Zwischenschritte der Schlussfolgerungen gut herausarbeiten und darstellen zu lassen.

Aus der EvaluatorInnen-Umfrage ergibt sich der im vorliegenden Zusammenhang durchaus interessante Hinweis, dass das Ausmaß, indem sich eine Programmevaluation in ihrer Analyse auf Gesichtspunkte beschränkt, die durch gesicherte Daten abgedeckt waren, oder aber über diese Datenlage hinausging, kaum als Einflussfaktor darauf betrachtet wird, ob und inwiefern die Evaluationsergebnisse genutzt wurden (vgl. Kapitel 3).

Von Seiten der AuftraggeberInnen wird der Qualität der Schlussfolgerungen und Empfehlungen hohe Bedeutung beigemessen. In den Worten zweier Interviewpartner: *„'Woraus schließen Sie das' ist die evaluatorische Frage, die sich ein Evaluator gefallen lassen muss. Wenn das nicht zur Genüge beantwortet werden kann, ist die Studie wertlos." (A3) „Wir nehmen dann die Approbation nicht vor, wenn wir faktische Fehler sehen, woher haben sie das, das ist mit der Datenlage nicht begründbar. Das passiert eher selten." (M1)* Die Plausibilität der Schlussfolgerungen und Empfehlungen für die AuftraggeberInnen wird immer wieder zum Anlass, um einen Evaluationsbericht auf Detailebene zu hinterfragen.

5. Aktuelle Herausforderungen in der FTI-politischen Arena

In den Gesprächen mit den AuftraggeberInnen an den verschiedenen politisch-administrativen Systemstellen sind Strukturmerkmale der Einbettung der Evaluationsfunktion und Bedarfslagen sichtbar geworden, die für die Nützlichkeit der Evaluationspraxis und ortbares Verbesserungpotenzial von Bedeutung sind. In den im Folgenden dargestellten Gesprächsinhalten wird zudem ersichtlich, dass die Evaluationspraxis nicht als statisch begriffen werden kann, sondern als Bestandteil von dynamischen Systemen selbst Weiterentwicklungen und Bedarfsveränderungen unterliegt. Aus den detailreichen Gesichtspunkten lassen sich jedenfalls die folgenden essentiellen Leitmotive extrahieren.

Ressourcen und Kapazitäten

Früher gehegte Erwartungen an die Leistungskraft von Programmevaluationen wurden im Zuge der Durchführung verschiedener Evaluationsprojekte zunehmend als unrealistisch erkannt. Die verfügbare Ressourcenausstattung von Programmevaluationen wird als wesentlicher Mitgrund dafür erachtet, dass immer wieder Informationsbedürfnisse nur eingeschränkt befriedigt werden konnten. *„Meiner Meinung nach sind die Studien recht unterdotiert dafür was man eigentlich wissen möchte. Dann kann ich kaum übelnehmen wenn in dieser Breite dann die Ergebnisse nicht wirklich vorliegen." (M1) „Dass man [Programmevaluationen] macht, um es abzuhaken, ohne jemals vorgehabt zu haben, es zu verwenden, das passiert bei uns nicht. Aber es gibt einen gewissen Spielraum beim Umfang der Aufträge. Da muss man sich schon vor Augen halten, für gewisse Summen bekommt man nur soundso viel. " (M2)*

Es liegen von allen Seiten Hinweise vor, dass die in den Ministerien und Agenturen vorhandenen Personalkapazitäten für Evaluation zwar für die Führung der Vergabe-, Durchführungs- und Abnahmeprozesse in der bisher geübten Form grundsätzlich ausreichen, aber doch bereits in Bezug auf diese knapp sind. Hier zieht sich eine kleinteilige Struktur von der Evaluationsplanung bis zur

Verwertung der Evaluationsergebnisse durch. Stellvertretend für verschiedene erhaltene Aussagen kann hier die folgende stehen: *„Man nimmt sich natürlich die Zeit, die Ausschreibungen zu machen und versucht das ... aber im Endeffekt gehen sich die Dinge aus, die einem wichtig sind, und das ist einfach etwas, was nicht Unwesentlich ist. (...) Aber wie weit die Anderen dann Zeit und Interesse haben können - also das Interesse wäre schon da glaube ich." (M1)*

In den auftraggebenden Ressorts liegt die Weitergabe von Evaluationsergebnissen in den Händen der für die evaluierten Programme bzw. die Beauftragung und Abnahme der Evaluationen zuständigen BeamtInnen. Damit hängt es, wie sich in den Gesprächen zeigt, stark von deren Arbeitskapazität und persönlicher Initiative ab, inwieweit die aus einer Evaluation gewonnene Information - über etablierte Mindestanforderungen der Informationsweitergabe hinaus - auch an weitere Fachabteilungen im Haus verteilt wird, und im Idealfall auch Gegenstand eines gemeinsamen fachlichen Austauschs wird. Auch die Möglichkeiten für nicht unmittelbar mit dem evaluierten Programm befasste Akteure, sich der Information zuzuwenden, sind deutlich limitiert. *„Man hat die Möglichkeit sich sehr breit zu informieren, aber man könnte die eigene Arbeit gar nicht erledigen, wenn man das alles lesen würde." (M1) „Uns ist allen klar, dass viele Evaluationen kaum gelesen werden, oder nur von den Auftraggebern und Auftragnehmern, und dann vielleicht ein paar Spezialisten. Deshalb ist es empfehlenswert, zumindest die wichtigsten Ergebnisse auch in anderen Formaten bekannt zu geben." (A3) „Berichte lesen, das macht nur die Handvoll unmittelbar Betroffene im Programm selbst, der Auftraggeber." (A1)*

Es zeigt sich, dass gute Intentionen zur Evaluationsplanung, die auch im Sinne der DeGeval-Standards durchaus zu begrüßen sind, bald in pragmatischer Hinsicht an Grenzen stoßen. So sagt etwa ein/e GepsrächspartnerIn zu einer Abstimmung der Informationsbedürfnisse aller beteiligten Seiten in der Principal-Agent-Beziehung, die der geplanten Evaluation gute Chancen einräumt, für alle Seiten auch ein gutes Ergebnis zu erbringen, sagt ein Interviewpartnern: *„Da gibt es Verbesserungspotenzial, ja. Jetzt gerade bei der laufenden Evaluation wird abgestimmt, wird diskutiert, wird die Gelegenheit gegeben an verschiedenen Stellen mitzumachen. Da versuchen wir eng anzubinden und kommen auch drauf, dass zu viel Information immer wieder gar nicht so ankommt. Wir sind alle gut beschäftigt und überarbeitet. Aber wichtig ist, offen zu sein, und das haben wir gelernt. Da offen zu sein, bringt uns viel mehr bei der Umsetzung der Evaluationsergebnisse, als da nicht offen zu sein." (A3).*

Zugleich ergeben sich aus der Verankerung der Programmevaluationen in Programmvereinbarungen festgelegte Evaluationszeitpunkte und Budgets. Es wird beschrieben, dass hier eine gewisse Flexibilität in der Handhabung des Einzelfalls möglich ist, doch wird im Gesamtbild davon auszugehen sein, dass im Großen und Ganzen enge Grenzen gesetzt sind. *„Wir haben sehr, sehr strenge, restriktive Regeln, was Begleitung von Programmen betrifft. Da muss ich schon sehr gut argumentieren, warum das jetzt notwendig ist. Die Gruppe der Abteilungsleiter macht das mit dem Sektionsleiter, dann muss es durch die Budgetabteilung und nochmals mit dem Minister verhandelt werden." (M1)* Evaluationsprojekte oder Studien evaluativen Charakters, die nicht in Programmdokumenten vorprogrammiert waren, wurden nur in seltenen Ausnahmefällen initiiert. Den vorprogrammierten Programmevaluationen attestieren GesprächspartnerInnen immer wieder eine gewisse Korsettierung, wie es in den folgenden Zitaten zum Ausdruck kommt. *„Da muss man schon ein bisschen aufpassen, weil das Vergaberecht und der Vergabeprozess schon ein Korsett ist, das an dem Punkt manchmal Schwierigkeiten bereitet. Man ist nicht in einem intensiven Iterationsprozess vor der Auftragsvergabe." (A1) „Bei der Erstellung der ToRs ist Beschränkung eines der schwierigsten und wichtigsten Dinge - was klammere ich von vornherein aus, was zwar interessant wäre, aber den Evaluationsgegenstand explodieren lassen würde. Wir haben nicht riesige Budgets. Wie komme ich zu einer kompakten Beschreibung der Evaluation, die es ermöglicht, eine interessante Studie zu machen, ohne Alles beantworten zu wollen. [Es ist auch] Mut zur Lücke [notwendig]." (A3)*

Spannungen zwischen unterschiedlichen Evaluationszwecken

Wie bereits im Kapitel zu eingetretenem Nutzen beschrieben wurde, erzeugt die Verankerung der Programmevaluationen als Bestandteile der Programmvereinbarungen auch eine Spannung zwischen vorgegebenen Evaluationsfragestellungen und aktuellen Informationsbedürfnissen in einem hochdynamischen System, die in der Vergangenheit wiederholt auf Kosten aktuell relevanter Erkenntnisse gegangen ist. Von manchen InterviewpartnerInnen wird eine Orientierung der Terms of Reference an der ursprünglichen Programmformulierung in den Vordergrund gestellt, von anderen ein Abgleich mit aktuellen Informationsbedürfnissen, die sowohl in Ressort als auch Agentur in der Zwischenzeit entstanden sein können. Stellvertretend für diese Problematik einer Tarierung zwischen eher an Rechenschaft und ursprünglichen Programmformulierungen ausgerichteten Evaluations-planungen und aktuellen Bezugspunkten für die anzustellende Analyse mag die folgende Darstellung

stehen: *„Das ist nicht trivial, eine solche Evaluierungsausschreibung, obwohl man glauben könnte, es ist jetzt drei Jahre nach dem Programmstart, dass es eine Zwischenevaluierung gibt, aber auch da ist noch nicht so ganz klar, wie die Ausschreibung dann ausschauen wird, welche Fragestellungen da schlussendlich den Schwerpunkt bilden." (A2)*

Von manchen InterviewpartnerInnen wurde auch auf die Doppelfunktion von Evaluation hingewiesen, einerseits gesicherte Fakten und Dokumentationsleistungen, und andererseits Bewertungen zu erbringen. Hier wird in manchen Anwendungsfällen von Evaluation die Erbringung einer guten Faktendokumentation in den Vordergrund gestellt. In der folgenden Formulierung wird freilich klar erkennbar, dass es sich beim Versuch, mehrere Zwecke gelichzeitig zu verfolgen, um ein nicht-triviales Problem für die Art der Erkenntnisse handelt, die durch die Programmevaluation überhaupt hervorgebracht werden können. *„Für mich wäre das Interessanteste die Lernfunktion, auf die möchte ich mich konzentrieren. Das Problem, das ich habe, [ist:] wenn das den Geruch von Kontrolle bekommt, werden sich die Auskunftspersonen anders verhalten." (A2)*

Zirkulation von evaluativer Information im FTI-politischen Governance-System

In allen Ressorts und Agenturen wurden Zuständigkeiten und Kapazitäten geschaffen, um Evaluationen durchführen und Evaluationsergebnisse auf einer strategischen Ebene handhaben zu können. Die Planung und Durchführung der Programmevaluationen, die primär in den Ressorts erfolgt, ist dort an die Fachzuständigkeiten für die evaluierten Programme gekoppelt. Die Weitergabe von Evaluationsergebnissen innerhalb der Hierarchien stellt sich als geregelter Vorgang dar. Dabei wird davon ausgegangen, dass Evaluationsergebnisse nur eine Informationsquelle unter vielen sind, auf die sich politische EntscheidungsträgerInnen stützen, und dass auch die politische Aufmerksamkeit für unterschiedliche Programme deutlich variiert. Auch hier werden die Spielräume als schmal beschrieben: *„Man stößt einfach immer wieder auf Ressourcenknappheit in der Hierarchie." (M1)*

Eine Zirkulation von Evaluationsergebnissen hin zu anderen Fachabteilungen, die zur Stärkung der Wissensbasis in systemischer Hinsicht beiträgt, bemisst sich stark am Engagement von Einzelpersonen. In jüngster Zeit sind verstärkte Bemühungen zu beobachten, durch übergreifende hausinterne Präsentationen Evaluationsergebnisse in Umlauf zu setzen und Diskussionen zu initiieren, in denen auch nicht direkt mit dem evaluierten Programm befasste Abteilungen von den Evaluationsergebnissen profitieren können und strategische Einschätzungen vorgenommen werden können (vgl. Kap. 1). Eine institutionelle Verankerung derartiger wertvoller Vorgänge ist allerdings nicht gegeben, und eine durchgehende Systematik liegt nicht vor.

Im Rahmen der bestehenden institutionellen Architektur bestehen einige wenige Berührungspunkte zwischen den Steuerungssegmenten im FTI-politischen Bereich, in denen zumindest potenziell Informationen über geplante und fertiggestellte Evaluationen ausgetauscht werden können. In erster Linie sind es jedoch Personen und Netzwerke, die einen übergreifenden Wissensfluss im Governance-System gewährleisten, sodass sich ein solcher Wissensfluss letztlich als akzidentiell darstellt. In mehreren erhaltenen Aussagen werden diesbezügliche Veränderungen als erstrebenswert bezeichnet:

„Was es nicht gibt, ist ein institutionalisierter Intensivaustausch. Manchmal wäre ein intensiverer Austausch sinnvoll, der glaube ich einerseits an den Ansprüchen des Tagesgeschäfts scheitert, zweitens aber durchaus auch durch eine Konkurrenzsituation, die sich in den letzten Jahren zwischen [Agenturen] aufgebaut hat." (A1)

„Wieviel von Evaluierungsergebnissen anderer Abteilungen, anderer Ministerien profitieren wir? Nehmen wir das auf, um selbst hier in der Programmentwicklung [darauf Bezug zu nehmen]? Ich glaube, da haben wir einen Schwachpunkt in Österreich, dass da eigentlich die Kommunikation abhängig von den handelnden Personen und nicht systematisiert ist." (M1)

„Die fteval ist eine Einrichtung, die [Informationsweitergabe] schon macht, aber es zirkuliert meiner Meinung nach zu wenig auf die Ebene der Programmanagement-Verantwortlichen." (M1)

„Es gibt eine Reihe von Befindlichkeiten, die sich als Hemmschuh für Entwicklungen in Richtung eines großen Flusses an Informationen erweisen." (A1)

Auffächerung von Untersuchungszwecken und Erkenntnistypen

Es wird von allen Seiten Bedarf an verstärkt systemisch orientierten Erkenntnissen artikuliert, durch die die Positionierung einer Maßnahme im breiten Kontext verschiedener Förderungs- und Steuerungsinstrumente aufgezeigt werden kann. Dabei geht es gerade auch um das Erkennen von Optionen, wie Bedarfslagen im systemischen Gesamtzusammenhang durch Einsatz und Konfiguration bestimmter Instrumente und Maßnahmen gezielt und in bestmöglicher Weise begegnet

werden kann. Aus den Berichtsanalysen ergibt sich, dass bisherige Evaluationen solche Gesichtspunkte zwar nicht völlig vernachlässigt haben, aber doch mit meist unzureichenden Mitteln zu behandeln versucht haben, mit entsprechend mäßigem Erfolg. Heute steht bei allen interviewten Akteuren im Vordergrund des Bewusstseins, dass es künftig darum zu tun sein muss, einzelne Maßnahmen unter den Rahmenbedingungen zu verstehen, die diese umgeben bzw. das Operationsfeld der jeweiligen Maßnahme mit strukturieren.*„Das wäre ein sehr interessantes Evaluationsergebnis für mich, dass mir eine Evaluation indikative Informationen darüber gibt, wo meine Programm in dem Gesamtsystem positioniert sind, was sie für eine strukturelle Funktion im Gesamtsystem haben. Wo die Leute herkommen, die in meine Programme gehen, was ist der Differenzierungsfaktor dieser Programme bezogen auf andere Instrumente oder Maßnahmen, die es im Gesamtsystem gibt. So eine Aussage habe ich noch nie gefunden." (M2) „Wo wäre mein ideales Programm, wenn ich jetzt ein neues Programm aufsetzen würde? Ich müsste diesen Kontext erst einmal abbilden." (A2)*

In einem anderen Gespräch heißt es: *„Dass man mehr evergleichende Struktur- und Systemanalysen macht, das wäre wahrscheinlich besser als nur Programmanalysen. Zum Beispiel Positionierung des nationalen Innovationssystems, wie sieht das gegenüber anderen aus? Die Evaluatoren lesen sich [bei Evaluierungen einzelner Programme] die Richtlinien [der jeweils evaluierten Einzelprogramme] durch und stellen fest: vom Text her gibt es kaum Überschneidungen. Ob das dann von der Wirkung her sehr wohl eine Konsequenz für das Programm hat, das wird dann oft im Vorspann nicht mehr angeschaut, oft in der Evaluierung nicht wirklich angeschaut." (M1)*

„Wenn ich die Evaluierung nur über das Programms selbst habe, und den systemischen Kontext nicht dabei habe, oder nicht in der Detailliertheit, dann kann ich natürlich nur über das Programm entscheiden, aber nicht sagen, ob es insgesamt zur Veränderung des Systems beiträgt." (M1)

„Bei der heutigen Struktur der Portfolien ist die Programmsicht bei weitem in allen Fällen nicht die richtige Antwort. Was mir in der Tendenz abgeht, ist etwas was in der Logik zwischen Einzelevaluierung und Systemevaluierung liegt. Da sehe ich ein echtes Defizit in der Struktur." (A1)

„Wir wissen viel zu wenig. Wir wissen Einiges, (...) aber wir wissen zu wenig systematisch, und wir wissen auch nicht wie groß die Bandbreite dieser Programme ist. Meinem Verständnis nach war das bei diesen klassischen Evaluationen gar nicht so das Thema. Die haben halt Patente gezählt, Publikationen gezählt, weitere Kooperationen vielleicht berücksichtigt, aber ich kann mich an keine erinnern, wo die wirtschaftliche Wirkungsmacht auf der Agenda gestanden wäre." (M2)

Von einigen GesprächspartnerInnen wurde darauf hingewiesen, dass in einer übergreifenden Systemperspektive auch Bedarf besteht, einen Brückenschlag zwischen Evaluationen von Programmen und Evaluationen von Institutionen zu schaffen. Dies erscheint mittlerweile als eine nicht mehr übergehbare Notwendigkeit für künftige Definitionen der Einsatzpunkte von Steuerungen und Anreizbildungen, und erste Schritte in dieser Richtung wurden an betreffenden Systemstellen bereits in Gang gesetzt.

In einzelnen Systemsegmenten besteht zudem Bedarf an intensiveren Auseinandersetzungen mit Zielgruppen und der Wirkungsweise von Maßnahmen auf einer detaillierten Ebene, die in Richtung einer wissenschaftlichen Begleitforschung weisen. *„Die Mikromechanismen zu beobachten, wie schafft man solche Erfolge, das ist enorm kompliziert." (A3)* In diesem Zusammenhang werden auch die Anlagen und Ergebnisse einzelner Programmevaluationen als Bestandteile eines Bündels an Informationsquellen perspektiviert, die sich aus mehreren ineinandergreifenden Typen der Informationserbringung zusammensetzen und sich auch auf wissenschaftliche Studien über Innovationsbereiche erstrecken.

Eine zweite aus den Gesprächen erkennbare Bedarfslage ist diejenige an möglichst frühzeitigen Klärungen zur intendierten Wirklogik von konzipierten Programmen und zum Ausmaß, in dem die Erreichung von intendierten Zielen mithilfe der vorgesehenen Programmaktivitäten auch als realistisch erachtet werden kann. Dies weist eher in Richtung von Analysestrategien, wie sie in ex ante-Evaluationen zur Anwendung kommen.

Zum Dritten werden zumindest in einzelnen Steuerungsbereichen hoch reaktive und schlanken Studien evaluativen Charakters gewünscht, die das FTI-politische Handeln in stark dynamischen Umwelten zeitnah unterstützen. *„Die Fragestellungen, die wir haben, sind viel zu spezifisch und zeitkritisch. (...) Evaluationen sind Blicke in die Vergangenheit, wo man versucht Schlüsse zu finden, wie man das Programm für die Zukunft gestaltet. Da sind einerseits die Zeiträume zu kurz, wo man sagt, was ist denn jetzt schon vergangen, wie groß können die Zeiträume überhaupt werden, die man betrachtet, wenn die Programme ständig wechseln? (...) Die Zugangsweise, in der man fünf oder mehr Jahre anschaut und acht Calls, wird immer schwieriger. (...) Mittlerweile ist es so, dass wir ständig Strategieprozesse haben und ständig in die Zukunft gerichtet sind" (M2)*

Programmevaluation erscheint somit stärker als in früheren Phasen des Aufbaus der Evaluationskultur im FTI-Bereich im Spannungsfeld zwischen umfassenden (*comprehensive*) und gezielt zugeschnittenen (*tailored*) Evaluationen, die unterschiedliche und genauer definierte Evaluationszwecke verfolgen und mit differenzierten Schwerpunktsetzungen auch differenziertere Rollen für die EvaluatorInnen implizieren, als es in einem Pauschalzugang zu den bislang dominierenden Multi-Purpose-Evaluationen (vgl. Kapitel 4.1) der Fall ist.

6. Schlussfolgerungen und Empfehlungen

6.1. Schlussfolgerungen

Bereits aus den konzeptuellen Grundlagen der internationalen Evaluationsforschung und Evaluationsstandards, auf die sich die vorliegende Untersuchung stützt, war von Anfang an davon auszugehen, dass Nützlichkeit und Nutzung von Programmevaluationen von einiger Komplexität gekennzeichnet sein werden. Die Ergebnisse der vorgenommenen Untersuchung mit ihren drei Erhebungsinstrumenten der qualitativen Analyse eines Samples von Evaluationsberichten, einer Online-Umfrage unter EvaluatorInnen und semistrukturierter Interviews im Auftraggeberbereich bestätigt diese hohe Komplexität. Während zentrale Charakterzüge der Evaluationspraxis sichtbar werden, wie Programmevaluationen im FTI-Bereich bislang geplant, durchgeführt, präsentiert und genutzt wurden, zeigen sich doch nicht einige wenige Stärken, denen dann einige eindeutige Schwächen gegenüberstünden. Es handelt sich vielmehr um eine Verflechtung vielfältiger Gesichtspunkte, die bei jeder einzelnen Programmevaluation in bestimmter Weise schlagend werden und so zu einer Vielfalt von Evaluationsprodukten und Nutzungsprozessen mit individuellen Profilen führen. In den folgenden Ausführungen werden übergreifende Charakterzüge dieser prinzipiellen Vielfalt kondensiert. Bei der Zeichnung eines solchen übergreifenden Bildes kann zwangsläufig nicht jedem Evaluationsprozess und jedem Nutzungsprozess in jeder Hinsicht Genüge getan werden. Zugleich beziehen sich die Schlussfolgerungen ausschließlich auf Programmevaluationen im österreichischen FTI-Bereich und können nicht unmittelbar verallgemeinert werden, auch wenn sich FTI-Evaluation über die letzten beiden Jahrzehnte europaweit entfaltet hat und internationale Verständnisweisen der evaluativen Vorgehensweise im Politikbereich nicht von der Hand zu weisen sind.

Nutzung von Programmevaluationen

Bisherige Programmevaluationen haben durchaus Nutzen generiert. Aus den Auskünften der AuftraggeberInnen und HauptaddressatInnen der Programmevaluationen und denen der EvaluatorInnen ergibt sich hier ein hochgradig konsistentes Bild. Dabei stehen Nutzungsweisen im Vordergrund, die von der Evaluationsforschung als instrumenteller und konzeptueller Nutzen von Evaluation bezeichnet werden. Auf Basis von Datenlagen, Schlussfolgerungen und Empfehlungen wurden Entscheidungen über Programme getroffen und neue Sichtweisen gewonnen, die zur Nachschärfung von Programmen oder zur Bereinigung von Zielkatalogen geführt haben. Entscheidungen über evaluierte Programme betrafen vor allem Adjustierungen von weiterlaufenden Programmen. Derartige Entscheidungen können sowohl zu Umsetzungsaspekten in den betrauten Agenturen fallen, entsprechend ihres jeweiligen Pouvoirs zum evaluierten Programm, oder seitens der Programmeigentümer für eine nachfolgende Programmphase in Programmdokumenten niedergelegt werden. Hinzu kommen konzeptuelle Einsichten über Merkmale von FTI-Segmenten ebenso wie über das Management von Programmen, die häufig auch da eintraten, wo nicht für außenstehende BeobachterInnen leicht erkennbare Entscheidungen gefällt wurden.

Die Programmevaluationen haben immer wieder Lerneffekte erzeugt, in deren Gefolge Themen verankert und Annahmen über Funktionsweisen von Teilen des FTI-Systems und der darauf gerichteten Steuerungs- und Anreizsysteme verändert wurden. Auch Evaluationen von Programmen, die in der Folge nicht weitergeführt wurden, haben solche systemisch wertvollen Einsichten und Lerneffekte erbracht. Evaluative Erkenntnisse zu einzelnen Programmen haben sich auf Konzeption und Gestaltung thematisch benachbarter Programme ebenso ausgewirkt wie auf Gestaltungsweisen von anderen Programmen im Portfolio derselben Agentur.

Nicht zuletzt ist es in der Entwicklung der Evaluationskultur zu organisatorischen Anpassungen gekommen, die den Umgang mit Evaluationen und deren Ergebnissen unterstützen. Insbesondere wurde in einer Agentur rezent ein Managementprozess für den gezielten Umgang mit Evaluationsergebnissen eingeführt, und in einer weiteren Agentur während der Laufzeit der

vorliegenden Metaevaluation die Planung, Durchführung und Präsentation von Evaluationen im Rahmen eines übergreifenden Monitoring- & Evaluationssystems weiter professionalisiert. Organisatorische Anpassungen, die die Evaluationskapazität erhöhen, liegen auch an anderen Systemstellen vor, in unterschiedlichem Ausmaß und insgesamt nicht auf einem gleichen Niveau.

In allen Gesprächen mit AuftraggeberInnen und in der EvaluatorInnen-Befragung wurde ersichtlich, dass die über die Jahre durchgeführten Programmevaluationen als wesentliche Beiträge zu einer Verbreiterung und Vertiefung der Wissensbasis eingeschätzt werden, auf die sich FTI-politisches Handeln gerade auch als aktualitätsbezogenes und voranschreitendes Handeln in dynamischen Umwelten stützt. Zugleich wird auch deutlich, dass es sich beim Eintreten von Nutzen aus Programmevaluationen um Gemengelagen handelt, sodass nicht nur eine Evaluation zu einer Nutzung führt, sondern multiple Effekte auftreten. Freilich handelt es sich bei Umsetzungen von Erkenntnissen aus Programmevaluationen nicht um Automatismen, sondern um Handlungsweisen in Multiakteurs-Konstellationen im Einzelfall, bei denen auch immer wieder Reibungsverluste auftreten. Evaluationsnutzungen sind deutlich von den Konfigurationen der Principal-Agent-Beziehungen geprägt, die sich für die parallel agierenden Segmente des politisch-administrativen Handelns im FTI-Bereich unterschiedlich darstellen. Ob und wie Evaluationsergebnisse in einem dieser Steuerungsbereiche auch über den Kreis der unmittelbar mit einem Programm befassten Personen breiter bekannt gemacht und aufgegriffen werden, erweist sich als unsystematisch und stark vom Engagement von Einzelpersonen abhängig. Zusätzlich erhöht wird die Komplexität der Nutzenentstehung im Überstieg zwischen der administrativen und der politischen Sphäre, wobei auch hier von einer beträchtlichen Variation von Einzelfall zu Einzelfall auszugehen ist.

Als deutlich weniger ausgeprägt erweist sich eine Nutzung von Evaluationsergebnissen, die über die Grenzen der jeweiligen Steuerungsbereiche mit ihren Principal-Agent-Verhältnissen hinaus reicht. Obwohl auch hier relevante Wissenszuwächse beschrieben werden und in den Steuerungsbereichen jeweils davon ausgegangen wird, dass interessierende Information aus anderweitig durchgeführten Evaluationen zumindest prinzipiell zugänglich ist, gibt es keinen systematischen Vorgang im FTI-politischen Governance-System, der das Aufgreifen von in anderen Zuständigkeitsbereichen erbrachten Evaluationsergebnissen und die Auseinandersetzung damit unterstützt. Ausstrahlungswirkungen von evaluativer Information auf interessierte Fachöffentlichkeiten bzw. auf Akteursgruppen in der FTI-Landschaft treten in noch geringerem Maß auf und müssen als volatil gelten, da sie abgesehen von der grundsätzlichen Verfügbarmachung derjenigen Evaluationen, zu denen die Berichte publiziert werden, in aller Regel nicht Gegenstand gezielter Vorgehensweisen sind.

Im FTI-politischen System dienen Programmevaluationen auch durchaus dazu, andere Akteure in der politischen Sphäre vom evaluierten Programm zu überzeugen oder Entscheidungen über Programme zu rechtfertigen („symbolischer Nutzen“). Hier geht es um eine Überzeugungsarbeit, die in der Multiakteurs-Arena eines von differenzierten Principal-Agent-Verhältnissen, Hierarchien und Kleinteiligkeit geprägten Systems stets zu leisten ist, wenn es um die Zukunft von FTI-Programmen bzw. Steuerunginstrumentarien geht. Zum Phänomenkomplex der Erzeugung von Evaluationsnutzen zählt freilich auch die Art der Verankerung der Evaluationsfunktion im rechtlich-institutionellen Rahmen der Bundesverwaltung. An der Schnittstelle zwischen Fachressorts und dem Bundesministerium für Finanzen (BMF) kommt den Programmevaluationen eine Funktion der Legitimation der Mittelausgaben zu. Diese Legitimationsfunktion ist im Motivbündel für die Planung und Durchführung von Programmevaluationen stets anwesend. Die Daten zeigen, dass eine symbolische Nutzung von Programmevaluationen andere Nutzenformen keineswegs ausschließt. Allerdings sorgt die gleichzeitige Anwesenheit von Lern- und Legitimationsfunktion für eine innere Spannung in jedem Evaluationsprojekt, die sich letztlich für eine Evaluationspraxis, die sich an der Erzeugung systematischer Wissenszuwächse im Governance-System orientiert, eher abträglich erweist.

Die Ergebnisse der Programmevaluationen können in ihrer Rolle als FTI-politische Informationsmittel als konkurrenzlos gelten, wenn sie auch oftmals nicht die alleinigen Grundlagen FTI-politischer Entscheidungen über Einsätze und Mittelzuweisungen sind. Im Verhältnis zu dieser Rolle von Evaluationsergebnissen ist das grundsätzliche Potenzial, bereits während der Evaluations-durchführung und unabhängig von den Evaluationsergebnissen aus der Durchführung von Programmevaluationen unmittelbar zu profitieren („Prozessnutzen“), bislang nur wenig ausgeschöpft worden.

Einflussfaktoren auf die Nutzung von Programmevaluationen

Faktoren, die in der bisherigen Evaluationspraxis Einfluss darauf gehabt haben, ob und wie Evaluationsergebnisse auch genutzt wurden, siedeln sich sowohl im Bereich dessen an, was innerhalb eines Evaluationsprojekts durch dessen Ausgestaltung beeinflusst werden kann, als auch im Bereich des umgebenden Kontexts, der außerhalb dieses Einflussbereichs verbleibt. Anhand von Daten aus der EvaluatorInnen-Befragung können die 20 wesentlichsten Einflussfaktoren bestimmt und gereiht werden. Sie finden in Auskünften von AuftraggeberInnen ihre Entsprechungen, wobei naturgemäß auch Perspektivunterschiede existieren sind und die EvaluatorInnen auch Faktoren bewertet haben, über die AuftraggeberInnen so nicht gesprochen haben. Diese 20 wesentlichsten Einflussfaktoren verteilen sich zu gleichen Teilen auf intrinsische Evaluationsmerkmale und auf Kontextfaktoren.

Unter jenen Faktoren, die sich im Verantwortungsbereich einer einzelnen Programmevaluation ansiedeln, steht die Glaubwürdigkeit der EvaluatorInnen bei den AuftraggeberInnen an erster Stelle. Diese Glaubwürdigkeit wird im österreichischen FTI-politischen Bereich durch die Heranziehung von auf FTI-Evaluation spezialisierten Instituten im In- und Ausland sowie durch die kontinuierliche Kommunikation von österreichischen FTI-EvaluationsspezialistInnen mit den Auftrggeber-Insitutionen in der Plattform fteval hergestellt. Ähnlich wichtig ist die Klarheit der Berichterstattung (Klarheit der Berichtsaussagen, Vorhandensein eines Executive Summary und dessen Aussagekraft), die von den FTI-EvaluatorInnen ebenso wie von deren AuftraggeberInnen als zentral erachtet wird.

Unter den Kontextfaktoren rangiert die Erwartung der AuftraggeberInnen, dass ihnen die konkrete Programmevaluation von Nutzen sein wird, an erster Stelle. Die bereits genannte Kombination von Lern- und Legitimationszwecken in der institutionell-rechtlichen Verankerung der Programmevaluationen kann als ein wesentlicher Grund dafür angesehen werden, dass deutliche Unterschiede im Umgang mit verschiedenen Programmevaluationen zu bemerken sind und immer wieder Fälle eingetreten sind, in denen Programmevaluationen von vornherein von ihren AuftraggeberInnen als notwendige Übungen betrachtet und kaum mit Nutzenerwartungen verbunden wurden, was sich dann von der Evaluationsplanung weg bis hin zum Umgang mit den Ergebnissen niederschlägt. Ebenfalls bedeutend für Art und Ausmaß der Nutzung von Evaluationsergebnissen ist der Umstand, ob eine Programmevaluation in direktem Konnex mit einem aktuellen FTI-politischen Entscheidungsbedarf oder Problemdruck steht. Ein derartiger Konnex besteht primär durch einen vorgegebenen Evaluations- und Verhandlungsrhythmus für Programmvereinbarungen der Ressorts mit dem Finanzministerium (BMF), während aktuelle FTI-politische Themenstellungen oder .„Windows of Opportunity“ kaum als solche zum Auslöser von direkt auf sie gemünzten Evaluationsaktivitäten werden. Aktuelle Informationsbedürfnisse der Ressorts und Agenturen werden innerhalb dieses Rahmens des Öfteren nur in eingeschränkter Weise befriedigt.

Eine starke Personenabhängigkeit der genaueren Umgangsweise mit einzelnen Programmevaluationen und ihren Ergebnissen tritt in der EvaluatorInnen-Umfrage mit vier Faktoren massiv zutage. Bei diesem „Human Factor“ in der Evaluationsnutzung geht es um die persönlichen Denkstile der jeweils Evaluationzuständigen, um die Konsistenz der Evaluationsergebnisse mit ihren Sichtweisen und Erwartungen, um ihre Erfahrenheit mit Evaluation, und um ihre Rolle in der jeweiligen Institution. Des Weiteren kommt organisatorischen Anpassungen, Ressourcen und Erfahrungen der auftraggebenden Institutionen ein erheblicher Stellenwert zu. AuftraggeberInnen haben hierauf mindestens ebenso stark hingewiesen wie die EvaluatorInnen, für die diese Kontextfaktoren mit zu den einflussreichsten zählen.

Einige hoch relevante evaluationsmethodische Gesichtspunkte wie die Angemessenheit der Evaluationskriterien, eine ausgewogene Darstellung von Stärken und Schwächen des untersuchten Programms oder die Art des Evaluationsansatzes sind in den 20 wesentlichsten Einflussfaktoren auf eine Nutzenentstehung aus der Sicht der EvaluatorInnen enthalten. Sie fallen jedoch im Gesamtbild hinter einige stärkere Einflussfaktoren merklich zurück, die durch die Vorgehensweise einer Evaluation nicht beeinflusst werden können. Methodenaspekte im engeren Sinn, wie die Anwendung eines Methodenmix, Triangulation oder die Finesse, mit der bestimmte Methoden eingesetzt werden, kommen unter den 20 wesentlichsten Einflussfaktoren auf Evaluationsnutzung, so wie die EvaluatorInnen sie einschätzen, nicht vor. AuftraggeberInnen sind auf evaluationsmethodische Aspekte nicht in einer vergleichbaren Detailliertheit eingegangen, haben aber doch gelegentlich auf Mängel hingewiesen, die in der Vergangenheit die Entstehung von Nutzen aus Programmevaluationen beeinträchtigt haben und in den Einzugsbereich der Methodenanwendung fallen. Insgesamt erhärtet sich das Bild, dass die traditionell vor allem in Methodendiskussionen

verankerte FTI-Evaluation die tatsächliche Entstehung von Nutzen aus durchgeführten Evaluationen nur in untergeordneter Weise diesem Hauptfokus ihrer Thematisierung der evaluatorischen Vorgehensweisen verdankt.

Nützlichkeit der Evaluationsberichte und Evaluationsprozesse

Die analysierten Evaluationsberichte entsprechen den herangezogenen DeGEval-Standards auf einem im Großen und Ganzen mittleren Niveau, und mit voranschreitender zeitlicher Entwicklung zunehmend besser. Verbesserungspotenzial ist dennoch vorhanden, wenn es um bestmögliche Programmevaluationen geht, die hohe Nützlichkeit erzielen und das im "Unternehmen Programmevaluation" angelegte Potenzial bestmöglich ausschöpfen. Eine sehr gute Erfüllung eines der 11 herangezogenen Standards konnte nur in einigen wenigen Fällen attestiert werden. Ebenso selten ist zugleich eine völlige Nichterfüllung eines der Standards, die auch in den letzten Jahren nicht mehr auftritt. Während zu allen herangezogenen Standards grundsätzlich noch Verbesserungspotenzial besteht, erscheinen die folgenden Gesichtspunkte als die relevantesten, um künftig noch nützlichere Programmevaluationen zu erzielen.

Die analysierten Programmevaluationen waren mit Ausnahme einer ex post-Evaluation Zwischenevaluationen oder Teile von Begleitevaluationen. Sie waren in den meisten Fällen sehr breit angelegt, Fragen von der Relevanz der Programme über ihre Effektivität bis hin zu ihrer Wirkung sollten verfolgt werden (sogenannte Multi-Purpose Evaluationen). Es wurden Outputs, Outcomes, und erste Wirkungen der Programme untersucht, sodass Erkenntnisse über die Programme durchaus erzielt wurden. Die Beobachtbarkeit von Programmwirkungen war auf Grund der früh gewählten Evaluationszeitpunkte fast immer deutlich eingeschränkt. Jedoch ist auch hinsichtlich dessen, was zu den Evaluationszeitpunkten bereits grundsätzlich zu den Programmen beobachtbar war, festzustellen, dass in vielen Fällen nicht von einer umfassenden und gründlichen Aufarbeitung der Programme gesprochen werden kann.

Die umfangreichen Evaluationsvorhaben wurden anhand von erhältlichen Monitoringdaten und weiteren, innerhalb der einzelnen Programmevaluationen jeweils selbst erhobenen Daten durchgeführt, die allerdings des Öfteren doch keine analytisch konsequente Ausleuchtung aller Programmkomponenten zuließen. Die Gesamtebenen aller relevanten Programmoutputs und -outcomes, die schrittweise hin zur Erreichung der Programmziele führen sollen, und insbesondere die Verbindungen zwischen diesen Ebenen, wurden nur mit teils deutlichen Einschränkungen greifbar gemacht (Standard N4). Aufgrund dieser Ausschnitthaftigkeit haben die meisten der untersuchten Programmevaluationen letztlich doch den Charakter einer sogenannten „black box"-Evaluation, durch die die genaue Art und Weise, wie ein Programm die intendierten Wirkungen erzielt bzw. an der Erzielung dieser Wirkungen gehindert ist, nicht oder zumindest nicht vollständig erfasst wird. Es zeigen sich des Öfteren Schwierigkeiten mit einer konzisen Gliederung von Programmkomponenten und Umsetzungsschritten zu Zielen unterschiedlicher logisch-hierarchischer Stellung (unmittelbare, intermediäre und übergeordnete Programmziele) und hinsichtlich der Art der Erreichung von direkten und indirekten Zielgruppen. Zugleich haben einige Evaluationen auch Fragestellungen behandelt, die nicht als zentrale Gesichtspunkte für ein tieferes Verständnis des evaluierten Programms zu erachten sind. Wie die EvaluatorInnen angeben, waren Auswahl und Umfang der in den Programmevaluationen herangezogenen Informationen häufig nicht ausreichend, um alle mitgegebenen Evaluationsfragen gut behandeln zu können, und noch weniger, um auch unbeabsichtigte Wirkungen der Programm erfassen zu können. Einige Evaluationsberichte tragen Züge eines „evaluability assessment", in dem die Bedingungen für eine zielführende Evaluation des Programms erst geklärt werden.

Fast alle analysierten Programmevaluationen haben sich auch mit dem Kontext der evaluierten Programme auseinandergesetzt, in unterschiedlicher Intensität und mit unterschiedlichen Perspektivierungen. Vor allem auf der Basis von qualitativen Untersuchungsstrategien wurden von manchen Evaluationen essentielle Randbedingungen greifbar gemacht, unter denen das jeweilige Programm in seinen Zielgruppen Wirkungen erreichen konnte bzw. daran gehindert war. Etliche Kontextanalysen leiden jedoch darunter, dass zwar einige Faktoren untersucht und für skizzenhafte Bilder fruchtbar gemacht wurden, aber der systematische Stellenwert dieser untersuchten Faktoren unklar bleibt bzw. keinen expliziten Bezug zu einer strukturierten und gesamthaft verstandenen Wirklogik des jeweiligen Programms aufweist (G2).

Die Evaluationsberichte geben trotz regelmäßig enthaltener Methodenbeschreibungen in der Mehrzahl doch nur unzureichend Auskunft darüber, was warum untersucht wurde, und als wie vollständig und tragfähig die erbrachten Ergebnisse eingeschätzt werden können. Im Verein mit nur sehr breiten und allgemein gehaltenen Angaben über die Untersuchungsschwerpunkte (Standard N2) und einer bemerkenswerten Enthaltsamkeit bei der Angabe von Evaluationsfragestellungen, die den jeweiligen Programmevaluationen zugrunde gelegt waren, ergibt sich so eine nur eingeschränkte Transparenz der Evaluationsergebnisse (Standard G3) und der Schlussfolgerungen, die aus ihnen gezogen wurden (Standard G8). Eine Transparenz der Vorgehensweise erscheint jedoch vor allem von Bedeutung, damit Evaluationsergebnisse auch von Akteuren aufgegriffen werden können, die nicht zum engen Kreis derjenigen Wenigen zählen, die unmittelbar mit der Konzeption und Umsetzung des untersuchten Programms und der dazu durchgeführten Evaluation befasst sind.

Ist anhand der Evaluationsberichte wegen ihrer Gestaltungsweise die Frage oft nicht gut beantwortbar, wie essentiell die erbrachten Ergebnisse im Hinblick auf die Gesamtlogiken der evaluierten Programme jeweils tatsächlich sind (N4, G3), so erscheint ebenso die Frage virulent, wie Programmen insgesamt Wert zugemessen wurde (N5). Hier offenbart sich ein „blinder Fleck" eines stark datenorientierten und zugleich oftmals eher unsystematischen Zugangs. Während manche Evaluationen nachvollziehbare Bewertungsmaßstäbe in konsequenter Weise in Anschlag gebracht haben, die in einer klaren Verbindung zu den Programmzielen standen, haben andere eher für sich stehende Einzelbewertungen zu einzelnen Beobachtungen vorgenommen, ohne dass in der Kombination von „üblichen" Betrachtungsweisen ein stringentes Gesamtkonzept greifbar würde. Es wird in der internationalen Evaluationstheorie allerdings davon ausgegangen, dass die Wahl der Bewertungsmaßstäbe ebenso eine tragende Säule jedes Evaluationskonzepts darstellt wie ihre Wissenschaftlichkeit und ihre gezielte Auseinandersetzung mit dem intendierten Nutzen.

Im Zusammenhang mit eingeschränkten Datenlagen waren die EvaluatorInnen immer wieder bestrebt, Lücken durch ihr Hintergrundwissen über das FTI-System und Annahmen über dessen Funktionsweisen oder Eigenschaften von Akteursgruppen wett zu machen (F3, G8). Dies beeinflusste oft merklich den Charakter von Schlussfolgerungen und Empfehlungen, die in unterschiedlicher Weise, aber doch teils recht deutlich, einen Zug von ExpertInnengutachten tragen, in denen das persönliche Wissen der AutorInnen zur Geltung gebracht wird. Dies deckt sich nicht mit dem Grundansatz der Evaluationsstandards, dass alle Aussagen einer Programmevaluation in transparenter Weise in von ihr herangezogenen Fakten und Quellen abgestützt sein sollten.

Evaluationsplanung

Festzustellen ist, dass in den Evaluationsprozessen zahlreiche Schritte, die den Standards zufolge vor allem im Planungsstadium einer Evaluation erfolgen können bzw. sollten, bislang nur ansatzweise wahrgenommen wurden. Hier zeigen sich unter Rückgriff auf Ergebnisse der EvaluatorInnen-Befragung unter anderem deutliche Verbesserungsmöglichkeiten bei der gezielten Auseinandersetzung damit, wie das Evaluationsprojekt auf eine konkret intendierte Nutzung zugeht (N8), und wie es entsprechend LernpartnerInnen einbindet (N1). Ebenso geht es aber um die Konfiguration von konzeptiv geschlossenen Studien, die nicht streckenweise letztlich ergebnisarme Unternehmungen bleiben, da sich im Verlauf der Durchführung herausstellt, dass Daten zur Beantwortung von Fragestellungen doch nicht ausreichend waren oder geplante Auswertungen so doch nicht durchgeführt werden konnten. Innerhalb von kurzen Vorbereitungsphasen der Evaluationen (Beantwortung von Terms of Reference und Hearing) kam es nur eingeschränkt zu einer Mitsprache der EvaluatorInnen, im Rahmen derer sie auf Basis ihrer Kompetenzen die Herangehensweise der Evaluation beeinflussen und schärfen konnten (N4). Wie die prominente Evaluationsforscherin C.Weiss feststellte, ist es ein übergreifender Charakterzug von Ausschreibungsverfahren („request for proposal"), dass sie vor allem fair sind, aber auch die Spielräume für Konzeptreflexionen einschränken, nicht zuletzt da auch der Aufwand der EvaluatorInnen für ihre Anträge in Grenzen des wirtschaftlich Verträglichen gehalten werden muss (Weiss 1998, S. 36f). Dahingehend unterscheidet sich die österreichische FTI-Evaluationspraxis nicht von internationalen Phänomenen in verschiedensten Politikbereichen.

Es hat sich allerdings in manchen Einsatzbereichen von Evaluation ein Modell entwickelt, mit dem nachteiligen Eigenschaften von Ausschreibungsverfahren durch eine spezifische Strukturierung der Evaluationsaufträge gegensteurt wird. Vor allem in internationalen Organisationen und im Politikbereich der Entwicklungszusammenarbeit hat sich eine sogenannte Inzeptionsphase (*inception phase*) etabliert, die einen Rahmen für eingehendere Konzeptualisierungsschritte am Beginn eines beauftragten Evaluationsprojekts einräumt.

Infobox: Eingangsphase einer Programmevaluation (Inception Phase)

UNODC - United Nations Office on Drugs and Crime, Evaluation Handbook, Chapter IV C. Inception Report (Auszug)

An Inception Report summarizes the review of documentation ("desk review") undertaken by an evaluator mandated by UNODC and specifies the evaluation methodology determining thereby the exact focus and scope of the exercise, including the evaluation questions, the sampling strategy and the data collection instruments. Consequently, the evaluator is expected to deliver an Inception Report as one of the key deliverables, which is shared with the Project Manager and the Independent Evaluation Unit for comments.

For Independent Project Evaluations, Project Managers check the quality of the Inception Report, provide extensive feedback and guidance to the evaluation team and finalize it. (...)

The Inception Report provides an opportunity to elaborate on the evaluation methodology proposed in the ToR and its related issues at an early stage of the evaluation exercise. It also ensures that evaluation stakeholders have a common understanding of how the evaluation will be conducted.

The evaluation team develops an Inception Report which contains the methodology used to answer the evaluation questions based on information derived from the ToR, the desk review and the evaluation team briefing. (...)

The Inception Report must explicitly and clearly state the limitations to the overall evaluation and to the chosen evaluation methods. A frequently encountered limitation is the lack of data (baseline and monitoring data) to address the evaluation questions. Alternative solutions have therefore to be found by the evaluation team to reconstruct the baseline data.

*

DANIDA Evaluation Guidelines, Ministry of Foreign Affairs of Denmark (Auszug)

INCEPTION: PLANNING THE EVALUATION

The purpose of the inception phase is for the evaluation team to prepare a detailed operational plan, i.e. the inception report, for the next phases of the evaluation: fieldwork and reporting.

Proper planning is essential to identifying those activities required to provide well-supported answers to the evaluation questions and to avoiding other unnecessary activities and related expenditures of time, effort and money.

The planning phase provides the evaluation team with the opportunity, and responsibility, to discuss methodological specificities, fieldwork activities and reporting strategy with, and where required obtain approval from, the Evaluation Department, and as well to consult with other stakeholders.

The inception report should present:

- An overall logic model of the intervention (the evaluand), depicting the linkages between resources (inputs), intervention activities (processes), intervention results (outputs or deliverables), intended outcomes (intervention objectives), overall impacts, and their relationships in terms of the criteria of relevance, efficiency, effectiveness and impact; an explanation of how the sustainability criterion is defined and operationalised.

- The methodology: design, approach, sufficiency and appropriateness of evidence, data collection strategy and methods, analytical framework and reporting outline.

- The hierarchy of evaluation questions starting from the general ones that are presented in the Terms of References through to the specific ones that will produce data and information.

- For each specific question the basis for assessment, i.e. indicator of minimum acceptable performance.

- A matrix indicating for each specific question the nature and source of evidence.

- A schedule of activities.

- A communication and consultation plan (with stakeholders).

Geschlossenheit des Feldes

Während sich v.a. in Europa die Praxis eingespielt hat, dass FachexpertInnen für bestimmte Themengebiete aufgrund dieser Spezialisierung als GutachterInnen und EvaluatorInnen eingesetzt werden (vgl. z.B. Widmer/de Rocchi 2012), steht in der US-amerikanischen Herangehensweise an Evaluation die Evaluations-Expertise, unabhängig von einer thematischen Spezialisierung auf ein Politikfeld, im Vordergrund. Die Evaluationsexpertise wird hier als Expertise sui generis betrachtet, die es ermöglicht, eine möglichst umsichtig konzipierte, methodisch sauber durchgeführte und möglichst nützliche Evaluation zu gestalten. Die Evaluationsexpertise geht gemäß dieser Betrachtungsweise weit über das Erheben und Analysieren von Daten hinaus und erfasst alle Aspekte, die in den Standards zum Ausdruck kommen, mit den dahinter stehenden Fachdiskussionen der Evaluationsforschung. Dementsprechend favorisiert die Evaluationspraxis US-amerikanischer Bauart, hohe Evaluationsexpertise in den Vordergrund zu stellen und diese allenfalls bedarfsgerecht mit themenspezifischer Fachexpertise zu kombinieren, je nach Bedarf der einzelnen Evaluation mit ihrer Ansiedlung in einem spezifischen Politikfeld (vgl. z.B. Weiss 1998, Stame 2013).

Ein der bisherigen FTI-Evaluationspraxis inhärentes Risiko betrifft augenscheinlich auch einen Lock-In in eingespielten Herangehensweisen. Wie die Analyse zeigt, wurden über weite Strecken Methoden zum Einsatz gebracht, die von den AuftraggeberInnen erwünscht waren oder von den EvaluatorInnen bzw. ihren Instituten regelmäßig eingesetzt werden. Lediglich in zwei der untersuchten Berichte wurden ungewöhnliche und innovative Methoden eingesetzt, die für die spezifische Aufgabenstellung der betreffenden Programmevaluation als produktiv erachtet und auch beauftragt wurden. Quasi-Kontrollgruppendesigns zur Auseinandersetzung mit Zielerreichungen sind selten, pre-post-Vergleiche kommen in den analysierten Evaluationen nicht vor. Da für einen Einsatz ersterer Herangehensweise ein Verstreichen längerer Zeiträume erforderlich ist, und für die zweitere Analysestrategie eine Evaluationsplanung und –beauftragung bereits vor dem Programmstart, können die eingespielten Evaluationsrhythmen als wesentliche Verantwortungsdresse für das Ausbleiben dieser Untersuchungskonzepte gelten. Die EvaluatorInnen geben in der Umfrage aber auch kaum wissenschaftliche Bezugspunkte an, die den durchgeführten Programmevaluationen zugrunde gelegen hätten (88% machen keine bzw. keine inhaltlich relevanten Angaben). Monitoringdaten bilden das von den Programmen ausgelöste Geschehen (Fördervergaben, Eigenschaften der Fördernehmer, Outputs, etc.) entsprechend früher Stadien der Programmkonzeption und -entwicklung ab und implizieren so auch eine potenzielle „Gefangennahme“ später möglicher Sichtweisen, sofern nicht durch ergänzende Erhebungen ausreichend gegensteuert wird, nach Maßgabe des jeweiligen Einzelfalls. Bewertungsmaßstäbe zur Einschätzung der Programme und Kriterien zur Einordnung von Beobachtungen wurden nur selten zwischen EvaluatorInnen und AuftraggeberInnen vorab gemeinsam geklärt (N5). Die Wahl von Bewertungsmaßstäben wurde oft den EvaluatorInnen überantwortet, und diese zogen entweder Maßstäbe heran, die in ihren Augen denen der AuftraggeberInnen entsprachen, oder verhielten sich unabhängig von solchen Annahmen.

So geht es auch um dominante Perspektivierungsweisen, die von den jeweils beteiligten Akteuren immer wieder aufs Neue ins Spiel gebracht werden. „Policy-relevant facts are the result of an intensive and complex struggle for political and epistemic authority. This is especially true where science and policy are difficult to distinguish and the guidelines for validating knowledge are highly contested.“ (Strassheim/Kettunen 2014, S.259) Ein Risiko, dass eine Erbringung von guten Grundlagen für evidenzbasierte Politik, die mit dem „Unternehmen Programmevaluation“ gemeint ist, zu einer politikgetriebene Evidenzerzeugung mutiert („policy-based evidence-making“), ist erst jüngst in internationalen Beobachtungen zur evaluativen Wissenproduktion unterstrichen worden (Kuhlmann 2015).

Im Wesentlichen kann somit heute für Evaluation im FTI-Bereich als nach wie vor gültig betrachtet werden, was in einem Bericht der *European Science Foundation* (ESF) vor einigen Jahren mit primärem Bezug auf Evaluation im Wissenschaftsbereich festgestellt wurde: „Nevertheless, the capacity worldwide, and the methods for carrying out evaluation are still poor compared with the amount spent on research and development. While there is the will to spend money to better understand the link between research and impact, there is not enough research, too few researchers and too few evaluation institutes to take up the questions. (...) Capacity development and new ideas are required. (...) We need to move away from symbolic or routine-based evaluation. Evaluation is most legitimate when it addresses a specific problem and can offer advice on decision-making. Evaluation exercises should have a specific goal and address a real problem. (...) Still, not all questions can be answered. (...) With every evaluation study there is the opportunity to expand the methodology. While it is sometimes useful to take the 'tried and tested' approach, at other times new pathways allow new

insights. Here, the boundaries to science studies are fluid. It is therefore useful to be in touch with the scientific community in the field." (ESF 2009, S.7f)

Wohl nicht nur für österreichischen FTI-Evaluationen können einige prinzipielle Forderungen als heute weiterhin gültig gelten, die für deren Anlage seit längerem von Seiten prominenter Forschung erhoben wurden (Kuhlmann/Meyer-Krahmer 1995, Georghiou/Roessner 2000, Georghiou 2003, Edler 2008, Kuhlmann 2009),. Die Berichtsanalysen bestätigen, dass Grundprobleme der FTI-Evaluation hier angesprochen sind, an denen auch die österreichische Evalautionspraxis nicht vorbeikommt:

- Ein langwährendes Problem der Evaluationen im FTI-Bereich besteht darin, dass Programme schlecht geklärte bzw. multiple und teilweise konfligierende Zielsetzungen aufweisen;
- Kurzfristig orientierte Analysen, die vor allem die Effektivität von Programmen im Auge haben, konfligieren mit längerfristig angelegten Erkenntnisinteressen, die ein gutes Verständnis von Zielen, Kontexten und Verbindungen zwischen Ergebnissen und Wirkungen voraussetzen;
- Politische Entwicklungen und Trends beeinflussen die Praxis der FTI-Evaluation mit;
- Einzelne Evaluationen sind auf Ergebnisebene kaum vergleichbar;
- Evaluation sollte sich zunehmend in die Lage versetzen bzw. in die Lage versetzt werden, eine essentielle Komponente in einem evolutionären Zugang zu FTI-Politik darzustellen.
- Probleme der kausalen Zuordnung von Effekten angesichts einer hohen Komplexität des Gegenstands erfordern evaluative Ansätze, die über die Effektivität einer Maßnahme als Teil des FTI-politischen Systems hinaus auch die Haltbarkeit tieferliegender Annahmen in den Blick nehmen.
- Evaluation steht kaum in Beziehung mit theoretischen Diskussionen und empirischen Arbeiten einschlägig befasster Disziplinen;
- Evaluationen haben das grundsätzliche Potenzial, zu einem besseren Verständnis von Forschung und Innovation als einer Vielzahl von Prozessen mit Feedback-Schleifen beizutragen. Dazu ist es allerdings notwendig, Evaluationsmethodik nicht mit Techniken der Datensammlung und -auswertung gleichzusetzen;
- Breiter angelegte Analysen sind notwendig, um den Stellenwert von Einzelmaßnahmen in Portfolios und in realen FTI-Systemen wirklich verstehen zu können;
- Im Zuge von Strukturreformen, aber auch einer beständigen graduellen Veränderung von Maßnahmenpaketen, wurde und wird eine Selbsttransformation des Systems mit offenem Ausgang und schlecht vorhersehbarer Dynamik in Gang gesetzt. Daraus resultiert ein Desiderat gerade für Evaluationen von Programmen und reformpolitischen Maßnahmen, ihren Einsatz als reflexives Instrument politischer Entscheidungsfindung in Multi-Akteurs-Szenarien zu stärken.

Diese Brennpunkte einer zukünftigen Aufmerksamkeit können zugleich mehreren Evaluationsstandards zugeordnet werden. Sie siedeln sich auf der Ebene der Definition der Evaluationszwecke und der diese möglichst stringent umsetzenden Evaluationsplanungen sowie der Ebene der Konzeptbildung und der Datenerfordernisse an (N2, N4, G3). Hier ist prinzipiell die logische Kette < *Evaluationszweck– Evaluationsschwerpunkte – Evaluationskriterien – Evaluations-fragen – Methodeneinsatz* > gefordert. Hinzu kommen Vorgehensweisen, die die evaluativen Unternehmungen in ihrem Stellenwert für eine Systemreflexion in Multi-Akteurs-Szenarien stärken können, indem sie an den Ebenen der oft siloisierten Evaluationsplanung und der Evaluationspräsentation ansetzen (N1, F5).

FTI-Evaluation ist wesentlich von dem Anliegen getrieben, die Funktionsweise von neuen Politken und Maßnahmen zu verstehen („What works"). Eine tragende Rolle der Evaluationsstandards darin ergibt sich insofern, als sie die anerkanntenmaßen produktivsten Wege zu einer Evalaution aufzeigen, die sodann hoffen kann, auch hohe Aufmerksamkeit zu finden. In diesem Zusammenhang ist es dennoch notwendig, auch Begrenzungen der Möglichkeiten einer noch so optimalen Evaluation hinsichtlich ihres Einflusses auf Politikgestaltungen zu sehen: „Evaluation standards in and by themselves do not generate good policy outcomes. [...] They may also be viewed as redundant where the value of the evaluation services provided can be reliably gauged in terms of the impact on the

quality of decisions reached (ascertained as an integral part of the evaluation process). On the other hand, just as one does not judge auditors by the profitability of the companies they serve, it is inappropriate to judge evaluators by the effectiveness of the programs and policies being evaluated." (Piciotto 2005: 34)

Was der „Goldstandard" für eine möglichst zielführende und aussagekräftige Evaluation ist, stellt eine jahrzehntelange Debatte unter EvaluationsexpertInnen dar. Waren es lange Zeit Kontrollgruppendesigns, die eine möglichst perfekte Messung von Zielerreichungen ermöglichen sollten, so wurden auch Nachteile dieser Herangehensweise erkannt. Echte Kontrollgruppendesigns und gute Counter-Factuals sind schwer zu realisieren, und Gruppenvergleiche auf der Effektebene können keine Auskunft darüber liefern, wie diese Effekte zustande gekommen sind, und erst recht nicht darüber, ob auch andere, ursprünglich so nicht erwartete Effekte zustande gekommen sind. Im Gegenzug wurden Ansätze entwickelt, die unter dem Begriff der „Theoriebasierten Evalaution" (*theory-based evaluation*) zusammengefasst werden. Diese Ansätze zielen darauf ab, die Operationsweise eines Programms anhand der ihm zugrunde liegenden Annahmen oder anhand von Annahmen, die zu seinem Operieren sinnvoll gemacht werden können, zu überprüfen und zu reflektieren, indem sie Kausalitätsbeziehungen zwischen Programmschritten und auf deren Basis entstehenden Effekten in den Blick nehmen. Mit dieser Herangehensweise eignen sich theoriebasierte Ansätze sowohl für die Auseinandersetzung mit Wirkungen (impacts) als auch für die Auseinandersetzung mit intermediären Programmstadien. Die strukturierte Aufarbeitung der Interventionslogik eines Programms wird hier zum zentralen Ansatz- und Ausgangspunkt für die gesamte Evaluationsanlage. Da oft keine gut ausformulierten bzw. in jeder Hinsicht vollständigen Programmdarstellungen vorliegen, wird die Auseinandersetzung mit der Logik der beabsichtigten Erzeugung von Outputs, Outcomes und Impacts auch als *Rekonstruktion der Programmtheorie* bezeichnet, die als zentrale Aufgabe einer Programmevaluation erachtet wird (Chen 1990, Rossi/Lipsey/Freeman 1999, Shadish/Cook/Campbell 2002, Donaldson 2007, Coryn et al. 2011, Chen 2015). Auf Möglichkeiten, solche Ansätze in Bezug auf Innovationspolitik zur Anwendung zu bringen, wurde bereits hingewiesen (Molas-Gallart/Davies 2006).

Der springende Punkt an theoriebasierter Evaluation ist die gezielte Auseinandersetzung mit einem Kausalmodell, das im Konzeptionsstadium eines Programms noch hypothetisch ist und in der konkreten Entfaltung des Programms auf reale Umsetzungsweisen und Umwelten trifft. „[T]heory-driven evaluation approaches share three fundamental characteristics: (a) to explicate the theory of a treatment by detailing the expected relationships among inputs, mediating processes, and short- and long-term outcomes, (b) to measure all of the constructs postulated in the theory, and (c) to analyze the data to assess the extent to which the postulated relationships actually occurred.'' (Shadish, Cook & Campbell 2002: 501) Sozialwissenschaftliche, ökonomische und andere wissenschaftliche Theorien im klassischen Sinn können dabei herangezogen werden und spielen je nach Ansatz eine unterschiedliche Rolle, doch geht es im Kern nicht um „Theorieanwendung".[11] Theoriebasierte Evaluation ist gegenüber Methoden neutral und gibt keiner Methode einen Primat dabei, die Kausalitätskette von beabsichtigten und beobachtbaren Outcomes, Outputs und Impacts zu verfolgen (vgl. Coryn et al. 2011). Neuere Konzepte differenzieren zwischen der *Theory of Change*, wie die Entstehung von beabsichtigten Verbesserungen vorgestellt wird, und der *Theory of Action*, welche Programmkonfiguration im Bezug auf vorfindliche Umwelten und Faktoren gewählt wird, um diese Verbesserungen auch tatsächlich eintreten zu lassen (vgl. Funnel/Rogers 2011). Für die Erfassung der Programmlogik, die meist in einem formalisierten Modell erfolgt, stehen heute unterschiedliche Ansätze zur Verfügung, die von einfachen Logic Charts bis hin zu differenzierten Outcome Chains reichen (ebd.).

Mit dem Ansatz der Realistischen Evaluation (Pawson/Tilley 1997) verlagert sich das Interesse der Analyse darauf, wie Programme mit ihren Kontexten interagieren: Was funktioniert für wen unter welchen Umständen? Die Analyse setzt sodann an der Identifiaktion von sogenannten Context-Mechanism-Outcomes (CMOs) an. Zentral ist dabei auch die Idee, Evaluation mit Forschungsarbeiten in Verbindung zu setzen und kumulativ nutzbar zu machen (vgl. u.a. Astbury/Leeuw 2010). Auch hier wurde bereits auf die Einsetzbar im FTI-Bereich hingewiesen (Edler et al 2014). Eine spezifische Spielart von theoriebasierter Evaluation liegt mit der Anfang der 2000er-

[11] Die Verwendung unterschiedlicher und unscharfer Begrifflichkeiten in der relevanten Literatur trägt zu einem verbreiteten Missverständnis bei, dass es sich stets um eine Arbeit mit Theorien wissenschaftlicher Disziplinen handeln müsse. Allerdings kann berechtigt davon ausgegangen werden, dass Theorien zu Forschung, Technologie und Innovation dazu beitragen können, die intendierte Handlungs- und Wirkungslogik eines FTI-Programms gut einzuordnen.

Jahre für den Bereich der Entwicklungszusammenarbeit entwickelten *Contrubution Analysis* vor. Hier wird die Rekonstruktion der Programmtheorie dazu genutzt, die wesentlichsten Gesichtspunkte einer grundsätzlich als hoch komplex anerkannten Entfaltung von Programmwirkungen zu fokussieren und in pragmatischer Weise auch Maßnahmenbündel evaluierbar zu machen (Mayne 2006, zur Relevanz für den FTI-Bereich Landsteiner 2014).

Aktuelle Herausforderungen in der FTI-politischen Arena

Früher gehegte Erwartungen an die Leistungskraft von Programmevaluationen wurden als unrealistisch erkannt. Die verfügbare Ressourcenausstattung von Programmevaluationen wird als wesentlicher Mitgrund dafür erachtet, dass immer wieder Informationsbedürfnisse nur eingeschränkt befriedigt werden konnten. Die Verankerung der Programmevaluationen als Bestandteile der Programmvereinbarungen erzeugt eine Spannung zwischen vorgegebenen Evaluationsfragestellungen und aktuellen Informationsbedürfnissen in einem hochdynamischen System, die wiederholt auf Kosten aktuell relevanter Erkenntnisse gegangen ist. Die Evaluationsfunktion ist im Governancesystem an feststehende Evaluationszeitpunkte und –budgets gebunden, die gemäß den Auskünften der AuftraggeberInnen zwar bisweilen mit einer gewissen Flexibilität gehandhabt werden können, im Großen und Ganzen aber jedenfalls enge Grenzen setzen. Evaluationsprojekte oder Studien evaluativen Charakters, die nicht in Programmdokumenten vorprogrammiert waren, wurden nur in seltenen Ausnahmefällen initiiert.

In allen Ressorts und Agenturen wurden Zuständigkeiten und Kapazitäten geschaffen, um Evaluationen durchführen und Evaluationsergebnisse auf einer strategischen Ebene handhaben zu können. Die Planung und Durchführung der Programmevaluationen, die primär in den Ressorts erfolgt, ist dort an die Fachzuständigkeiten für die evaluierten Programme gekoppelt. Abstimmungsprozesse intern und innerhalb der Principal-Agent-Beziehungen sind erforderlich, die im allgemeinen auf Grund vorhandener Kooperationsbereitschaft erfolgreich verlaufen, aber doch keiner institutionell klar verankerten Systematik folgen. Das Engagement, das für eine einzelne Programmevaluation aufgebracht wird, bemisst sich nicht zuletzt an den zum jeweiligen Zeitpunkt gegebenen Möglichkeiten der fachzuständigen Einzelpersonen im Rahmen auch anderer Agenden. Durchgehend wird dargestellt, dass im Rahmen der gegebenen Kapazitäten keine weiteren Spielräume mehr bestehen.

Die Weitergabe von Evaluationsergebnissen innerhalb der Hierarchien stellt sich als geregelter Vorgang dar. Dabei wird davon ausgegangen, dass Evaluationsergebnisse nur eine Informationsquelle unter vielen sind, auf die sich politische EntscheidungsträgerInnen stützen, und dass auch die politische Aufmerksamkeit für unterschiedliche Programme deutlich variiert. Eine Zirkulation von Evaluationsergebnissen hin zu anderen Fachabteilungen, die zur Stärkung der Wissensbasis in systemischer Hinsicht beiträgt, bemisst sich stark am Engagement von Einzelpersonen. In jüngster Zeit sind verstärkte Bemühungen zu beobachten, durch übergreifende hausinterne Präsentationen Evaluationsergebnisse in Umlauf zu setzen und Diskussionen zu initiieren, in denen auch nicht direkt mit dem evaluierten Programm befasste Abteilungen von den Evaluationsergebissen profitieren können und strategische Einschätzungen vorgenommen werden können. Eine institutionelle Verankerung derartiger wertvoller Vorgänge ist allerdings nicht gegeben, und eine durchgehende Systematik liegt nicht vor.

Im Rahmen der institutionellen Architektur bestehen einige wenige Berührungspunkte zwischen den Steuerungssegmenten im FTI-politischen Bereich, in denen zumindest potenziell Informationen über geplante und fertiggestellte Evaluationen ausgetauscht werden können. In erster Linie sind es jedoch Personen und Netzwerke, die einen übergreifenden Wissensfluss im Governance-System gewährleisten, sodass sich ein solcher Wissensfluss letztlich als akzidentiell darstellt.

Es besteht allseitiger Bedarf an verstärkt systemisch orientierten Erkenntnissen, durch die die Positionierung einer Maßnahme im breiten Kontext verschiedener Förderungs- und Steuerungsinstrumente ebenso aufgezeigt werden kann wie Optionen, in welcher Weise Bedarfslagen im systemischen Gesamtzusammenhang durch Einsatz und Konfiguration bestimmter Instrumente und Maßnahmen gezielt und in bestmöglicher Weise begegnet werden kann. In einzelnen Systemsegmenten besteht zudem Bedarf an Typen von Politikinformation, die mit den routinisierten Multi-Purpose-Evaluationen nicht gut abgedeckt werden können. Es geht hier (1) um intensivere Auseinandersetzungen mit Zielgruppen und Wirkungsweisen von Maßnahmen auf einer detaillierten Ebene, die in Richtung einer wissenschaftlichen Begleitforschung weisen, (2) um ein möglichst frühzeitiges Erkennen der Realitätshaltigkeit von Annahmen über die Wirkungsweise von programmen, und (3) um hoch reaktive und schlanke Studien evaluativen Charakters, die das FTI-politische Handeln in dynamischen Umwelten zeitnah unterstützen.

FTI-Programmevaluation siedelt sich so deutlicher als vor einem Jahrzehnt in einem Spannungsfeld an, das in den Standards un din der Evaluationstheorie als ein grundsätzliches Orientierungsfeld für Programmevaluationen angesehen wird. Eine Positionierung eher auf der wissenschaftlichen Seite von Intensivstudien oder auf der pragmatischeren Seite der prozessorientierten Unterstützungsleistungen bedeutet demnach nicht, dass die Denkprinzipien der Programmevaluation vollkommen verlassen werden müssen. Umfangreiche Programmevaluationen bieten ein grundsätzliches Leistungsspektrum, das von keiner anderen Vorgehensweise ersetzt werden kann, insbesondere dann, wenn sie sich nach längerem Zeitablauf mit Prozessen der tatsächlichen Entfaltung eines Programms auseinander setzen, die erst dann beobachtbar werden können.

Evaluationsqualität einschätzen

Die Konzeptualisierung der Nutzungsformen durch die internationale Evaluationsforschung erstreckt sich auch auf Formen und Gründe der Nicht-Nutzung. Diese Diskussion weist nachdrücklich darauf hin, dass Evaluationen auch mit Recht nicht genutzt werden, wenn sie in unzureichender Weise erstellt wurden. Sollte ein Auftraggeber zur Ansicht gelangen, dass eine Evaluation unzureichend durchgeführt wurde, so wäre die Nutzung von deren Ergebnissen als missbräuchliche Nutzung einzustufen (vgl. z.B. Alkin & Taut 2003). In den USA wurden auch schon Metaevaluationen einzelner Evaluationen durchgeführt, um ihre berechtigte Verwendbarkeit zur Begründung von politischen Veränderungen zu klären, indem ihre daten nochmals neu berechnet wurden („evaluation audit" von House 1997). Die Standards gehen davon aus, dass eine Evaluation günstigenfalls in ihrer Planung und Durchführung begleitend beraten werden sollte (begleitende Metaevaluation), so wie etwa auch die vorliegende Metaevaluation durch zwei eminente Experten in ihrem Planungsstadium unterstützt wurde. Die Problematik der gerechtfertigten oder weniger gerechtfertigten Nutzung von Evaluationsergebnissen verweist aber auch auf Qualitätsüberprüfungen der Evaluationsberichte durch ihre Auftraggeber.

In der österreichischen FTI-Evaluationspraxis gibt es, wie InterviewpartnerInnen dargelegt haben, eine Aufmerksamkeit für Evaluationsqualität jedenfalls heute durchaus. Manche InterviewpartnerInnen drücken unmissverständlich aus, dass sie höchste Evaluationsqualität für unabdingbar erachten, wenn der Anspruch einer strategisch hochwertigen FTI-Politik erhoben wird. Ansatzpunkt für eine kritische Betrachtung von Evaluationsergebnissen waren oft die Schlussfolgerungen und Empfehlungen, die auf ihre Absicherung in erbrachten Daten hin reflektiert wurden. Nicht eindeutig klar geworden ist in den angestellten Erhebungen, inwiefern hier auch ein Irritationsfaktor mitgespielt hat bzw. mitspielt, sodass eine Übereinstimmung der Evaluationsergebnisse mit vorhandenen Sichtweisen oder eine Abweichung davon die Motivation zur Überprüfung mit beeinflussen. Auf Basis vorliegender internationaler Ergebnisse der Evaluationsforschung wird grundsätzlich davon auszugehen sein, dass auch ein derartiger Einflussfaktor ab und an wirksam sein wird. Von den InterviewpartnerInnen, die die Hauptakteure des Evaluationsgeschehens und der Plattform fteval darstellen, wurde allerdings überzeugend vermittelt, dass in ihren Handlungsbereichen ein Interesse an neuen Inputs durchaus vorhanden ist, da diese die Voraussetzung für einen Zugewinn an Erkenntnissen und Perspektiven bilden. Wurden hier teilweise Begriffe wie „Tabus brechen" und „überrascht werden" verwendet, so wurde andererseits doch in der EvaluatorInnenbefragung eine offene Angabe erhalten, dass große Kritikbereitschaft der AuftraggeberInnen nicht erlebt wurde. Im Gesamtbild werden die Gesamtergnibsse zur Nützlichkeit der Programmevaluationen dadurch nicht schwer beschädigt. Eine Variationsbreite zwischen unterschiedlichen Evaluationen wird immer vorliegen, da es sich stets um individuell konfigurierte Fälle in komplexen Kontexten handelt. Auf nicht-kritische Evaluationsergebnisse nicht mit einem gewissen Vorbehalt zu reagieren, kann genauso Probleme verursachen, wie sachlich fundierte Kritik in die Schranken weisen zu wollen. Stets aber kann der späte Zugriff auf die Evaluationsergebnisse, wie kritisch oder unkritisch diese auch ausgefallen sein mögen, eine qualitätsorientierte Zugangsweise im Planungs- und Durchführungsstadium der Evaluationen nicht ersetzen.

6.2 Empfehlungen

Es wird auf Basis der drei Datenquellen und deren integrierender Analyse ersichtlich, dass es sich bei der Frage der Evaluationsqualität in Bezug auf Nützlichkeit und tatsächlich zustande kommenden Evaluationsnutzen nicht um Einzelursachen handelt, sondern um Syndrome und Faktorenbündel von erheblicher Komplexität. Die bisherige Evaluationspraxis im FTI-Bereich erweist sich als gleichermaßen durch Gestaltungsmerkmale einzelner Programmevaluationen wie durch Kontextfaktoren bedingt. Damit ist auch nicht die *eine* Lösung greifbar, die eine entscheidende Weiterentwicklung über den bisher erreichten Stand hinaus bewirken könnte.

Limitierungen für die Gestaltung von Programmevaluationen und die Entstehung von Evaluationsnutzen ergeben sich aus Merkmalen des institutionellen Arrangements. Evaluationsberichte und die hinter diesen Produkten stehenden Evaluationsprozesse gehen auf das, was die Evaluationsstandards als optimale Schritte hin zu hoher Nützlichkeit bezeichnen, bislang nur bedingt zu. Damit lassen auch Evaluationsprodukte und –prozesse Nutzen entstehen, die die von den Standards empfohlenen bzw. als notwendig erachteten Evaluationseigenschaften nicht optimal verwirklichen. Unter dem Gesichtspunkt einer größtmöglichen Nützlichkeit auf der Basis hervorragender Evaluationsqualität muss es zweifellos angelegen sein, von einer strukturell kompromisshaften Situation zu verbesserten Bedingungen für die Planung, Durchführung, Kommunikation und Nutzung von Programmevaluationen zu gelangen. Die Evaluationsstandards sind als praktische Anleitung zur Bewältigung von Problemen bei der Nutzenentstehung konzipiert, doch können sie Probleme nicht lösen, die außerhalb der Reichweite eines konkreten Evaluationsprojekts liegen, und gute Lösungen entlang der Standards müssen von EvaluationsauftraggeberInnen in der Gestaltung der Evaluationsaufträge auch ermöglicht werden. Strukturell ermöglichte Potenziale für die Planung, Durchführung, Kommunikation und Nutzung von Programmevaluationen bleiben sodann in den jeweiligen Projekten auf der Basis von Kapazitäten und Kompetenzen auszufüllen.

Die Metaevaluation gelangt daher zu Empfehlungen, die sich sowohl auf einer evaluations-theoretischen Ebene als auch auf der Ebene der institutionellen Einbettung der Evaluationsfunktion ansiedeln. Die vorgelegten Empfehlungen sind an der Weiterentwicklung einer in sich dynamischen und systemevolutiven Evaluationspraxis orientiert. Da die bisherige Evaluationspraxis in nachvollziehbarer Weise bereits Nutzen erzeugt hat, setzen die Empfehlungen nicht auf eine radikal-disruptive Veränderung, die aus einer Orientierung an Governancemodellen anderer Länder grundsätzlich abgeleitet werden könnte, aber hinsichtlich tatsächlicher Transferierbarkeit und Eintreten der erhofften Effekte doch auch mit einigen Ungewissheiten einhergeht. Für die evaluationsmethodische Ebene würde eine Benennung aller denkbaren Verbesserungsoptionen freilich darauf hinauslaufen, den gesamten Gehalt der Standards zu referieren. Diesbezügliche Empfehlungen werden nur für diejenigen Gesichtspunkte ausgesprochen, die als die wesentlichsten erscheinen. Letztlich beruht eine hoch entwickelte Evaluationskultur auch auf gesellschaftlich-kulturellen Faktoren wie der Offenheit für sachlich fundierte Kritik und der Bereitschaft zur offenen Diskussion, die sich freilich einer gezielten Beeinflussung entziehen.

Die folgenden 20 Empfehlungen werden ausgesprochen:

1. Programmevaluationen sollten in Zukunft weiterhin durchgeführt werden, da sie in der Vergangenheit wertvolle Beiträge zur zielgerechten Umgestaltung und Neukonzeption FTI-politischer Maßnahmen erbracht haben, die noch über die Ebene der jeweils evaluierten Programme hinaus reichen. Um die Produktivität der Programmevaluationen über das bisherige Maß hinaus weiter steigern zu können, sollten sie mit den folgenden Empfehlungen benannten Schritten einhergehen.
2. Die derzeit gegebene Verankerung der Evaluationsfunktion bei den Institutionen, die für die Konzeption und Umsetzung von FTI-Programmen zuständig sind, sollte beibehalten werden. Entscheidende Kapazitäten für die Planung, Durchführung und Verwertung von Programmevaluationen wurden hier über Jahre hinweg aufgebaut. Die Verankerung bei den Programmverantwortlichen sorgt auch für ein Commitment zu den Programmevaluationen, das für in der Vergangenheit entstandenen Evaluationsnutzen wesentlich war. Eine Weiterentwicklung der Evaluationskultur im FTI-Bereich sollte als pfadabhängige Entwicklung auf dieser wertvollen Grundlage gedacht werden.

3. Programmevaluationen sollten in Zukunft mit denjenigen Ressourcen ausgestattet werden, die eine konzeptgemäße Analyse des evaluierten Programms unter Heranziehung aller für die Evaluationsschwerpunkte und –fragestellungen benötigten Informationsquellen tatsächlich ermöglichen und eine gute Durchführung gemäß dem Qualitätsverständnis der internationalen Standards für Programmevaluation gewährleisten.

4. Programmevaluationen sollten künftig stärker auf eingegrenzte Evaluationsschwerpunkte fokussiert werden. Dadurch können unter Bedingungen begrenzter Ressourcen intensivere und genauere Untersuchungen zu den gewählten Schwerpunkten durchgeführt werden. Jeweils nicht gewählte Evaluationsschwerpunkte können gegebenenfalls durch eine weitere Evaluation verfolgt werden. Dabei können dann auch andere Evaluationsteams zum Einsatz kommen, was zu einer Anreicherung der Sichtweisen auf das untersuchte Programm auf Basis unterschiedlicher Kompetenzen beitragen kann.

5. Programmevaluationen sollten verstärkt in ihrer Prozessqualität begriffen und auf dieser Ebene in Planung und Durchführung gestärkt werden. Die DeGEval-Standards mit ihrem Interpretationshintergrund der Joint Committee-Standards weisen auf Schritte hin, durch die im Planungs- und Durchführungsstadium von Programmevaluationen Qualität in unterschiedlichen Hinsichten gestärkt und sicherstellt werden kann. Die *Plattform fteval* sollte sich mit solchen Möglichkeiten auseinandersetzen, da sie Voraussetzungscharakter für die Erzielung späterer Evaluationsergebnisse und deren Nutzungspotenzial für verschiedene Akteursgruppen haben.

6. Evaluationsberichte sollten in jeder Hinsicht klar und in einer auch für Außenstehende gut verständlichen Weise abgefasst werden. Dies ist insbesondere als Voraussetzung dafür zu verstehen, dass es zu einer verstärkten Nutzung von Programmevaluationen in anderen FTI-politischen Bereichen und nach dem Denkprinzip einer vermehrten systemreferentiellen Selbststeuerung der FTI-Akteure kommen kann.

7. Alle Evaluationsberichte sollten systematisch ein Kapitel beinhalten, in dem die Gesamtvorgehensweise der Evaluation in methodischer wie organisatorischer Hinsicht konzis und vollständig dargestellt wird und auch auf Vor- und Nachteile der tatsächlich durchgeführten Analyse hingewiesen wird. Eine derartige kompakte Übersicht über die Gesamtvorgehensweise erscheint insbesondere hinsichtlich einer stärkeren Nutzung von Evaluationsergebnissen in einem gesamt-systemischen Zusammenhang relevant, damit auch Akteure, die mit den unmittelbaren AuftraggeberInnen nicht identisch sind, auf die erbrachten Evaluationsergebnisse gut zugreifen können. In der Darstellung der Vorgehensweisen sollte es auch Mut zum Ausweis von Lücken geben, da keine Programmevaluation alles beleuchten kann, was theoretisch zu einem Programm untersucht werden könnte. Auch ein abholbares Wissen darüber, was noch nicht intensiv untersucht werden konnte, sollte als produktiver Beitrag zum FTI-politischen Wissens- und Informationssystem betrachtet werden, damit dieses im Weiteren produktiv ausgestaltet werden kann.

8. Es sollte eine verstärkte Auseinandersetzung mit der gezielten Anwendung von Bewertungsmaßstäben auf die evaluierten Programme angestrebt werden. Dabei geht es nicht nur darum, wie Zielerreichungen gemessen und eingeschätzt werden, was oft zum Evaluationszeitpunkt in dieser Form noch gar nicht möglich ist, sondern auch und gerade um die wohlbegründete Einordnung der Beobachtungen zu Aspekten der Programmentfaltung. Die von einer Evaluationsstudie angewendeten Bewertungsmaßstäbe sollten als vitale Konzeptfrage begriffen und im Planungsstadium als integraler Bestandteil des übergreifenden Evaluationskonzepts vereinbart und festgelegt werden. Konsistente Bewertungsmaßstäbe verkörpern sich unter anderem in der Verfolgung von Kohärenz und Konsistenz von Programmzielen und Programmkomponenten in ihrer Umsetzung, in einer Festlegung, wie die Sichtweisen verschiedener Akteursgruppen auf das evaluierte Programm zur Gesamteinschätzung führen, in vorab festgelegen Kriterien zur Einordnung späterer Messergebnisse, in oder in der gezielten Bestimmung von Messgrößen (etwa bei einem Programmziel „Kooperation“ die Quantität von Kooperationsbeziehungen versus qualitative Eigenschaften von eingegangenen Kooperationen).

9. Dem Risiko eines Lock-Ins in üblichen Herangehensweisen an Evaluation, die mit Ermüdungserscheinungen der Evaluationspraxis in Zusammenhang stehen, sollte durch eine systematische professionelle Beratung von Evaluationsplanungen und –prozessen gegengesteuert werden. Eine solche Beratung wird vor allem dann ein probates Mittel darstellen, wenn sie nicht nur FTI-spezifische Kompetenzen heranzieht, sondern auch evaluationsmethodische Kompetenzen, die den Konnex zu Entwicklungen und Know-How anderer Bereiche herstellen.

10. Eine Intensivierung der Planungsphasen der Programmevaluationen sollte angestrebt werden, um das Risiko zu minimieren, dass beschränkte Ressourcen in letztlich ergebnisarme Untersuchungsschritte fließen. Dafür bietet sich das international anzutreffende Modell einer sogenannten *„Inception Phase"* am Beginn einer Programmevaluation an, in der sich die beauftragten EvaluatorInnen intensiv mit der Datenlage, methodischen Möglichkeiten im Rahmen der gegebenen Ressourcen, und der Beantwortbarkeit der vorgesehenen Evaluationsfragen auseinandersetzen. Diese genaue Abwägung bildet sodann die Grundlage für ein bestmögliches Evaluationsdesign, das der im Anschluss durchgeführten Evaluation zugrunde gelegt wird. Das Modell zielt darauf ab, so realistische Erwartungen wie möglich an eine Evaluation zu entwickeln und die für die Evaluation verfügbaren Ressourcen so gut wie möglich zu nutzen. Vergaberechtliche Voraussetzungen für die Nutzbarkeit dieses Modells bleiben zu prüfen.

11. Die institutionelle Verankerung der Evaluationsfunktion in den Ressorts und Agenturen sollte weiter gestärkt werden. In den auftraggebenden Ressorts und Agenturen existieren HauptansprechpartnerInnen für Evaluationsangelegenheiten und VertreterInnen der Institutionen in der *Plattform fteval*, doch ist bis heute keine dieser Personen ausschließlich mit Evaluationsangelegenheiten betraut, um sich dieser komplexen und anforderungsreichen Materie vollständig widmen zu können. Ressourcen von fachzuständigen MitarbeiterInnen für die Auseinandersetzung mit anderweitig erarbeiteten Programmevaluationen sind kaum vorhanden. Eine spezialisierte, hoch professionelle Evaluationsabteilung oder Stabstelle, die sich mit der Planung der Programmevaluationen, dem Evaluationsmanagement, einer Qualitätskontrolle und der Verwertung und Weitergabe der Evaluationsergebnisse für das ganze Haus befasst, stellt in diesem Zusammenhang das Idealbild dar, das einen entscheidenden Schritt zur Überwindung der Variabilität im Umgang mit einzelnen Evaluationen verkörpern würde.

12. Die Lernfunktion der Programmevaluationen sollte künftig durch eine Flexibilisierung der Auslösung und Intensität der einzelnen Evaluationen weiter gestärkt werden. Frei allozierbare Evaluationsbudgets könnten die Gestaltung von Programmevaluationen im Aktualitätsbezug sowie unter Gewichtung von Informationsbedarfslagen ermöglichen. Nicht alle Programme brauchen in gleicher Weise evaluiert zu werden, um in einem übergreifenden FTI-politischen Informationssystem wesentliche Erkenntnisse zu erzielen. Eine Flexibilisierung würde somit zu zielgerechten Investitionen in anspruchsvollere Evaluationen und Studien und zu einer effektiveren Nutzung der im System vorhandenen Ressourcen beitragen. Programmwirkungen könnten zu passenderen Zeitpunkten analysiert werden, als es bislang der Fall war. Thematische Evaluationen, etwa zu Programmfamilien oder Zielgruppen, und Instrumentenevaluationen könnten verstärkt durchgeführt werden.

13. Programmevaluationen sollten verstärkt als übergreifende und konzise Analysekonzepte verstanden und angelegt werden. Evaluationsmethodische Konzepte und Tools, die für eine möglichst zielführende Evaluation von Programmen über die letzten beiden Jahrzehnte international entwickelt wurden, sollten dabei herangezogen werden. Zu empfehlen ist eine Zuwendung zu Ansätzen, die unter dem Sammelbegriff der Theorie-basierten Evaluation (*theory-based evaluation*) bekannt sind. Diese Ansätze sind gezielt dafür konzipiert, die Einlösbarkeit von Programmannahmen in der realen Programmentfaltung zu beleuchten und geschärfte Umgangsweisen mit der Kausalitätsproblematik zu ermöglichen, wie und inwieweit ein Programm zu intendierten Veränderungen beiträgt. Mit der Zuwendung zu ihnen würde die Evaluationspraxis im FTI-Bereich Analysestrategien zur Anwendung bringen, die in anderen Politikbereichen auf internationaler Ebene und in internationalen Organisationen bereits eingesetzt werden. Die avancierten Ansätze der Realistischen Evaluation (*realistic evaluation*) und der *Contribution Analysis* könnten aufgegriffen werden, um zu einem vertieften Verständnis der Wirkungsweise von Programmen in ihrer Kontextabhängigkeit zu gelangen und komplexe Programme, Programmfamilien, Portfolien und Maßnahmenbündel zielführend und in pragmatischer Weise zu analysieren. Allseitige Ressourcen für die Arbeit mit qualitativen Daten und notwendige Interaktionen zwischen EvaluatorInnen und AuftraggeberInnen während der Evaluationsdurchführung sind freilich vorausgesetzt..

14. Programmdokumente sollten so eingehend wie möglich darlegen, wie Zielsetzungen systematisch gegliedert sind, welche Outputs die verschiedenen Programmaktivitäten erzeugen sollen, und welche Annahmen darüber gemacht werden, wie diese Outputs zu Outcomes und weiteren Entwicklungen hin zu Zielerreichungen führen. Eine möglichst gute Darstellung der intendierten Wirkungsweise der Programme durch die Programmeigentümer bei der Programmkonzeption bildet den Gegenpol zur evaluatorischen Aufarbeitung einer Programmlogik und deren Ausgestaltung in der Programmwirklichkeit. Die Konzeptualisierung der intendierten

Wirkungsweise der Programme kann im Planungsstadium durch ex ante-Evaluationen unterstützt werden. Freilich können ex ante-Evaluationen spätere Überprüfungen nicht ersetzen, wie sich Programmeffekte im realen Operieren des Programms herstellen oder mit Hindernissen konfrontiert sind.

15. Programmevaluationen sollten gemeinsam mit allen verwandten und ergänzenden Bestandteilen eines übergreifenden FTI-politischen Wissens- und Informationssystems durch Publikation verfügbar gemacht werden, um auch Synergien zwischen Studien unterschiedlichen Typs allgemein nutzbar zu machen. Die konkrete Bezeichnung von Programmevaluationen, Reviews, Assessments oder wissenschaftlichen Studien evaluatorischen Charakters sollte nicht zum Anlass werden, wertvolle Informationspotenziale zu beschneiden. Ein Repositorium für alle evaluativen und wissenschaftlichen Studien kann im Bedarfsfall in allgemein zugängliche Bereiche und Bereiche mit Zugangsbeschränkungen gegliedert werden. Nicht-Publikation ist gerechtfertigt und angebracht, wenn in einer systematischen Qualitätskontrolle zum Schluss gekommen wird, dass durch die Publikation unzuverlässige oder irreführende Information zur Nutzung freigegeben würde. Die Nutzbarkeit jedweder evaluativer Information wird von einer adäquaten Dokumentation über den genauen Charakter dieser Information abhängig bleiben. Im Verständnis der Evaluationsstandards ist jeder Nutzung von Evaluationsergebnissen eine umfassende Auseinandersetzung mit der genauer Vorgehensweise und Durchführungsqualität der betreffenden Evaluation vorausgesetzt. Eine bloße Verfügbarkeit von Datenbeständen, die unter nicht genau verstehbaren Ausgangsbedingungen in Bezug auf nicht genau bekannte Informationsbedürfnisse erarbeitet wurden, sollte nicht als ausreichend erachtet werden.

16. Jede Programmevaluation sollte bei ihrer Publikation von einer „Management Response“ begleitet werden, die die Kenntnisnahme der Evaluationsergebnisse auf Ebene des Top Managements bestätigt, eine Positionierung zu diesen Ergebnissen angibt, und damit auch Verbindlichkeit erzeugt. Dabei geht es nicht etwa um eine automatische Übernahme von Evaluationsergebnissen, sondern im Gegenteil um das Produkt einer aktiven Auseinandersetzung mit ihnen. Dieser Weg wird beispielsweise von der Deutschen Forschungsgemeinschaft (DFG) bereits beschritten und wurde neuerdings auch von einer Agentur im österreichischen FTI-Governancesystem eingeschlagen.

17. Der RFTE sollte die ihm zur Verfügung stehenden Mittel nützen, um in der evaluativen Wissensproduktion offen bleibenden Informationsbedarf durch gezielte Vergabe von Studien in aktualitätsbezogener und flexibler Weise zu befriedigen. Dies erscheint im Hinblick auf intensive Analysen zu Themen und Segmenten des FTI-Systems ebenso relevant wie im Hinblick auf übergreifende, systemisch ausgerichtete Analysen. Ein Charakter wissenschaftlicher Begleitforschung, die in den Ressorts und Agenturen keinen Ort hat, könnte dabei zum Tragen kommen. Im Hinblick auf den systemischen Stellenwert solcher Studien erscheint eine Abstimmung mit den relevanten FTI-politischen Akteuren sinnvoll und wichtig.

18. Eine Koordinationsfunktion für FTI-Evaluationen sollte geschaffen werden, die sich mit möglichen Synergiebildungen zwischen an verschiedenen Systemstellen angesiedelten Evaluationsaufgaben und -ressourcen befasst, um durch Abstimmungen und Beratungen die gegenwärtige Zersplitterung der Evaluationsaktivitäten und Kleinteiligkeit im Analytischen zu überwinden. Dadurch kann ein Potenzial ausgeschöpft werden, das aus einer Bündelung von Ressourcen und Erkenntnisinteressen resultiert. Ressort- und Agentur-übergreifende Abstimmungsleistungen könnten erbracht werden, deren Machbarkeit unter den gegebenen Bedingungen eingeschränkt ist. Erträge hinsichtlich stärker systemisch ausgerichteter Fragestellungen zum Stellenwert von einzelnen Maßnahmen und Steuerungen sind zu erwarten.

    Dies kann zugleich als sinnvolle Alternative zu ebenso seltenen wie schwer initiierbaren Großunternehmungen wie der Systemevaluation 2009 erachtet werden, indem systemische Fragestellungen zum Gegenstand eines rollenden Verfahrens werden. Eine solche Koordinationsfunktion ist jedenfalls mit hohen fachlichen Kompetenzen und adäquaten Ressourcen auszustatten. Es bleibt zu prüfen, ob eine Einrichtung möglich ist, ohne bestehende Rechtsbestände anzutasten. Die Konfiguration und Einrichtung sollte durch eine Studie vorbereitet werden, die sich mit internationalen Beispielen auch außerhalb des FTI-politischen Bereichs befasst.

19. Ein Diskussionsforum sollte geschaffen werden, das Evaluationsergebnisse an ein breiteres Fachpublikum heranträgt, das über den engen Kreis der in der Plattform fteval versammelten Akteure hinausreicht und ProgrammangerInnen und Programmverantwortliche an unterschiedlichen Systemstellen genauso anspricht wie Akteursgruppen im FTI-System. Hierdurch

können Wissensflüsse in Gang gesetzt und Diskussionen ausgelöst und angereichert werden, die für ein System systemreferentieller und selbstreflexiver Akteure relevant sind. Der derzeitigen starken Abhängigkeit von Wissensflüssen im FTI-politischen Governancesystem von Personen und Netzwerken würde damit gegengesteuert. Ebenso würde der Umstand, dass auf der Basis einer bloßen Publikation evaluative Information eine Holschuld für etwaige InteressentInnen bleibt, behoben. Ein solches Diskussionsforum kann optional mit der vorgenannten Koordinationsfunktion verbunden werden, aber auch eine getrennt angesiedelte Systemfunktion darstellen.

20. Hinsichtlich einer substantiellen Stärkung der Lernfunktion der Programmevaluationen ist die derzeitige Kombination der unterschiedlichen Evaluationszwecke des Lernens und der Rechenschaftslegung, die für die Programmevaluationen durch deren institutionell-rechtliche Verankerung als Schnittstellenfunktion zwischen Fachressorts und Finanzressort stets gegeben ist, nicht als produktiv zu erachten. Nachdem mit der Wirkungsorientierten Folgenabschätzung (WFA) eine andersartige Evaluationsfunktion im Bezug auf Rechenschaftslegung geschaffen wurde, könnte überlegt werden, inwiefern die Lernfunktion der Programmevaluationen von Zwecken der Rechenschaftslegung künftig getrennt werden kann. Zwecke der Programmdokumentation könnten verstärkt in die Hände der Agenturen gelegt werden, die bereits jetzt wesentliche Teile der Datenbasen erarbeiten, die in Programmevaluationen verwendet werden. Im Gegenzug könnten Evaluationen dann verstärkt Analyseschritte setzen, die nicht der Gefahr eines Lock-Ins in vorab festgelegten Datenstrukturen ausgesetzt sind.

LITERATUR

Alkin, M. C. (2012). Evaluation roots (2nd ed.). Thousand Oaks, CA: Sage.

Alkin M.C. (1990) Debates on Evaluation. Thousand Oaks, CA: Sage.

Alkin M.C. , Taut S.M. (2003), Unbundling Evaluation Use. Studies in Educational Evaluation, 29 (1), 1-12.

Alkin M. C., Daillak R., White P. (1979). Using evaluations: Does evaluation make a difference? Beverly Hills: Sage.

Arnold, E. (2004). Evaluating research and innovation policy: a systems world needs systems evaluations. In: Research Evaluation, 13(1), 3-17.

Astor M., Fischl I., Hoffmann J., Koglin G., Kulicke M., Sheikh S., Wessels J., Whitelegg K. (2014), Evaluation von Forschungs-, Technologie und Innovationspolitik in Deutschland und Österreich – ein Überblick, In: Böttcher W., Kerlen C., Maats P., Schwab O., Sheikh S. (DeGEval-Vorstand) (Hg.), Evaluation in Deutschland und Österreich. Stand und Entwicklungsperspektiven in den Arbeitsfeldern der DeGEval – Gesellschaft für Evaluation, Münster- New York, 139-149.

Astbury B., Leeuw F.L. (2010) Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation. American Journal of Evaluation 31(3), 363-381.

Balthasar A. (2007), Institutionelle Verankerung und Verwendung von Evaluationen: Praxis und Verwendung von Evaluationen in der schweizerischen Bundesverwaltung, Zürich-Chur: Rüegger.

Barjak F. (2013), Wirkungen innovationspolitischer Fördermassnahmen in der Schweiz. Staatssekretariat für Bildung, Forschung und Innovation SBF.

Beywl W. (2006), The Role of Evaluation in Democracy: Can it be Strengthened by Evaluation Standards? A European Perspective. Journal of MultiDisciplinary Evaluation, Number 6, November 2006, 10-29.

Beywl W. (2001), Die Standards für Evaluation der DeGEval - Vorstellung und Einladung zum Dialog. In: Plattform Forschungs- und Technologieevaluierung Newsletter Nr. 14 Dezember 2001, S. 16-19.

Beywl W., Speer S. (2004), Data- and Literature-Based Reflections on Western European Evaluation Standards and Practices. New Directions for Evaluation No. 104, Winter 2004, 3-54.

Beywl W., Taut S. (2000), Standards: Aktuelle Strategie zur Qualitätsentwicklung in der Evaluation. Vierteljahrshefte zur Wirtschaftsforschung 69. Jahrgang, Heft 3/2000, S. 358–370.

Biegelbauer P. (2013), Wie lernt die Politik - Lernen aus Erfahrung in Politik und Verwaltung, Wiesbaden: Springer VS Verlag.

Bovens M., 't Hart P., Kuipers S. (2006), The Politics of Policy Evaluation. In: Moran M., Rein M., Goodin R.E. (eds.), The Oxford Handbook of Public Policy, New York:Oxford University Press, 319-335.

Chen H.T. (2015), Practical Program Evaluation. Theory-Driven Evaluation and the Integrated Evaluation Perspective, 2nd ed., Sage.

Chen, H. T. (1990). Theory-driven evaluations. Newbury Park, CA: Sage.

Cooksy L.J., Caracelli V.J. (2009), Metaevaluation in Practice. Selection and Application of Criteria, Journal of MultiDisciplinary Evaluation, Volume 6, Number 11: 1-15.

Cooksy L.J., Caracelli V.J. (2005), Quality, Context, and Use Issues in Achieving the Goals of Metaevaluation. American Journal of Evaluation, Vol. 26 No. 1, 31-42.

Coryn L.S., Noakes L.A., Westine C.D., Schröter D.C. (2011), A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009. American Journal of Evaluation June 2011 vol. 32 no. 2 199-226.

Cousins J.B. (2006), Non-academic Impact of Research through the Lens of Recent Developments in Evaluation. Paper presented at the conference 'New Frontiers in Evaluation', Vienna, April 2006.

Cousins J. B., Leithwood K. A. (1986), Current empirical research on evaluation utilization. Review of Educational Research, 56, 331–364.

Cousins, B.J., Whitmore E. (1998): Framing Participatory Evaluation. New Directions for Evaluation, 80, S. 87-105.

Cronbach L. (1984), Ninety-five theses for reforming program evaluation. In: Madaus G.F., Scriven M.S., Stufflebeam D.L. (1984), Evaluation Models. Viewpoints on Educational and Human Services Evaluation, 2nd ed., Boston-The Hague-Dordrecht-Lancaster:Kluwer-Nijhoff Publishing: 405-412.

DeGEval – Gesellschaft für Evaluation, 2008, Standards für Evaluation, 4. Aufl., Mainz: DeGEval.

DeGEval – Gesellschaft für Evaluation, 2001, Standards für Evaluation, Mainz: DeGEval.

Donaldson S.I. (2007), Program Theory-Driven Evaluation Science: Strategies and Applications, New York-London: Taylor & Francis.

Edler J. (2008), Evaluation of systems and portfolios: using existing evaluation to make sense at systems level A concept development. Key Lecture at the OECD Workshop "Enhancing Research Performance Through Evaluation and Priority Setting", 15-16 September 2008, OECD.

Edler J., Cunningham P., Gök A., Shapira P. (2014), Innovation Policy Impact. Lessons From a Comparative Study on Innovation Policy Instruments. Presentation to the EU SPRI Annual Conference Manchester, June 19 2014.

Edler J., Ebersberger B., & Lo V. (2008). Improving policy understanding by means of secondary analyses of policy evaluation. In: Research Evaluation, 17(3), 175-186.

Elg L., Hakansson S. (2012), Impacts of Innovation Policy - Lessons from VINNOVA's impact studies: VINNOVA –Verket för Innovationssystem.

EPEC - European Policy Evaluation Consortium (2011), Understanding the Long Term Impact of the Framework Programme. Final Report to the European Commission DG Research, 5 December 2011.

ESF – European Science Foundation (2009), Evaluation in Research and Research Funding Organisations: European Practices. A report by the ESF Member Organisation Forum on Evaluation of Publicly Funded Research, Brussels: ESF.

European Policy Evaluation Consortium (EPEC) (2011), Understanding the Long Term Impact of the Framework Programme. Final Report to the European Commission DG Research, 5 December 2011.

Finn Jr., C. E., Stevens, F. I., Stufflebeam, D. L., & Walberg, H. J. (1997). A meta-evaluation. In H. L. Miller, Jr. (Guest Ed.), The New York City Public Schools Integrated Learning Systems Project: Evaluation and meta-evaluation. International Journal of Educational Research, 27(2), 159-174.

Fleischer D. N., Christie C. A. (2009), Evaluation Use - Results From a Survey of U.S. American Evaluation Association Members. American Journal of Evaluation Volume 30 (2), 158-175.

Funnel S.C., Rogers J.P. (2011), Purposeful Program Theory. Effective Use of Theories of Change and Logic Models, San Franciso: John Wiley & Sons.

Georghiou L. (2003), Evaluation of research and innovation policy in Eurpoe - new policies, new frameworks? In: Shapira P., Kuhlmann S. (eds), Learning from science and policy evaluation: experiences from the Unitied States and Europe, Cheltenham – Northampton MA, 65-80.

Georghiou L., Roessner D. (2000), Evaluating technology programs: tools and methods, Research Policy 29 (2000), 657–678.

Gök A., Mollas-Gallart J. (2014), STI Policy Evaluation: An Isolated Academic and Practice Field. EUSPRI Conference 18/06/2014.

Good B. (2012), Assessing the effects of a collaborative research funding scheme: An approach combning meta-evaluation and evaluation synthesis. Research Evaluation 21, 381-391.

Good B. (2006), Technologie zwischen Markt und Staat. Die Kommission für Technologie und Innovation und die Wirksamkeit ihrer Förderung, Zürich-Chur:Rüegger.

Hanberger A. (2013), Framework for exploring the interplay of governance and evaluation. Scandinavian Journal of Public Administration 16(3): 9-27

Hansson, F. (2006), Organizational use of evaluations: Governance and con- trol in research evaluation, Evaluation 12 (2): 159–178.

Henry G.T. Mark M.M. (2003), Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions. American Journal of Evaluation, Vol. 24, No. 3, 2003, 293–314.

Hense J.U., Widmer T. (2013), Ein Überblick zum internationalen Stand der Forschung über Evaluation. In: Hense J.U., Rädiker S., Böttcher W., Widmer T. (Hrsg.), Forschung über Evaluation. Bedingungen, Prozesse und Wirkungen, Münster - New York: Waxmann, 251-277.

Hertting,N., Vedung E. (2012), Purposes and criteria in network govern- ance evaluation: How far does standard evaluation vocabulary takes us? Evaluation 18(1), 27-46.

House, E. R. (1987). The evaluation audit. Evaluation Practice, 8(2), 52-56.

Hyvärinen J. (2011), TEKES impact goals, logic model and evaluation of socio-economic effects. Research Evaluation, 20(4), 313-323.

Johnson K., Greenside L.O., Toal S.A., King J.A., Lawrenz F. and Volkov B. (2009), Research on Evaluation Use: A Review of the Empirical Literature from 1986 to 2005. American Journal of Evaluation, 30(3): 377- 410.

Joint Committee On Standards for Education & James R.Sanders (eds.) (1994), The Program Evaluation Standards: How to Assess Evaluations of Educational Programs, Thousand Oaks, CA: SAGE.

Joint Committee on Standards for Educational Evaluation / James R. Sanders (Hg.) (2006), Handbuch der Evaluationsstandards. 3., erweiterte und aktualisierte Auflage, übersetzt und für die deutsche Ausgabe erweitert von Wolfgang Beywl und Thomas Widmer, Wiesbaden: VS Verlag für Sozialwissenschften.

Kirkhart E.K. (2000), Reconceptualizing evaluation use: An integrated theory of influence. New Directions for Evaluation Special Issue: The Expanding Scope of Evaluation Use, Volume 2000, Issue 88, 5–23.

Kuhlmann S. (2015), Jenseits kruder Evidenzmessung: Evaluation als Lernmedium. Keynote auf der 18. Jahrestagung der DeGEval – Gesellschaft für Evaluation e.V. am Deutschen Forschungsinstitut für öffentliche Verwaltung Speyer 16.-18. September 2015.

Kuhlmann S. (2009), Evaluation von Forschungs- und Innovationspolitik in Deutschland - Stand und Perspektiven. In: Widmer T., Beywl W., Fabian C. (Hrsg.), Evaluation - ein systematisches Handbuch, Wiesbaden: VS Verlag für Sozialwissenschften, 283-294.

Kuhlmann S., Meyer-Krahmer F. (1995), Introduction. In: Becher G., Kuhlmann S. (eds) (1995), Evaluation of Technology Programmes in Germany, Doordrecht, 3-32.

Landsteiner G. (2014), Engaging with dynamism and uncertainty of innovation pathways: Towards realistic accounts of policy interventions' contributions to complex systems. Presentation held at the EuSpri 2014 Conference, Manchester 18-20 June 2014. http://www.euspri-manchester2014.com/wp-content/uploads/2014/07/ EuSPRIProgramme_PPTsLink2.pdf

Lehmann L., Balthasar a. (2004), Quality Assessment of External Evaluation Reports Commissioned by the Swiss Agency for Development and Cooperation. A Case of Evaluation Standards Put to Practice. Paper presented at the 6th conference of the European Evaluation Society (EES) in Berlin, September 30 - October 2, 2004.

Leeuw F., Furubo J.E. (2008), Evaluation Systems - What Are They and Why Study Them? Evaluation vol. 14 no. 2, 157-169.

Leeuw F.L., Rist R.C, Sonnichsen R.C. (2000), Can Governments Learn? Comparative Perspectives on Evaluation & Organizational Learning, New Brunswick – London: Transaction Publishers.

Leviton L.L., Hughes E.F.X (1981), Research on the Utilization of Evaluations. Evaluation Review Vol 5, No 4, 525-548.

Lynch, D. C., Greer, A. G., Larson, L. C., Cummings, D. M., Harriett, B. S., Dreyfus, K. S., & Clay, M. C. (2003), Descriptive metaevaluation: Case study of an interdisciplinary curriculum. Evaluation & the Health Professions, 26, 447-461.

Madaus G.F., Scriven M.S., Stufflebeam D.L. (1984), Evaluation Models. Viewpoints on Educational and Human Services Evaluation, 2nd ed., Boston-The Hague-Dordrecht-Lancaster:Kluwer-Nijhoff Publishing.

Mark M.M., Henry G.T. (2004), The Mechanisms and Outcomes of Evaluation Influence, Evaluation Vol 10(1): 35–57.

MIOIR - Manchester Institute of Innovation Research (ed.) (2013), Compendium of Evidence on the Effectiveness of Innovation Policy Intervention, funded by the National Endowment for Science, Technology and the Arts (NESTA), Manchester:MIOIR

MIOIR - Manchester Institute of Innovation Research, Atlantis Consulting, ISI-Fraunhofer, Joanneum Research, Wise Guys Ltd. (2010), INNO-Appraisal. Understanding Evaluation of Innovation Policy in Europe. Final Report February 2010.

Molas-Gallart J., Davies A. (2006), Toward theory-led evaluation - The experience of european science, technology, and innovation policies, Americen Journal of Evaluation Vol. 27 No 1, 64-82.

OECD DAC Network on Development Evaluation (2010), Evaluating Development Co-operation. Summary of key norms and standards, 2nd ed, OECD: Paris.

Ottoson J.,Martinez D. (2010), An Ecological Understanding of Evaluation Use. A Case Study of the Active for Life Evaluation. Robert Wood Johnson Foundation Evaluation Series.

Owen, J. M., Rogers P.J. (1999): Program Evaluation: Forms and Approaches, Thousand Oaks: Jossey Bass.

Patton M.Q. (2010), Developmental Evaluation. Applying Complexity Concepts to Enhance Innovation and Use. New York: Guilford Press.

Patton M.Q. (1997), Utilization-Focused Evaluation. The New Century Text, Thousand Oaks: Jossey Bass.

Pichler R. (2013), Wirkungsorientierung und Evaluierung. Erste Erfahrungen aus der Forschungspolitik nach der Haushaltsrechtsreform in Österreich. Präsentation auf der 16. Jahrestagung der DeGEval: Komplexität und Evaluation, vom 11. bis zum 13. September 2013 an der Ludwig-Maximilians-Universität München.

Pichler R. (2009), Institutionelle Dimensionen von Evaluierung in Österreich. In: Widmer T., Beywl W., Fabian C. (Hrsg.), Evaluation - ein systematisches Handbuch, Wiesbaden: VS Verlag für Sozialwissenschaften, 40-51.

Piciotto R. (2005), The Value of Evaluation Standards: A Comparative Assessment. Journal of MultiDisciplinary Evaluation Number 3, October 2005: 30-59.

Plattform Forschungs- und Technologieevaluierung (Hg.) (2003), Standards der Evaluierung in Österreichs Forschungs- und Technologiepolitik, Wien: fteval.

Plattform Forschungs- und Technologieevaluierung (Hg.) (2005), Standards der Evaluierung in Österreichs Forschungs- und Technologiepolitik, Wien: fteval.

Plattform Forschungs- und Technologieevaluierung (Hg.) (2013), Standards der Evaluierung in Österreichs Forschungs- und Technologiepolitik, Wien: fteval.

Preskill H., Caracelli V. (1997). Current and developing conceptions of use: Evaluation use TIG survey results. Evaluation Practice, 18, 209-225.

Rist R.C., Stame N. (eds.) (2006), From Studies to Streams: Managing Evaluative Systems, New Brunswick, NJ: Transaction Publishers.

Rossi P.H. (2013), My views of evaluation and their origins. In: Alkin M.C. (ed.), Evaluation Roots, S.106-112.

Rossi P.H., Freeman H.E., Lipsey M.W. (1999), Evaluation: A systematic approach, 6th ed., Thousand Oaks, CA: Sage.

Shadish, W.R., Cook T.D., Campbell D.T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton-Mifflin.

Shadish W.R, Cook T.D., Leviton L.C. (1991), Foundations of Program Evaluation: Theories of Practice, Thousand Oaks, CA: SAGE.

Shapira P., Kuhlmann S. (2003), Learning from science and policy evaluation. In: Shapira P., Kuhlmann S. (eds), Learning from science and policy evaluation: experiences from the Unitied States and Europe, Cheltenham – Northampton MA: 1-17.

Shulha, L. M., Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. Evaluation Practice, 18, 195-208.

Stame N. (2013), A European Evaluation Theory Tree. In: Alkin M.C. (ed.) (2012), Evaluation roots (2nd ed.): 355-370.

Stamm, M. (2003), Evaluation und ihre Folgen für die Bildung - eine unterschätzte Herausforderung, Münster: Waxmann.

Strassheim, H., Kettunen, P. (2014): When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms, Evidence & Policy 10(2) S. 259-277.

Stufflebeam D.L. (2001a), The Metaevaluation Imperative, American Journal of Evaluation Vol. 22, No. 2, 2001, pp. 183–209.

Stufflebeam D.L. (2001b), Evaluation Models. New Directions for Evaluation Special Issue: Evaluation Models, Volume 2001, Issue 89: 7–98.

Stufflebeam D.L. (1999), Program Evaluations Metaevaluation Checklist (Based on The Program Evaluation Standards), https://www.wmich.edu/sites/default/files/attachments/u350/2014/program_metaeval_short.pdf

Stufflebeam D.L., Coryn C.L.S. (2014), Evaluation Theory, Models, and Applications, 2nd Edition, San Francisco, CA: Jossey-Bass.

Stufflebeam, D.L., Shinkfield A.J. (2007), Evaluation Theory, Models, and Applications, San Francisco, CA: Jossey-Bass.

Taut S.M., Alkin M.C. (2003), Program Staff Perceptions of Barriers to Evaluation Implementation, American Journal of Evaluation vol. 24 no. 2, 213-226.

Vedung E., Hansen M.B., Kettunen P.T. ( ), Five Political Science Contributions to Evaluation Research. Scandinavian Journal of Public Administration, 16 (3): 3-8.

Weiss C.H. (1998a), Have We Learned Anything New About the Use of Evaluation? American Journal of Evaluation, Vol. 19, No. 1, 21-33.

Weiss C.H. (1998b), Evaluation. Methods for Programs and Policies, 2nd ed., Prentice Hall.

Weiss C.H. (1977), Using social research in policy making. Policy Studies Organisation series 11, Lexington-Toronto: D.C.Heath.

Weiss C.H. (1973), Where politics and evaluation research meet. Evaluation vol 1 no 3, 37-45.

Widmer, T. (2001), Qualitätssicherung in der Evaluation – Instrumente und Verfahren. LeGes – Gesetzgebung & Evaluation 12(2): 9-39.

Widmer T. (1996), Meta-Evaluation: Kriterien zur Bewertung von Evaluationen. Bern: Haupt.

Widmer, T., Beywl, W. (2006), Die Übertragbarkeit der Evaluationsstandards auf unterschiedliche Anwendungsfelder. In: Joint Committee on Standards for Educational Evaluation (Hrsg.): Handbuch der Evaluationsstandards. 3., erweiterte und aktualisierte Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften, 247-261.

Widmer T., Landert C., Bachmann N. (2001), Evaluations-Standards der Schweizerischen Evaluations-gesellschaft (SEVAL-Standards). Bern/Genève: SEVAL.

Widmer, T., Leeuw, F. L. (2009), Die institutionelle Einbettung der Evaluationsfunktion: Deutschland, Österreich und die Schweiz im Vergleich. In: Widmer, Thomas/Beywl,

Wolfgang/Fabian, Carlo (Hrsg.): Evaluation. Ein systematisches Handbuch. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 64-71.

Widmer T., Neuenschwander P. (2004), Embedding Evaluation in the Swidss Federal Administration. Purpose, Institutional Design and Utilization. Education Vol. 10(4): 388-409.

Widmer T., De Rocchi T. (2012), Evaluation - Grundlagen, Ansätze und Anwendungen, Zürich-Chur: Rüegger.

Yarbrough D. et al., 2011, The Program Evaluation Standards. A Guide for Evaluators and Evaluation Users, 3rd Edition, Thousand Oaks, SAGE Publications.

Zinöcker K, Dinges M. (2009), Evaluation von Forschungs- und Technologiepolitik in Österreich. In: Widmer T., Beywl W., Fabian C. (Hrsg.), Evaluation - ein systematisches Handbuch, Wiesbaden: VS Verlag für Sozialwissenschften, 295-304.