

2 Evaluationsforschung

Das folgende Kapitel verdeutlicht die Ausrichtung der vorliegenden Untersuchung, indem die theoretischen Grundlagen zur Evaluationsforschung erarbeitet werden. Im ersten Schritt werden Besonderheiten der Evaluationsforschung herausgearbeitet, Ergebnis- und Prozessevaluation voneinander abgegrenzt und konzeptuelle und methodische Herausforderungen bei der Durchführung von Evaluationsstudien dargestellt. Zur Konkretisierung im Hinblick auf die vorliegende Untersuchung wird im Anschluss daran der Untersuchungsgegenstand, das X-Change for Competence Programm, genauer beschrieben. Es werden die Annahmen und Ziele des Programms expliziert. Im Anschluss daran wird der Evaluationsauftrag dargestellt und es werden die übergreifenden Forschungsfragen formuliert.

2.1 Wissenschaftliche Evaluation von Interventionsprogrammen

Zunächst werden grundlegende Begriffe geklärt und Funktionen von Evaluationsforschung erläutert. Verschiedene Evaluationsvarianten werden voneinander abgegrenzt und es wird die Unterscheidung zwischen Ergebnis- und Prozessevaluation vertieft. Anschließend werden methodische Überlegungen zur Planung und Durchführung von Evaluationsstudien angestellt.

2.1.1 Definition, Funktionen und Varianten von Evaluation

Da der Begriff Evaluation in der Literatur nicht einheitlich verwendet wird, dient die folgende Begriffsklärung einer Eingrenzung im Rahmen der vorliegenden Arbeit. Evaluationsforschung wird anschließend anhand ihrer zentralen Funktionen beschrieben und von anderen Formen angewandter Forschung abgegrenzt. Es folgt eine überblicksartige Darstellung verschiedener Varianten von Evaluation.

2.1.1.1 Definition und Funktionen von Evaluation(sforschung)

Der Begriff Evaluation wird in der Literatur unterschiedlich definiert. Aus diesem Grund wird hier zunächst eine allgemeine Definition voran gestellt, welche den gemeinsamen Kern vieler Definitionen enthält, bevor auf Unterschiede eingegangen wird: „Mit dem Begriff der Evaluation wird die systematische, datenbasierte und kriterienbezogene Bewertung von Programmen, Projekten und einzelnen Angeboten

bezeichnet“ (Heiner, 2007, S. 825).² Evaluation soll den Verdienst, Wert oder die Bedeutung dieser Maßnahmen feststellen (Scriven, 1991) und zu einer rationalen Entscheidungsfindung über die Fortführung oder Umgestaltung dieser beitragen (Heiner, 2007). Evaluation beinhaltet die Erhebung deskriptiver und bewertender Informationen, die notwendig sind, um Entscheidungen über den Nutzen von Interventionen zu treffen und Veränderungs- und Verbesserungspotenziale zu identifizieren (Goldstein & Ford, 2002). Die Bewertung kann das Konzept, die Implementierung und/oder die Wirksamkeit von Interventionen umfassen (Rossi, Lipsey & Freeman, 2004). Häufig werden Begriffe wie Erfolgskontrolle, Wirkungskontrolle, Qualitätskontrolle oder Begleitforschung synonym verwendet oder als Unterformen von Evaluation verstanden (Wottawa & Thierau, 2003, S. 13).

Uneinheitlichkeit herrscht insbesondere bezüglich der Frage, inwieweit Evaluation mit wissenschaftlichen Methoden zu erfolgen hat. Während einige Forscher eher für eine „Kunst des Möglichen“ plädieren, die sich an pragmatischen Anforderungen orientieren muss (Cronbach, 1982, S. 321ff.) und in der Zuordnung der Evaluation zur angewandten Sozialforschung eine Einengung sehen (Scriven, 1991, S. 1), fordert die wissenschaftliche Literatur in der Mehrheit einen klaren Forschungsbezug. Evaluationen müssen demnach denselben methodischen Standards genügen wie andere empirische Forschungsarbeiten (Bortz & Döring, 2006; Wottawa & Thierau, 2003). Mehrere Autoren verwenden daher den Begriff *Evaluationsforschung* und bezeichnen damit den Prozess der systematischen Bewertung von Maßnahmen unter expliziter Verwendung (sozial)wissenschaftlicher Methoden und Techniken (Bortz & Döring, 2006; Gollwitzer & Jäger, 2014; Wottawa & Thierau, 2003). Dementsprechend wird Evaluationsforschung als ein anwendungsorientierter Teilbereich der empirischen Forschung verstanden, der durch seine speziellen Fragestellungen – die Bewertung von Interventionen – gekennzeichnet ist (Bortz & Döring, 2006). Dieses Verständnis von Evaluation liegt auch der vorliegenden Arbeit zugrunde.

In Anlehnung an Stockmann (2000) differenzieren Bortz und Döring (2006, S. 97) fünf zentrale Ziele oder Funktionen von Evaluation: (1) Evaluationsforschung hat eine *Erkenntnisfunktion*, das heißt sie trägt zur Entstehung wissenschaftlicher Erkenntnisse über Charakteristika, Prozesse und Wirkungen von Interventionen bei. (2) Evaluationsforschung hat eine *Optimierungsfunktion*: Sie identifiziert Stärken und

² Sehr breite Definitionen von Evaluation beziehen als Evaluationsgegenstand auch Produkte, Umweltfaktoren oder Personen mit ein. Diese Arbeit fokussiert jedoch auf Programmevaluati on und grenzt den Evaluationsgegenstand daher auf soziale Interventionen ein.

Schwächen einer Intervention im Hinblick auf deren angestrebte Ziele und leitet Empfehlungen daraus ab. (3) Evaluationsforschung hat eine *Kontrollfunktion*, indem sie überprüft, ob die Implementierung der Planung entspricht, inwieweit Interventionsziele erreicht werden und ob nicht intendierte positive wie negative „Nebenwirkungen“ auftreten. (4) Evaluationsforschung hat eine *Entscheidungsfunktion*, das heißt die durch sie erzeugten Erkenntnisse werden dazu genutzt zu entscheiden, ob und welche Interventionen implementiert, fortgeführt, modifiziert oder eingestellt werden. (5) Schließlich hat Evaluationsforschung auch eine *Legitimationsfunktion*. Schon die Durchführung einer Evaluation soll häufig dazu beitragen, die Intervention gegenüber relevanten Stakeholdern zu legitimieren und aufzuwerten. Die aus der Evaluation resultierenden Befunde dienen dazu, Rechenschaft über die Verwendung von Mitteln abzulegen oder Entscheider, insbesondere Geldgeber, zu beeinflussen.

Diese Zusammenstellung von Funktionen der Evaluation verdeutlicht, dass der wissenschaftliche Erkenntnisgewinn nur eines von mehreren Zielen von Evaluationsforschung ist. Daher ist Evaluationsforschung stärker als andere Formen angewandter Forschung zweckgebunden. Bortz und Döring (2006, S. 99) grenzen Evaluationsforschung dementsprechend über die *Gebundenheit vs. Offenheit der Forschungsziele* von anderen Formen angewandter Forschung ab. Evaluationsforschung orientiert sich primär an den vom Auftraggeber gestellten Fragen; die Forschungsbemühungen sind darauf ausgerichtet, die Wirksamkeit einer konkreten Maßnahme und/oder mögliche Einflussfaktoren auf die Wirksamkeit zu erfassen (Beywl, 2002; Bortz & Döring, 2006). Dagegen sind die Forschungsziele der angewandten Forschung offener. Zum Beispiel kann sich die Trainingsforschung mit generellen Einflussfaktoren auf die Trainingsqualität oder Bedingungen für den Transfer des Gelernten in den Arbeitsalltag beschäftigen. Die angewandte sozialpsychologische Forschung kann sich mit den im Intergruppenkontakt relevanten kognitiven und affektiven Prozessen beschäftigen und daraus neue Forschungsfragen entwickeln. In der Evaluationsforschung hingegen wird der Stand der Forschung nicht primär dazu genutzt, Forschungsfragen abzuleiten, sondern dazu, den Evaluationsgegenstand in den aktuellen Wissensstand einzuordnen, dem Evaluator Hintergrundwissen zu liefern, um Annahmen und Hypothesen begründet zu formulieren und sich mit Designs und Methoden vertraut zu machen, die sich in anderen Untersuchungen mit ähnlicher Thematik bewährt haben (Bortz & Döring, 2006, S. 109). Vor diesem Hintergrund erfolgt auch die Fokussierung der vorliegenden Arbeit: Die aus der Fülle aller möglichen Forschungsfragen herausgegriffenen Schwerpunkte

orientieren sich an den Programmzielen des X-Change for Competence Programms. Der Ausblick am Ende der Arbeit verweist auf weiterführende Forschungsfragen, die sich aus der vorliegenden Evaluationsstudie ergeben und teilweise auch mit den erhobenen Daten eingehender untersucht werden könnten.

2.1.1.2 Evaluationsvarianten

Zur Beschreibung verschiedener Varianten von Evaluation findet sich in der Literatur ebenfalls eine Begriffsfülle, die teils ähnliche, teils unterschiedliche Konzepte bezeichnet. Im Folgenden werden die etabliertesten und für diese Arbeit relevanten Begriffsunterscheidungen eingeführt. Die Unterscheidung erfolgt entlang zweier Dimensionen: des Zeitpunktes und des Gegenstandes der Evaluation (Kauffeld, 2010, S. 111). Auf diese Weise werden die Begriffspaare *formative* vs. *summative* Evaluation und *prozess-* vs. *ergebnisbezogene* Evaluation voneinander abgegrenzt (Scriven, 1991). Während die Unterscheidung *formative* vs. *summative* Evaluation sich vor allem auf den Evaluationszeitpunkt bezieht, geht es bei der Unterscheidung zwischen *prozess-* und *ergebnisbezogener* Evaluation um den Evaluationsgegenstand.

Die auf Scriven (1991) zurückgehende Unterscheidung zwischen *summativer* und *formativer Evaluation* bezieht sich vor allem auf den Zeitpunkt der Evaluation:

Summative Evaluation erfolgt nach dem Abschluss einer Intervention. Sie wird daher auch retrospektive Evaluation genannt (Bortz & Döring, 2006) und bezeichnet die post hoc stattfindende Bilanzierung hinsichtlich einer Intervention (Brown & Gerhardt, 2002; Wottawa & Thierau, 2003). Diese Bilanzierung kann sich sowohl auf die Durchführung (Nähe zum ursprünglichen Implementierungskonzept, Qualität der Durchführung, Abbrecherquoten etc.) als auch auf die Wirksamkeit der Intervention (z.B. bei Programmteilnehmern erzielte Veränderungen) beziehen. *Summative* Evaluation liefert nach Abschluss der Intervention eine Rechtfertigung für den Einsatz von Mitteln und kann als Entscheidungsgrundlage für die Fortführung, Ausweitung oder Beendigung von Interventionen dienen (Kauffeld, 2010). Daher adressiert sie über die unmittelbar am Programm beteiligten Personen hinaus häufig auch Geldgeber oder andere einflussreiche Entscheider.

Formative Evaluation hingegen findet während der Durchführung einer Intervention statt und zielt auf ihre kontinuierliche Verbesserung (Scriven, 1991). Sie wird daher auch begleitende Evaluation oder Begleitforschung genannt (Bortz & Döring, 2006; Kauffeld, 2010; Scriven, 1991) und wird vor allem bei der Entwicklung und Implementierung neuer Programme eingesetzt. Die *formative* Evaluation liefert

regelmäßig Zwischenergebnisse und dient dazu, die Implementierung (z.B. Durchführungsqualität, Akzeptanz) des Programms zu erfassen und seine unmittelbaren Wirkungen fortlaufend zu kontrollieren. Sie kann dazu beitragen, dass die ursprüngliche Programmkonzeption verändert und das Programm kontinuierlich weiterentwickelt wird (Bortz & Döring, 2006; Craig & Hannum, 2007; Kauffeld, 2010). Formative Evaluation adressiert in der Regel die an der Programmdurchführung unmittelbar beteiligten Personen.

Scriven bringt die Unterscheidung zwischen summativer und formativer Evaluation unter Rückgriff auf Robert Stake wie folgt auf den Punkt: „The distinction between formative and summative evaluation has been well summed up by Bob Stake: ‘When the cook tastes the soup, that’s formative; when the guests taste the soup, that’s summative.’” (Scriven, 1991, S. 169).

In der Literatur wird summative Evaluation teilweise mit Ergebnisevaluation und formative Evaluation mit Prozessevaluation gleichgesetzt (z.B. Bortz & Döring, 2006; Craig & Hannum, 2007; Ely, Boyce, Nelson, Zaccaro, Hernandez-Broome & Whyman, 2010). Allerdings verschwimmt dabei die Differenzierung zwischen dem Zeitpunkt der Evaluation und dem Gegenstand der Evaluation.

Die Unterscheidung zwischen *Ergebnis-* und *Prozessevaluation* bezieht sich, unabhängig vom Zeitpunkt, auf den Gegenstand der Evaluation (Kauffeld, 2010; Scriven, 1991). Ergebnisbezogene Evaluation überprüft die Wirkung oder Wirksamkeit von Interventionen. Alternativ finden sich hierfür auch die Begriffe Erfolg, Effekt, Effektivität oder Nutzen in der Literatur, im englischsprachigen Raum Impact, Outcome und Effectiveness. Ergebnisbezogene Evaluation prüft, inwieweit intendierte Veränderungen und angestrebte Ziele tatsächlich erreicht wurden.

Die Bewertung der Wirkung einer Intervention reicht jedoch nicht aus, um zu beurteilen, welche Bestandteile der Maßnahme (z.B. Inhalte, Design, Teilnehmer, Rahmenbedingungen) zum Ergebnis beigetragen oder es verhindert haben und welche Stellschrauben im Prozess existieren, um die Intervention zu verbessern. Prozessbezogene Evaluation zielt darauf ab, Einflussfaktoren auf das Ergebnis einer Intervention zu identifizieren. „Katalysatoren und Barrieren“ für die Programmwirksamkeit sollen überprüft werden (Kauffeld, 2010, S. 111). Dazu untersucht die prozessbezogene Evaluation Charakteristika der Programmgestaltung, der Teilnehmer und des Kontexts, die sich förderlich oder hinderlich auf das Ergebnis auswirken. Ziel ist es, Ursachen für erzielte und nicht erzielte Ergebnisse zu finden, um

diese bei künftigen Programmplanungen zu berücksichtigen und darüber die Wahrscheinlichkeit zu erhöhen, dass die angestrebten Ziele zukünftig erreicht werden. Eine Unterscheidung, die der Abgrenzung von prozess- vs. ergebnisbezogener Evaluation verwandt ist, ist die zwischen Input- und Implementations-Evaluation einerseits und Output- und Outcome-Evaluation andererseits (Scriven, 1991).

Eine summative Evaluation zu Programmende muss also nicht notwendigerweise nur das Ergebnis bewerten. Sie kann auch bilanzierend den Prozess analysieren und zusammenfassend förderliche und hinderliche Einflussfaktoren ermitteln. Im weiteren Verlauf der Arbeit bezeichnet daher formativ vs. summativ den Zeitpunkt der Evaluation. Wenn von Prozess- vs. Ergebnisevaluation die Rede ist, ist der Evaluationsgegenstand gemeint.

Die vorliegende Evaluationsstudie begleitete das Programm von Beginn an und wurde als formative und summative Evaluation durchgeführt. Vorgehen und Beitrag der formativen Evaluation werden in Abschnitt 2.2.2.1 kurz zusammengefasst. Die vorliegende Arbeit fokussiert auf die Darstellung der summativen Evaluation. Diese umfasste sowohl eine ergebnisbezogene Evaluation, orientiert an den Programmzielen, als auch eine prozessbezogene Evaluation, orientiert an den Grundannahmen des Programms. Daher wird im nächsten Abschnitt genauer auf Ebenen der Ergebnisevaluation und Facetten der Prozessevaluation eingegangen.

2.1.2 Ergebnis- und Prozessevaluation

Im Hinblick auf die Ergebnisevaluation können mehrere Ebenen unterschieden werden. Zur Prozessevaluation werden die in der Literatur diskutierten Faktoren zusammenfassend dargestellt. Beide werden in Kapitel 3 im theoretischen Teil konkret auf Führungsentwicklung sowie auf Intergruppenkontakt bezogen und vertieft.

2.1.2.1 Ergebnisevaluation: Differenzierung von Wirkebenen

Die Ergebnisevaluation beschreibt und bewertet die Wirkung von Interventionen. Welche Art von Wirkungen durch eine Intervention erwartet wird, hängt unmittelbar mit der Art der Intervention und deren Zielen zusammen. Daher wird hier in einem ersten Schritt dargestellt, wie Ergebnisse von Interventionsmaßnahmen im Allgemeinen systematisiert werden können. Grundsätzlich besteht Einigkeit, dass im Rahmen von Evaluationsforschung die Bewertung der Wirkung über bloße Zufriedenheitsbewertungen hinausgehen sollte (u.a. Ely et al., 2010; Kirkpatrick, 1967, 1994).

Eine erste Differenzierung der Wirkebenen bezieht sich auf die *Reichweite der angestrebten Wirkungen*. Martineau, Hannum und Reinelt (2007) spannen einen weiten Bogen, der von der individuellen bis zur gesellschaftlichen Ebene reicht: Sie unterscheiden Wirkungen auf der Ebene von Individuen, Gruppen und Teams, Organisationen, Gemeinschaften und Netzwerken, Sektoren (z.B. privatwirtschaftlicher oder Nonprofit-Sektor) und Gesellschaften als Ganzes. Aufgrund des doppelten inhaltlichen Fokus des X-Change for Competence Programms als Führungs-entwicklungs- und Begegnungsprogramm bietet sich für eine Konkretisierung möglicher Wirkebenen über die generelle Reichweite hinaus der Bezug zu diesen beiden Forschungsfeldern an.

In der Evaluationsforschung zu Trainings und anderen berufsbezogenen Entwicklungsmaßnahmen hat sich insbesondere die Taxonomie nach Kirkpatrick (1967, 1994) etabliert (Alliger, Tannenbaum, Bennet, Traver & Shotland, 1997; Goldstein & Ford, 2002; Kraiger, Ford & Salas, 1993; Salas & Cannon-Bowers, 2001). Diese fokussiert auf die ersten drei von Martineau et al. (2007) benannten Ebenen, also vom Individuum bis zur Organisation und füllt diese inhaltlich stärker. Kirkpatrick (1994) unterscheidet die vier Ebenen Reaktions-, Lern-, Verhaltens- und Ergebnisebene. Die *Reaktionsebene* (1) erfasst das Ausmaß, in dem Teilnehmer positiv auf die Trainings- bzw. Personalentwicklungsmaßnahme reagieren. Auf der *Lernebene* (2) wird geprüft, in welchem Umfang die Teilnehmer angestrebtes Wissen, Einstellungen, Kompetenzen, Fertigkeiten etc. erworben haben. Auf der *Verhaltensebene* (3) wird der Transfer vom Lern- ins Arbeitsfeld erfasst, d.h. in welchem Ausmaß das im Training Erlernte im beruflichen Alltag angewandt und in Verhalten umgesetzt wird. Die *Ebene der Resultate* (4) bezieht sich auf den Einfluss der Trainingsmaßnahme auf organisationale Ziele, wie z.B. Produktivität, Return on Investment, Teamleistung, Mitarbeiterbindung oder Kundenzufriedenheit. Auf der Grundlage dieser Vier-Ebenen-Taxonomie haben verschiedene Autoren weitere Differenzierungen vorgenommen. So kann die Reaktionsebene weiter unterteilt werden in Zufriedenheit und wahrgenommenen Nutzen (*affective reactions* vs. *utility judgments*; Alliger, Tannenbaum, Bennet, Traver & Shotland, 1997; *satisfaction* vs. *utility*; Grohmann & Kauffeld, 2013). Kraiger et al. (1993) haben die Lernebene weiter ausdifferenziert und unterscheiden dabei unter anderem kognitives Lernen, d.h. Veränderungen in Wissen und kognitiven Strategien (z.B. Selbstreflexivität, Selbstkenntnis, kognitive Flexibilität) von affektivem Lernen,

also Veränderungen in Einstellungen und Motivation (z.B. Selbstwirksamkeitsüberzeugung, berufsbezogene Einstellungen wie organisationales Commitment oder Arbeitszufriedenheit).³

Im Gegensatz zur Trainingsforschung hat sich im Bereich des Intergruppenkontakts und der Evaluation von Encounter-Programmen keine vergleichbar einheitliche Systematisierung von Wirkebenen etabliert. Aber auch hier lassen sich verschiedene Reichweiten – Veränderung bei individuellen Teilnehmern, Veränderungen zwischen den am Programm teilnehmenden Subgruppen, Veränderungen in generalisierten Einstellungen und Verhalten gegenüber der Outgroup als ganzer – unterscheiden (Pettigrew & Tropp, 2011).

Neben der Reichweite der Wirkung sind die *Art der angestrebten Veränderung* (kognitive, affektive oder Verhaltensänderungen, z.B. Kraiger et al., 1993; Pettigrew & Tropp, 2011) und der *Zeithorizont* der Wirkung (kurz-, mittel-, oder langfristig; z.B. Martineau et al., 2007; Greif, 2012; König & Vollmer, 2003) weitere Differenzierungsmöglichkeiten im Hinblick auf die Ergebnisevaluation.

2.1.2.2 Prozessevaluation: Differenzierung von Einflussfaktoren auf die Wirksamkeit

Die Bewertung der Wirkungen einer Intervention sagt noch nichts darüber aus, wie diese zustande gekommen sind. Bei der Prozessevaluation geht es darum zu ermitteln, welche Einflussfaktoren (z.B. Inhalte und Design der Intervention, Teilnehmercharakteristika, Rahmenbedingungen) zu den Wirkungen beigetragen oder sie verhindert haben und somit förderliche und hinderliche Einflussfaktoren auf die Programmwirksamkeit zu identifizieren (Goldstein & Ford, 2002).

Eine Differenzierungsmöglichkeit im Hinblick auf Einflussfaktoren ist die Unterscheidung zwischen *Inputfaktoren* und *Prozessfaktoren im engeren Sinn*. Im von Scriven (1991) zitierten Bild bedeutet dies die Unterscheidung zwischen den Zutaten für ein Gericht (Inputfaktoren) und dem Prozess des Zubereitens durch den Koch und weitere Beteiligte (Prozessfaktoren). Zusätzlich können *Kontextfaktoren* (organisationale und gesellschaftliche Rahmenbedingungen) die Programmwirksamkeit beeinflussen.

Aufgrund des doppelten Fokus des X-Change for Competence Programms wird zur Konkretisierung bzgl. der Systematisierung von Einflussfaktoren auf die

³ Zur Ausdifferenzierung der Verhaltensebene sind zahlreiche Taxonomien entwickelt worden. Abschnitt 3.1.2.1 dieser Arbeit geht genauer auf Taxonomien effektiven Führungsverhaltens ein.

Programmwirksamkeit auf die Literatur sowohl zur Trainingsevaluation als auch zum Intergruppenkontakt Bezug genommen.

In der Trainings- und Führungsentwicklungsliteratur hat sich die Unterteilung in drei Gruppen von Einflussfaktoren etabliert – in Anlehnung an das Rahmenmodell des Transferprozesses nach Baldwin und Ford (1988) (Bass & R. Bass, 2008; Burke & Hutchins, 2007). Es werden Charakteristika der Trainingsteilnehmer und der Teilnehmerzusammensetzung (Inputfaktoren), der Trainings- oder Programmgestaltung (Prozessfaktoren im engeren Sinn) sowie des organisationalen Kontexts herangezogen. Merkmale der Teilnehmer umfassen u.a. Persönlichkeitsfaktoren, Fähigkeiten und Fertigkeiten und Motivation; Merkmale des Trainings beinhalten z.B. Lernprinzipien und die Passung der Trainingsinhalte und -methoden zu den beruflichen Anforderungen; unter Merkmale der Arbeitsumgebung fallen etwa die Unterstützung durch Vorgesetzte und Kollegen sowie Anwendungsmöglichkeiten.

In der Literatur zum Intergruppenkontakt orientieren sich berücksichtigte Einflussfaktoren eng an der jeweils zugrundeliegenden Theorie (u.a. Allport, 1954; Pettigrew, 1998), auf die in Abschnitt 3.2.1 genauer eingegangen wird. Im Wesentlichen können die Einflussfaktoren ebenfalls den drei Gruppen Inputfaktoren (Teilnehmercharakteristika wie z.B. Status, Persönlichkeit, Vorerfahrungen mit der Outgroup), Prozessfaktoren (z.B. soziale Kategorisierungsprozesse während des Intergruppenkontakts) und Kontextfaktoren (institutionelle und gesellschaftliche Rahmenbedingungen des Intergruppenkontakts) unterteilt werden.

2.1.3 Methodische Überlegungen zur Gestaltung von Evaluationsstudien

Nach der Beschreibung verschiedener Evaluationsvarianten wird im Folgenden genauer auf methodische Überlegungen zur Gestaltung von Evaluationsstudien eingegangen. Nach Überlegungen zur Ziel- und Auftragsklärung werden verschiedene Designvarianten dargestellt. Schließlich werden Methoden der Datenerhebung in den Blick genommen.

2.1.3.1 Ziel- und Auftragsklärung

Die Aussagekraft einer Evaluationsstudie steht und fällt mit der Auswahl der erfassten Kriterien (Goldstein & Ford, 2002). Insofern Evaluationsforschung immer auch in dem oben beschriebenen Verwertungskontext steht, muss die Auswahl der Evaluationskriterien sich an den Zielen und Anliegen der Stakeholder orientieren (Ely et al., 2010). Allerdings gehen in der Praxis Interventionsprogramme häufig mit recht weit gesteckten

oder vage formulierten Zielen einher, die nur schwer messbar sind (Bortz & Döring, 2006; Gutiérrez & Tasse, 2007, S. 48ff.). Darüber hinaus besteht oft Unklarheit und Uneinigkeit auf Seiten der Stakeholder darüber, welche Ziele genau im Vordergrund stehen sollen, woran Veränderungen festgemacht werden können und wie und warum spezifische Programmaktivitäten zur Zielerreichung beitragen. Bortz und Döring (2006, S. 117) betonen daher: „Eine gelungene Operationalisierung setzt eine sorgfältige Explikation der Ziele voraus, die mit der Maßnahme angestrebt werden.“ Erfolgreiche Evaluationsstudien sind nicht ohne eine sorgfältige Zielexplication möglich. Dies bedeutet auch, dass der Evaluator den Prozess der Zielexplication ggf. aktiv unterstützen muss (Bortz & Döring, 2006, S. 122). „Part of an evaluator’s role is to help ensure that stakeholders have a shared understanding of the program’s objectives and outcomes” (Martineau et al., 2007, S. 24). Idealerweise unterstützt die Zielklärung so die Zielerreichung des Programms dadurch, dass klar definierte Ziele bewusster verfolgt werden können. Für die Evaluationsstudie ist die Konkretisierung der Ziele notwendig, um die Forschungsfragen präzise formulieren zu können.

Verschiedene Autoren empfehlen, Ziele und dem Programm zugrundeliegende Annahmen gemeinsam mit den Stakeholdern zu erarbeiten. Dadurch soll die Programmlogik (Martineau et al., 2007), die Theory of Change, das Pfadmodell (Gutiérrez & Tasse, 2007) bzw. das Model of Change (Riemer & Bickman, 2011) des Programms herausgearbeitet werden. Gemeinsamer Kern dieser unterschiedlichen Begriffe ist, dass die Annahmen der Programmverantwortlichen expliziert und in eine nachvollziehbare Struktur gebracht werden. So kann zum Beispiel zwischen Annahmen über die Ausgangssituation, kurz- und langfristigen Programmzielen und Interventionsstrategien unterschieden und diese zueinander in Bezug gesetzt werden. Gutiérrez und Tasse (2007, S. 48 ff.) beschreiben diesen Prozess des *Pathway Mappings* genauer: Stakeholder des Programms explizieren ihre Theorien und Annahmen über das Programm und dessen Verlauf. Daraus wird ein Ablaufdiagramm erstellt, das Kernpunkte einer Situationsanalyse zusammenfasst, aus der sich der Bedarf für das Programm ableitet. In diesem Ablaufdiagramm werden kurz- und langfristige Programmziele ausformuliert und verdeutlicht, wie die Aktivitäten des Programms im Laufe der Zeit spezifische Ergebnisse herbeiführen sollen. Auf dieser Grundlage können dann die Evaluationsfragen und -kriterien konkretisiert werden. Das Model of Change des X-Change for Competence Programms, das im Rahmen der vorliegenden Evaluationsstudie erstellt wurde und in Abschnitt 2.2.1.2 beschrieben wird, folgt diesem von Gutiérrez und Tasse (2007) empfohlenen Vorgehen.

2.1.3.2 *Evaluationsdesigns*

Auch die Untersuchungsplanung und das Design von Evaluationsstudien leiten sich aus der Art der Evaluation und den Fragen, die durch die Evaluation beantwortet werden sollen, ab (Craig & Hannum, 2007). Bortz und Döring (2006) unterscheiden grundsätzlich zwischen *hypothesenprüfenden* und *explorativ angelegten Evaluationsstudien*. Hypothesenprüfende Evaluationsstudien überprüfen im Vorfeld getroffene Annahmen über die Wirksamkeit (oder auch Einflussfaktoren auf die Wirksamkeit) von Interventionen. Explorative Evaluationsstudien sind erkundender Natur und dienen dazu, Unerwartetem auf die Spur zu kommen und Veränderungsprozesse nachzuvollziehen.

Ergebnisevaluationen enthalten immer ein hypothesenprüfendes Element: Sie dienen dazu, die Hypothese zu überprüfen, dass die Intervention wirkt bzw. dass sie so wirkt, wie man es erwartet hat (Bortz & Döring, 2006). Dementsprechend ist das Grundbestreben von Evaluationsdesigns, die auf Wirksamkeitsüberprüfung angelegt sind, Störquellen – und somit alternative Erklärungen für die gefundenen Ergebnisse – möglichst auszuschließen (Craig & Hannum, 2007; Wottawa & Thierau, 2003). Das Evaluationsdesign soll eine hohe interne Validität gewährleisten, d.h. gefundene Veränderungen sollen kausal auf die durchgeführte Intervention zurückgeführt werden können. Es geht darum, die „Netto-Wirkung“ einer Intervention zu ermitteln (Rossi, Lipsey & Freeman, 2004). Neben einem hypothesenprüfenden Vorgehen ist für Evaluationen aber auch ein explorativer Zugang bedeutsam: In der ergebnisbezogenen Evaluation, um auch solche positiven wie negativen Wirkungen erfassen zu können, die nicht erwartet wurden und dementsprechend auch nicht Bestandteil der Wirkhypothese waren. Für Fragestellungen der Prozessevaluation ist ein hypothesenprüfendes Vorgehen oft nicht ausreichend, sondern sind explorative Evaluationszugänge nötig, wenn komplexe Veränderungsprozesse interessieren (Bortz & Döring, 2006, S. 109). Die vorliegende Studie integriert einen hypothesenprüfenden mit einem explorativen Zugang.

Daher werden im Folgenden Anforderungen an ein hypothesenprüfendes Vorgehen verdeutlicht. Grundsätzlich steht Evaluationsforschung vor der Herausforderung, ein Design auszuwählen, das unter den gegebenen Rahmenbedingungen umsetzbar ist (Praktikabilität) und gleichzeitig Störquellen möglichst gering hält, welche den Kausalbezug zwischen Programm und Wirkung (d.h. die interne Validität)

einschränken.⁴ Daher werden zunächst verschiedene Störfaktoren und anschließend verschiedene Designvarianten zur Überprüfung der Wirksamkeit von Interventionen erläutert.

Störquellen, welche die interne Validität potenziell gefährden, können in programmexterne Störfaktoren, Selektionseffekte und Störfaktoren, die durch das Forschungsdesign bedingt sind, unterteilt werden (Craig & Hannum, 2007; Rossi, Lipsey & Freeman, 2004). *Programmexterne Störfaktoren* umfassen alle Einflussfaktoren, die die Zielvariable(n) zusätzlich zur Intervention beeinflussen. Darunter fallen unter anderem Reifungsprozesse (z.B. zunehmende Berufserfahrung, die unabhängig von der Intervention zu einem Kompetenzzuwachs führt) und exogener Wandel, d.h. Veränderungen im unmittelbaren Umfeld (z.B. eine organisationale Umstrukturierung, die zu einem Verlust an führungsbezogener Selbstwirksamkeit führt) oder gesellschaftliche Ereignisse (z.B. ein eskalierender Konflikt, der die wechselseitigen Einstellungen zweier Gruppen negativ beeinflusst). Auch *Selektionseffekte*, die zu systematischen Unterschieden zwischen Teilnehmern und Nicht-Teilnehmern führen, sind problematisch. So kann sich zum Beispiel eine Überschätzung der Programmwirksamkeit ergeben, wenn sich nur besonders veränderungsmotivierte Teilnehmer für eine Programmteilnahme entscheiden. Störfaktoren, die *durch das Forschungsdesign bedingt* sind, umfassen u.a. Messprobleme (Lerneffekte, mangelnde Reliabilität und Validität), systematische Verzerrungen durch fehlende Werte, Regression zur Mitte oder den Hawthorne-Effekt⁵ (Craig & Hannum, 2007; Rossi, Lipsey & Freeman, 2004; Wottawa & Thierau, 2003). Verschiedene Designvarianten bieten unterschiedlich gute Möglichkeiten, diese Störfaktoren zu kontrollieren. Experimentelle Designs mit randomisierter Gruppenzuweisung, welche die beste Absicherung gegen viele der genannten Störfaktoren bieten, sind in der Evaluationsforschung meist nicht möglich. Daher überwiegen quasi-experimentelle Designs, bei denen verschiedene „natürliche“ Gruppen miteinander verglichen werden und Eingruppen-Designs mit einem oder mehreren Messzeitpunkten (Bortz & Döring, 2006; Craig & Hannum, 2007; Wottawa

⁴ Die Bedeutung der externen Validität, also des Ausmaßes, in dem Schlüsse über das konkret evaluierte Programm hinaus, z.B. für andere Teilnehmer und spätere Zeitpunkte gültig sein sollen, kann je nach Evaluationsauftrag variieren. Manchmal wird eine Generalisierung über das spezielle Programm hinaus zumindest von Auftraggeberseite aus überhaupt nicht angestrebt, häufig jedoch, gerade wenn es sich um Pilotprojekte handelt, ausdrücklich gefordert (Craig & Hannum, 2007, S.37).

⁵ Der Hawthorne-Effekt bezeichnet das Phänomen, dass Untersuchungsteilnehmer allein aufgrund der Tatsache, dass sie untersucht oder beobachtet werden, ihr Verhalten ändern.

& Thierau, 2003). Kontrollgruppen-Designs bieten gegenüber Eingruppen-Designs den Vorteil, dass zwischen der Interventionswirkung und Reifung unterschieden werden kann. Bei Eingruppen-Designs kann diese Konfundierung nicht aufgelöst werden. Um die „Netto-Wirkung“ einer Intervention zu ermitteln, wird daher der Einsatz von Kontrollgruppendesigns wo immer möglich empfohlen (Bortz & Döring, 2006; Craig & Hannum, 2007; Hannum & Martineau, 2008; Wottawa & Thierau, 2003). Allerdings liegt hier genau eine Problematik vieler Evaluationsprojekte: Geeignete Vergleichsgruppen sind häufig nicht verfügbar und Kontrollgruppendesigns in der Praxis oft nicht durchsetzbar (Frese, Beigel & Schoenborn, 2003; Hannum & Craig, 2010). Frese, Beigel und Schoenborn (2003) schlagen daher das *Nonequivalent Dependent Variable Design* (Cook & Campbell, 1979), das auch als interne Referenzstrategie bezeichnet wird (IRS, Haccoun & Hamtiaux, 1994), als Alternative zum Kontrollgruppendesign vor und wenden es auf die Evaluation von Führungstrainings an. In diesem Eingruppen-Prä-Post-Design werden trainierte Verhaltensweisen mit Verhaltensweisen verglichen, die nicht trainiert wurden. Innerhalb einer Treatmentgruppe werden also zwei abhängige Variablen miteinander verglichen; für eine dieser abhängigen Variablen wird ein Einfluss durch das Training erwartet (Zielvariable), für die andere Variable (Kontrollvariable) hingegen nicht. Damit von einem Trainingseffekt gesprochen werden kann, müssen intendierte Veränderungen in der Zielvariablen die Veränderungen in der Kontrollvariablen übersteigen. Durch diese Strategie können Störquellen (Lern- und Entwicklungseffekte, exogener Wandel) im Gegensatz zu einem einfachen Prä-Post-Vergleich kontrolliert werden. Im Gegensatz zu einer Wartelisten-Kontrollgruppe wirkt der Hawthorne-Effekt auf Ziel- und Kontrollvariablen gleichermaßen und kann so ebenfalls kontrolliert werden. Darüber hinaus ist im Gegensatz zu quasiexperimentellen Kontrollgruppendesigns die Gefahr ausgeschlossen, dass eine nicht erhobene Variable, die zwischen Treatment- und Kontrollgruppe differenziert, konfundierend wirkt (z.B. wenn Treatment- und Kontrollgruppe aus unterschiedlichen Organisationseinheiten stammen und aufgrund dessen von einer Umstrukturierung unterschiedlich stark betroffen sind). Nachteile der internalen Referenzstrategie sind jedoch, dass der Forscher selbst definieren kann, wie schwierig oder leicht es ist, signifikante Unterschiede in der Veränderung von Ziel- und Kontrollvariablen zu erhalten, je nachdem, wie ähnlich oder unähnlich er die beiden Variablen wählt. Darüber hinaus lässt das Design keine Interpretation bei Nicht-Veränderung zu, sondern ist vollständig

abhängig von entgegengesetzten Mustern der Veränderung und Nicht-Veränderung (Cook & Campbell, 1979).

In der Regel erfordern Evaluationsdesigns einen Kompromiss zwischen Praktikabilität und dem bestmöglichen Ausschluss konfundierender Variablen. Über die Designwahl hinaus empfiehlt es sich daher, mögliche konfundierende Variablen statistisch zu kontrollieren (Rossi, Lipsey & Freeman, 2004). Mögliche fördernde und hemmende Einflussfaktoren auf die Wirksamkeit der Intervention sollten in den Untersuchungsplan aufgenommen werden, um zumindest korrelativ einen Beitrag zur Aufklärung des Bedingungsgefüges zwischen Intervention, Kontext und Wirkung leisten zu können. Auch der Einsatz qualitativer, offener Verfahren der Datenerhebung bietet sich an, um auf mögliche Störquellen aufmerksam zu werden. (Wottawa & Thierau, 2003, S. 129).

2.1.3.3 Auswahl von Datenquellen und Methoden der Datenerhebung

Nach der Klärung der Evaluationsziele, -kriterien und grundlegender Designfragen müssen Datenquellen und Erhebungsmethoden ausgewählt werden. Zunächst wird dargestellt, warum es in der Evaluationsforschung in der Regel notwendig ist, multiperspektivisch und multimethodal vorzugehen. Im Anschluss daran werden Anforderungen an die Instrumentauswahl und Schwierigkeiten bei der Operationalisierung und Messung der angestrebten Veränderungen umrissen. Der methodische Zugang, der in der vorliegenden Studie gewählt wurde, wird im Methodenteil (Kapitel 4) genauer begründet.

Die Datenerhebung soll auf der Grundlage des aktuellen Forschungsstandes die bestmögliche Beantwortung der grundlegenden Evaluationsfragen sicherstellen. Grundsätzlich wird in der Evaluationsforschung ein multiperspektivischer und multimethodaler Ansatz empfohlen (Bortz & Döring, 2006; Hannum & Martineau, 2008).

Die *Multiperspektivität* bezieht sich darauf, dass mehrere Datenquellen herangezogen werden sollten, um einer häufigen Kritik an Evaluationsstudien, nämlich dass diese einseitig auf Teilnehmereinschätzungen zurückgreifen, zu begegnen. Neben der Bewertung durch die Teilnehmer sollten auch die Perspektiven anderer Stakeholdergruppen berücksichtigt werden (Ely et al., 2010; Hannum & Martineau, 2008; Russon & Reinelt, 2004). Dabei muss bei der Wahl der Datenquelle bedacht werden, worüber unterschiedliche Stakeholder überhaupt Auskunft geben können. Während zum Beispiel die Reaktions- und Lernebene in der Taxonomie von Kirkpatrick (1994) mittels Selbstauskunft der Teilnehmer erhoben werden kann, erfordert die Einschätzung von

Verhaltensänderungen Fremdeinschätzungen. Für die Evaluation von Führungsentwicklungsmaßnahmen wird hier insbesondere der Einsatz von 360°-Assessments empfohlen (u.a. Smither, London, Flautt, Vargas & Kucine, 2003).

Die Forderung nach einem *multimethodalen Ansatz* wird damit begründet, dass mittels einer Methodentriangulation die Stärken unterschiedlicher Verfahren komplementär genutzt werden können. Häufig wird dabei für eine Kombination aus quantitativen und qualitativen Erhebungsmethoden, d.h. für einen Mixed Methods Ansatz plädiert (Craig & Hannum, 2007, S. 40ff.; Maoz, 2000, S. 136). Quantitative Methoden der Datenerhebung kommen überwiegend bei hypothesenprüfenden Aspekten der Evaluation zum Einsatz, um den Grad an Veränderung oder Zielerreichung auf aggregierter Ebene zu messen. Sie ermöglichen eine ökonomische Datenerhebung (Kuckartz, Dresing, Rädiker & Stefer, 2008). Den Einsatz qualitativer Methoden empfehlen viele Autoren besonders dann, wenn es darum geht, Veränderungsprozesse zu explorieren, zu verstehen und anhand konkreter Beispiele zu veranschaulichen (Bortz & Döring, 2006; Craig & Hannum, 2007) sowie Wirkfaktoren für beobachtbare Veränderungen und hinderliche Einflussfaktoren zu identifizieren (Van de Walk & Conostas, 2010). Bortz und Döring (2006, S. 110) plädieren besonders dann für einen qualitativen Zugang, „wenn die mit einer Maßnahme verbundenen Wirkungen sehr komplex sind, sodass ‚eindimensionale‘ quantitative Wirkindikatoren die eigentlichen Effekte bestenfalls verkürzt abbilden können.“ Quantitative Ergebnisse auf aggregierter Ebene können durch das Herausarbeiten subjektiver Erfahrungen von Programmteilnehmern vertieft und veranschaulicht werden, subjektive Erfahrungen einzelner durch den Rückbezug zu aggregierten Auswertungen auf Gruppenebene in den Kontext gesetzt werden (Craig & Hannum, 2007).

Im Hinblick auf quantitative Verfahren der Datenerhebung sei auf zwei besondere Schwierigkeiten im Rahmen der Evaluationsforschung hingewiesen.

Die erste Schwierigkeit betrifft die Auswahl und/oder Konstruktion von Fragebögen zur Operationalisierung der ausgewählten Kriterien. Vor dem Hintergrund klassischer Testgütekriterien wünschenswert wären wissenschaftlich fundierte, sorgfältig konstruierte, etablierte Testverfahren mit überprüften psychometrischen Eigenschaften (Wottawa & Thierau, 2003, S. 131). Verfügbare Testverfahren sind jedoch für Evaluationszwecke oft nicht geeignet, da sie inhaltlich nicht ausreichend auf das Evaluationskriterium abgestimmt sind („asymmetrisch“, das heißt entweder zu eng oder zu weit gefasst sind, vgl. Wittmann, 1990). Dies bedeutet, dass es zuweilen sinnvoller sein kann, „mit einer ‚weichen‘, weniger exakten und elaborierten Methode ein

interessantes Kriterium zu messen als mit hoher Präzision etwas, was inhaltlich nicht zu den eigentlichen Evaluationszielen passt“ (Wottawa & Thierau, 2003, S. 132), also zugunsten der inhaltlichen Passung Einschränkungen bei den klassischen Testgütekriterien und der Vergleichbarkeit mit anderen Untersuchungen hinzunehmen (Craig & Hannum, 2007). In der Evaluationsforschung kommt Nebengütekriterien wie Akzeptanz bei den Stakeholdern, Zumutbarkeit und Ökonomie der eingesetzten Verfahren eine besondere Bedeutung zu (Amelang & Zielinski, 1997, S. 139ff.; Hannum & Martineau, 2008; Wottawa & Thierau, 2003).

Die zweite Schwierigkeit betrifft Probleme der Veränderungsmessung. Hier lassen sich grundsätzlich zwei Arten der Veränderungsmessung unterscheiden: die direkte und die indirekte Veränderungsmessung (Craig & Hannum, 2007; Hannum & Martineau, 2008; Riedelbauch, 2011). Indirekte Veränderungsmessung erfolgt über die Ermittlung der Differenz von Einschätzungen vor und nach der Intervention (Prä-Post-Vergleich). Bei der direkten Veränderungsmessung wird die Veränderung mittels retrospektiver Einschätzung nach der Durchführung der Intervention erfragt. Vor- und Nachteile beider Formen der Veränderungsmessung werden in der Literatur zum Teil kontrovers diskutiert. Die indirekte Veränderungsmessung gilt als strenger und kann dazu führen, dass Veränderungen unterschätzt werden, unter anderem durch den *Response Shift Bias*,⁶ Regressionseffekte oder dadurch dass – besonders bei kleinen Stichproben – die Teststärke nicht ausreicht, um tatsächlich vorhandene Unterschiede aufzudecken. Die direkte Veränderungsmessung ist zwar deutlich ökonomischer, da sie keine Prä-Messung erfordert, ist aber auch anfälliger für *demand characteristics* und überschätzt daher Veränderungen tendenziell (Bortz & Döring, 2006; Craig & Hannum, 2007; Hannum & Martineau, 2008; Riedelbauch, 2011). Auch aufgrund dieser Schwierigkeiten empfehlen einige Autoren, zusätzlich zur Veränderungsmessung die Zielerreichung, zum Beispiel mittels *Goal Attainment Scaling*, zu erfassen (Bortz & Döring, 2006, S. 118; Riedelbauch, 2011; Wottawa & Thierau, 2003, S. 128).

⁶ Der *Response Shift Bias* bedeutet, dass der Bewertungsmaßstab sich durch die Teilnahme an der Intervention ändert. Ein Wert von drei auf einer Skala von eins bis fünf kann nach dem Programm etwas anderes bedeuten als vor dem Programm. Wird zum Beispiel der innere Bewertungsmaßstab strenger, indem einem Teilnehmer durch das Programm deutlich wurde, über welche Führungskompetenzen er noch nicht verfügt, kann ein Postwert von drei auf einer Skala von eins bis fünf eine Verbesserung indizieren, obwohl er mit dem Prä-Wert identisch ist, weil sich der Teilnehmer im Rückblick vor dem Programm lediglich bei zwei einstufen würde. Diese Anpassung der Selbsteinschätzung kann zwar ein im Sinne der Intervention wünschenswertes Ergebnis sein, verdeckt aber bei Prä-Post-Vergleichen tatsächliche Veränderungen (Hannum & Martineau, 2008).

2.2 Untersuchungsgegenstand der vorliegenden Evaluationsstudie

Nachdem grundlegende Begriffe der Evaluationsforschung und generelle methodische Herausforderungen dargestellt wurden, wird im Folgenden der Untersuchungsgegenstand der vorliegenden Studie, das X-Change for Competence Programm, genauer beschrieben. Daraus werden die übergreifenden Forschungsfragen abgeleitet.

2.2.1 *Evaluationsgegenstand: Das X-Change for Competence Programm*

Das Programm X-Change for Competence richtet sich an deutsche und israelische Führungskräfte aus dem Nonprofit-Bereich und verfolgt ein doppeltes Ziel: Es soll die Führungsfähigkeit stärken (*Developing Leadership*) und die Beziehungen zwischen Deutschland und Israel verbessern (*Connecting Societies*). Jede der drei Kohorten wurde zudem unter einem speziellen thematischen Fokus durchgeführt: Bildung im Jahr 2011, Diversity im Jahr 2012 und soziale Gerechtigkeit im Jahr 2013. Zunächst werden der Hintergrund und die Entstehungsgeschichte des Programms erläutert. Anschließend wird auf Programmkonzeption und -ziele genauer eingegangen. Schließlich wird der Ablauf des Programms zusammenfassend beschrieben.

2.2.1.1 *Entstehungsgeschichte des Programms*

Das Programm X-Change for Competence wurde über drei Jahre (2011-2013) hinweg gemeinsam von der Bertelsmann Stiftung und dem Deutsch-Israelischen Zukunftsforum (DIZF) finanziert und durchgeführt.

Die Bertelsmann Stiftung knüpfte dabei an Vorerfahrungen aus mehreren anderen Programmen an, die zum Teil bereits eine langjährige Tradition hatten. Zum einen war das deutsch-israelische Begegnungsprogramm „German-Israeli Young Leaders Exchange (GIYLE)“ über zehn Jahre hinweg fester Bestandteil des Programmportfolios der Bertelsmann Stiftung. Zum anderen existierten konzeptuelle Vorarbeiten zu einem – in der ursprünglich geplanten Form nicht umgesetzten – Programm „Führungsfähigkeit stärken“. Nach zehn Jahren Durchführung des GIYLE-Programms wurde nach Möglichkeiten der Weiterentwicklung gesucht, wobei die bestehenden Erfahrungen in die Neukonzeption des Programms einfließen sollten. Das DIZF, das erst 2009 seine Arbeit als Stiftung aufnahm, hatte als Institution keine Geschichte vergleichbarer Vorläuferprogramme, wenngleich die Programmverantwortlichen des DIZF ihre individuellen Erfahrungen aus früheren beruflichen Zusammenhängen einbrachten.

Ein weiterer Unterschied zwischen den Stiftungen liegt in ihrer grundsätzlichen Ausrichtung: Die Bertelsmann Stiftung ist eine operative Stiftung, die neben dem X-Change for Competence Programm eine Vielzahl unterschiedlicher Projekte und Programme initiiert und implementiert und ein bedeutender Player in der deutschen Stiftungslandschaft ist. Das DIZF hingegen ist eine kleine, fördernde Stiftung, für das die Durchführung eines Programms eine Ausnahme darstellt. Die Durchführung des X-Change for Competence Programms sollte für die Stiftung Netzwerke in Deutschland und Israel erschließen und als Plattform für die Entwicklung von Kooperationsprojekten zwischen den Teilnehmern dienen, die sich dann im Anschluss an das Programm um Fördergelder beim DIZF bewerben könnten.

Durch die Kooperation versprachen sich beide Stiftungen die Möglichkeit, ein Programm in einem Umfang durchführen zu können, der für keine der beiden Stiftungen alleine finanziell umsetzbar gewesen wäre.

2.2.1.2 Programmkonzeption und „Model of Change“ des Programms

Im Folgenden werden die wichtigsten Schritte des Planungsprozesses und die der Programmkonzeption zugrunde liegenden Annahmen, das *Model of Change* des Programms, dargestellt.

Nachdem Mitte 2010 die beiden Trainer auf der Basis der jeweils vorgeschlagenen Konzepte zum Führungstraining von den Stiftungen ausgewählt worden waren,⁷ konstituierte sich die Steuerungsgruppe des X-Change for Competence Programms. Diese bestand aus den Programmverantwortlichen – je zwei Stiftungsvertretern (Programmleiter und Projektleiterin der Bertelsmann Stiftung, Stiftungsvorstand und Projektleiterin des DIZF) – den beiden Trainern und wurde teilweise ergänzt durch für die Hintergrundorganisation verantwortliche Assistenzkräfte. Begleitet wurde die Steuerungsgruppe durch die Evaluatorin. Die Konzeption des Programms erfolgte dabei arbeitsteilig: Während die Trainer verantwortlich waren für die inhaltliche und didaktische Konzeption des Führungstrainings, lag die Verantwortung für die Gestaltung der Programmkomponenten, die auf deutsch-israelische Beziehungen und die kohortenspezifischen Themen bezogen waren (Auswahl von Referenten und Aktivitäten) bei den Stiftungen.

⁷ Dabei handelte es sich um einen israelischen Trainer und eine in Deutschland lebende österreichische Trainerin. Beide hatten zuvor noch nicht miteinander gearbeitet, sondern wurden durch die Stiftungen für das Programm erstmalig als Trainertandem zusammen gebracht.

Im Prozess der Planung kristallisierte sich eine Reihe von Grundannahmen heraus, welche die Schwerpunktsetzung und die konzeptionellen Entscheidungen der Verantwortlichen innerhalb der Steuerungsgruppe leiteten. Wie in Abschnitt 2.1.3.1 erläutert wurde, ist es im Rahmen von Evaluationsstudien sinnvoll, Grundannahmen der Programmverantwortlichen in einem Wirkmodell, dem *Model of Change* des Programms, zu explizieren und zu dokumentieren. Dieses Vorgehen verdeutlicht einerseits Ziele und mögliche Zieldivergenzen der verschiedenen Stakeholder und schafft andererseits eine gemeinsame, verbindliche Grundlage für die Evaluation. In der vorliegenden Evaluationsstudie wurden zunächst die Diskussionen innerhalb der Steuerungsgruppe mitprotokolliert und die Grundannahmen durch die Evaluatorin in einem ersten Entwurf, orientiert an der von Gutiérrez und Tasse (2007) vorgeschlagenen Struktur, systematisiert. Dieser Entwurf wurde der Steuerungsgruppe zur erneuten Diskussion vorgelegt, überarbeitet und Anfang 2011 als Grundlage für die Evaluation verabschiedet.

Das *Model of Change* expliziert die zugrundeliegenden Annahmen der Steuerungsgruppe darüber, aus welcher Ausgangssituation sich der Bedarf für ein solches Programm ergibt, welche Wirkungen das Programm anstrebt und wie das Programm gestaltet sein soll, damit diese Wirkungen erreicht werden können. Abbildung 2.1 verdeutlicht die Struktur des *Model of Change* im Überblick. Das vollständige Wirkmodell findet sich in Anhang A.

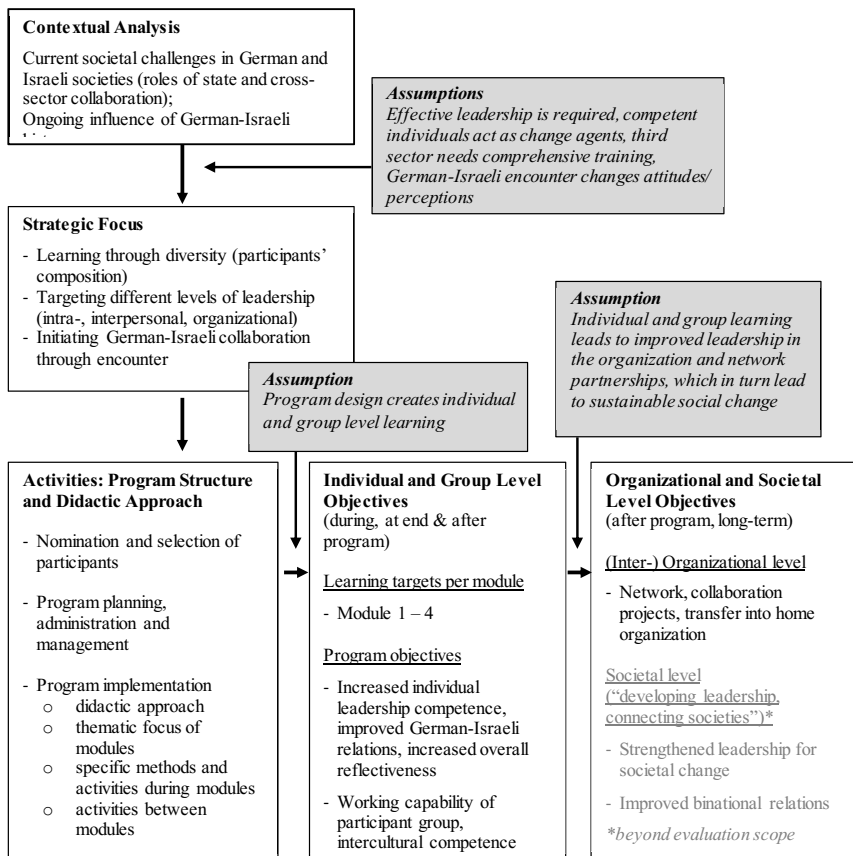


Abbildung 2.1: Überblick über das Model of Change des X-Change for Competence Programms
Anmerkung. Third sector = Nonprofit-Sektor

Zunächst werden hier die Annahmen zur Ausgangssituation und der daraus abgeleitete strategische Fokus des Programms beschrieben. In den folgenden beiden Abschnitten werden Programmziele und -ablauf genauer erläutert.

Begonnen wurde mit einer Analyse der Ausgangssituation, das heißt des generellen gesellschaftlichen Kontexts, aus dem der Bedarf für ein solches Programm abgeleitet wurde. Daraus wurde zunächst ersichtlich, dass Deutschland und Israel aktuell ähnlichen gesellschaftlichen Herausforderungen, zum Beispiel wachsender sozialer Ungleichheit, gegenüberstehen. In beiden Ländern gibt es einen Wandel

gesellschaftlicher Institutionen aus Politik, Zivilgesellschaft, Bildung, Kultur und Wirtschaft und eine Verschiebung der Verantwortlichkeiten weg vom Staat hin zu anderen gesellschaftlichen Akteuren. Diese Veränderungen erfordern eine sektorübergreifende Zusammenarbeit, werden in den beiden Ländern allerdings mit teilweise unterschiedlichen Lösungsstrategien beantwortet. Gleichzeitig ergab die Kontextanalyse, dass deutsch-israelische (bzw. deutsch-jüdische) Beziehungen noch immer von der Vergangenheit beeinflusst werden, wie etwa die von der Bertelsmann Stiftung in Auftrag gegebene Studie “Germans and Jews – United by Their Past, Divided by Their Present?” (Bertelsmann Stiftung, 2009) verdeutlicht.⁸ Vor dem Hintergrund der geschichtlichen Erfahrungen und angesichts unterschiedlicher gegenwärtiger Existenzlagen (u.a. bedingt durch den Nahostkonflikt) wird deutlich, dass auch zwischen Deutschen und Israelis der zweiten und dritten Generation nach dem Holocaust Schwierigkeiten bestehen, einander zu verstehen und miteinander in Beziehung zu treten.

Demgemäß wurde geschlossen, dass das Zusammentreffen von Teilnehmern aus Deutschland und Israel die wechselseitigen Wahrnehmungen verändert. Es wurde angenommen, dass die Diskussion aktueller Fragestellungen und Herausforderungen beide Seiten bereichert. Ebenfalls wurde aus der Kontextanalyse deutlich, dass in beiden Ländern ein Bedarf an effektivem Leadership insbesondere im Nonprofit-Bereich besteht, um künftige Veränderungsprozesse innerhalb von Organisationen und letztlich auf gesellschaftlicher Ebene umzusetzen. Eine Grundannahme dabei war, dass die Entwicklung von Führungskompetenz nicht ohne einen intensiven persönlichen Lernprozess möglich ist und es insbesondere dem Nonprofit-Bereich an systematischen Führungsentwicklungsprogrammen mangelt, in denen ein solcher Lernprozess stattfinden kann. Diese beiden Aspekte wurden im bereits erwähnten doppelten Programmfokus integriert.

Die grundlegende Ausrichtung des Programms, der strategische Fokus, ergibt sich aus dieser Analyse und folgt drei Grundprinzipien: (1) Diversität innerhalb der Teilnehmergruppe ist ein zentrales Element des Lernprozesses. (2) Deutsche und Israelis in einer intensiven Begegnung zusammen zu bringen ist wichtig, um ein nachhaltiges Netzwerk für die Zukunft entstehen zu lassen. (3) Führung umfasst immer mehrere Ebenen – die individuelle, interpersonelle, organisationale und gesellschaftliche Ebene – und erfordert diesen Ebenen zugeordnete Kompetenzen.

⁸ Auf diese Untersuchung wird in Abschnitt 3.2.2.3 näher eingegangen.

Vor diesem Hintergrund sprach das Programm X-Change for Competence überwiegend Führungskräfte aus zivilgesellschaftlichen Organisationen, d.h. aus dem Nonprofit-Sektor an, aufgrund des Bedarfs an sektorübergreifender Zusammenarbeit aber auch einige Teilnehmer aus dem öffentlichen und vereinzelt dem privaten Sektor. Die tatsächliche Zusammensetzung der Teilnehmergruppe wird im Methodenteil im Rahmen der Stichprobenbeschreibung dargestellt.

2.2.1.3 Ziele des Programms

Im Hinblick auf die Ziele des Programms können drei Ebenen von Zielen unterschieden werden: (1) auf der untersten Ebene die Lernziele für jedes Modul, die bereits im Verlauf des Programms erreicht werden sollten, (2) die Programmziele, die bei den Teilnehmern auf individueller Ebene sowie auf der Ebene der Teilnehmergruppe als ganzer zum Programmende bzw. nach dem Programmabschluss angestrebt wurden, (3) langfristige Ziele auf der interorganisationalen Ebene.

Auf der untersten Ebene definierten die Trainer für das Führungstraining jeweils spezifische Lernziele, die im Programmverlauf erreicht werden sollten.

Was die Programmziele angeht, sollte das Programm auf der individuellen Ebene die Führungsfähigkeit verbessern, die Wahrnehmung und Gestaltung deutsch-israelischer Beziehungen positiv beeinflussen und die generelle Reflexivität der Teilnehmer stärken. Auf der Gruppenebene verfolgte das Programm vor allem das Ziel, Arbeitsfähigkeit innerhalb der Teilnehmergruppe als ganzer zu erreichen, was die Fähigkeit der Gruppe zum Umgang mit Diversität und Konflikten einschließen sollte.

Die Lernerfolge von Individuum und Gruppe sollten sich darüber hinaus in nachhaltigen Effekten auf organisationaler und interorganisationaler Ebene niederschlagen: in verbesserten organisationalen Führungskompetenzen und langfristiger, interorganisationaler Vernetzung. Der Aufbau und die Aufrechterhaltung eines sektorübergreifenden, binationalen Netzwerks über die Programmlaufzeit hinaus war ausdrückliches Ziel des Programms und sollte unter anderem die Basis für die Entwicklung konkreter Kooperationsprojekte sein.

Auf der gesellschaftlichen Ebene strebte das Programm letztlich eine verbesserte Beziehung zwischen den beiden Nationen sowie die Stärkung von Führungskompetenzen zum Umgang mit gesellschaftlichen Herausforderungen an. Die gesellschaftliche Ebene übersteigt allerdings den Rahmen der vorliegenden Evaluationsstudie und ist daher nicht mehr Teil des Untersuchungsgegenstandes.

Abbildung 2.2 zeigt die detaillierte Zielformulierung aus dem Model of Change.

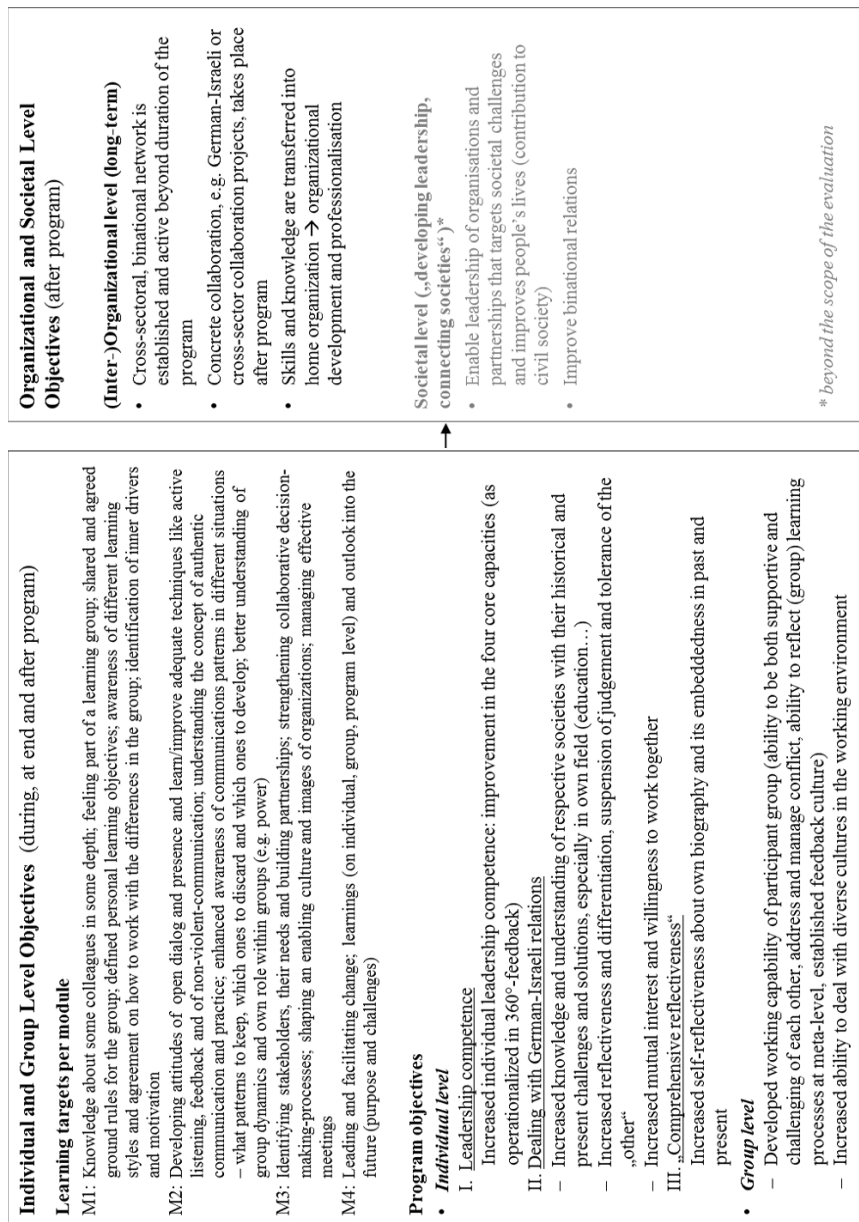


Abbildung 2.2: Ausschnitt aus dem Model of Change: Programmziele

2.2.1.4 Programmablauf

Der nächste Abschnitt geht genauer auf den Ablauf des X-Change for Competence Programms ein. Das Programm wurde in englischer Sprache in den drei aufeinanderfolgenden Jahren 2011, 2012 und 2013 für je 20 Teilnehmer durchgeführt. Jede Kohorte bestand aus zehn deutschen und zehn israelischen Teilnehmern, die über mehrere Monate hinweg an verschiedenen Trainingsmodulen teilnahmen. Die Module fanden abwechselnd in Deutschland und Israel statt und wurden von zwei Trainern, je ein bis zwei Stiftungsvertretern und der Evaluatorin⁹ begleitet. Alle Module enthielten sowohl Trainingsbausteine zum Thema Führung als auch Bausteine, die sich auf deutsch-israelische Beziehungen bezogen.¹⁰

Der Programmablauf der Kohorten 2011 und 2012 gliederte sich in vier fünftägige Module über den Zeitraum von ca. 10 Monaten. Dabei fanden das erste und dritte Modul in Deutschland, das zweite und vierte Modul in Israel statt.¹¹

Was das Führungstraining angeht, weitete sich der inhaltliche Fokus über die Module hinweg von der individuellen bis zur gesellschaftlichen Perspektive: Im ersten Modul ging es überwiegend um die intrapersonale Ebene von Führung. Hier standen Selbstreflexion, Feedback und das Aufspüren eigener (Führungs-)Motive im Vordergrund. Im zweiten Modul ging es vor allem um interpersonale Aspekte von Führung wie Kommunikation und Konflikt. Das dritte Modul umfasste gruppenbezogene Aspekte von Führung, unter anderem Rollen in Gruppen, Macht und Einfluss sowie Entscheidungsstrategien. Das vierte Modul schließlich fokussierte auf Change Management auf organisationaler Ebene und den Beitrag zu gesellschaftlichen Veränderungsprozessen.

Diese führungsbezogenen Trainingsbausteine wurden jeweils durch Themen ergänzt, die auf den Bereich der deutsch-israelischen Beziehungen bezogen waren. Die Themen umfassten dabei einerseits eine Beschäftigung mit der deutsch-israelischen Vergangenheit und dem Holocaust. Andererseits wurden durch Exkursionen und Vorträge mit anschließender Diskussion innerhalb der Gruppe aktuelle gesellschaftliche Themen inklusive des israelisch-palästinensischen Konflikts bearbeitet. Der inhaltliche

⁹ Aus Ressourcengründen war die Evaluatorin im Jahr 2012 nur in den beiden in Deutschland stattfindenden Modulen (Modul 1 und 3) anwesend.

¹⁰ Bausteine zum kohortenspezifischen Thema werden hier dem Bereich deutsch-israelischer Beziehungen zugeordnet. Auf die Gewichtung dieses Themas im Rahmen der Programmdurchführung wird im Ergebnisteil in den Abschnitten 6.3 und 6.4 eingegangen.

¹¹ Die genauen Durchführungsdaten waren im Jahr 2011: 03.-08.02., 07.-12.04, 01.-05.07, 21.-26.10.; im Jahr 2012: 03.-08.05, 05.-10.07., 11.-16.10, 07.-12.02. (2013).

Schwerpunkt bezogen auf die aktuellen gesellschaftlichen Themen orientierte sich dabei an dem jeweiligen kohortenspezifischen Thema. In der Kohorte 2011 handelte es sich dabei um Bildung, in der Kohorte 2012 um Diversity.

Das Programm war darüber hinaus gerahmt von einem 360°-Feedback zu Führungskompetenzen, das einerseits als Intervention genutzt wurde, indem die individuellen Ergebnisberichte im ersten Modul als Bestandteil des Führungstrainings bearbeitet wurden, und andererseits im Prä-Post-Vergleich als Evaluationsinstrument diente.

Abbildung 2.3 verdeutlicht den Ablauf und die inhaltliche Ausgestaltung der Module am Beispiel der ersten Kohorte 2011, deren thematischer Fokus Bildung war.

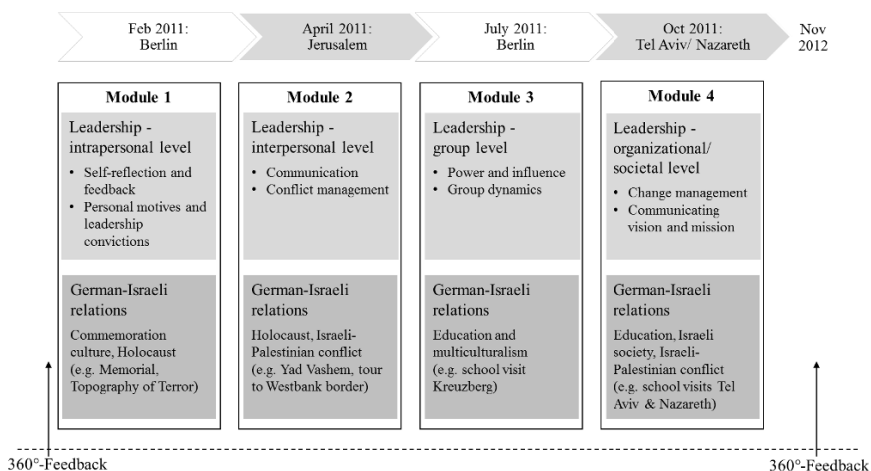


Abbildung 2.3: Ablauf des Programms am Beispiel der Kohorte 2011

Für den dritten Durchführungsjahrgang im Jahr 2013 erfolgte eine Zusammenlegung des Programms auf zwei Trainingsmodule à neun Tage. Das erste Modul fand dabei in Deutschland (Juni 2013) und das zweite Modul vier Monate später (Oktober 2013) in Israel statt. Im Hinblick auf das Führungstraining kann inhaltlich weitestgehend von einer Zusammenlegung des ersten und des vierten Moduls sowie der Module zwei und drei ausgegangen werden. Im ersten Modul des Jahrgangs 2013 ging es um Führung auf intrapersonaler Ebene und Change Leadership. Weiterhin wurden der Holocaust sowie aktuelle gesellschaftlich-politische Themen in Deutschland bearbeitet. Im zweiten Modul folgte die Behandlung des Themas Leadership sowohl auf interpersonalen Ebene

als auch auf Gruppen- und organisationaler Ebene. Darüber hinaus erfolgte die Thematisierung aktueller politischer und gesellschaftlicher Fragestellungen in Israel und die Beschäftigung mit dem israelisch-palästinensischen Konflikt. Das übergeordnete Thema war hier Soziale Gerechtigkeit.

Das Follow-up-360°-Feedback zum Zweck der Evaluation und einer erneuten Standortbestimmung für die Teilnehmer wurde an die unterschiedlichen Programmlaufzeiten angepasst und erfolgte zwölf (2011 und 2012) bzw. sieben (2013) Monate nach Programmende.

2.2.2 Evaluationsauftrag und Forschungsfragen

Grundlage der Evaluation war ein Kooperationsvertrag zwischen der Universität Kassel und der Bertelsmann Stiftung. Die Evaluation sollte einerseits dazu dienen, das Programm fortlaufend zu begleiten, kontinuierlich Feedback zu Stärken und Schwächen in der Implementierung zu geben und insbesondere nach der erstmaligen Durchführung im Jahr 2011 Empfehlungen zur weiteren Programmgestaltung (und ggf. -anpassung) auszusprechen. Dies entspricht der in Abschnitt 2.1.1.2 beschriebenen *formativen* Evaluation. Darüber hinaus sollte die Evaluation aber auch die Zielerreichung, also die Ergebnisse des Programms nach drei Durchführungsjahrgängen zusammenfassend bewerten und förderliche und hinderliche Faktoren für die Zielerreichung identifizieren. Dieser Auftrag entspricht einer *summativen* Evaluation, die sowohl ergebnis- als auch prozessbezogene Elemente umfasst. Vor diesem Hintergrund wurde das Evaluationsdesign längsschnittlich angelegt und umfasste sowohl Prä-Post-Messungen als auch eine kontinuierliche, durchführungsbegleitende Datenerhebung. Als Grundlage für die Evaluation diente das bereits beschriebene Model of Change. Zusätzlich zur Evaluation selbst führte die Universität Kassel im Rahmen einer Zusatzvereinbarung mit der Bertelsmann Stiftung das 360°-Feedback durch.¹² Dies beinhaltete das Projektmanagement, die Datenerhebung mittels Online-Umfrage sowie das Erstellen individueller Feedbackberichte für die Programmteilnehmer.

Der Beitrag, der im Rahmen der formativen Evaluation geleistet wurde, wird im Folgenden kurz zusammengefasst. Gegenstand der vorliegenden Arbeit ist die summative, ergebnis- und prozessbezogene Evaluation des X-Change for Competence Programms. Daher werden im Anschluss an die kurze Zusammenfassung zur formativen Evaluation die leitenden Forschungsfragen für die summative Evaluation formuliert, die

¹² Lediglich die Prä-Erhebung 2011 erfolgte durch ein externes Beratungsunternehmen.

aus dem Model of Change abgeleitet wurden und den Rahmen für die vorliegende Arbeit bilden. In den darauffolgenden Kapiteln wird dann eine theoretische Einordnung vorgenommen, das methodische Vorgehen erläutert und die Ergebnisse der summativen Evaluation dargestellt.

2.2.2.1 Formative Evaluation: Beiträge im Rahmen der Programmbegleitung

Zunächst werden die Beiträge der formativen Evaluation, die über alle drei Kohorten hinweg geleistet wurden, zusammengefasst. Anschließend werden zusätzliche Leistungen in den Kohorten 2011 und 2012 erläutert. Die Erarbeitung des Model of Change mit den Mitgliedern der Steuerungsgruppe war bereits Bestandteil der formativen Evaluation und diente dazu, implizite Annahmen zu explizieren, diese innerhalb der Steuerungsgruppe zu diskutieren und aus dem Interessensabgleich eine gemeinsame Programmgrundlage zu entwickeln.

Über alle drei Durchführungsjahrgänge hinweg nahm die Evaluatorin im Rahmen der Module, bei denen sie anwesend war (d.h. alle Module außer Modul 2 und 4 im Jahr 2012), an allen Staff-Sitzungen teil und steuerte Eindrücke und Anregungen bei. Darüber hinaus wurde den Mitgliedern der Steuerungsgruppe nach jedem Modul ein etwa zehnsseitiger Evaluationsbericht auf Grundlage der von den Teilnehmern ausgefüllten Modulevaluationsbögen¹³ zur Verfügung gestellt. Am Ende eines jeden Durchführungsjahres wurden der Steuerungsgruppe Programmevaluationsberichte zugesandt, die neben den Teilnehmerrückmeldungen zum jeweiligen Durchführungsjahrgang auch Vergleiche über die verschiedenen Module der Kohorte sowie über die Kohorten hinweg beinhalteten.

Die erstmalige Programmdurchführung im Jahr 2011 wurde darüber hinaus besonders intensiv begleitet, um Lessons Learned herauszuarbeiten und diese für die Folgejahrgänge nutzen zu können. Tabelle 2.1 gibt einen Überblick über Beiträge der formativen Evaluation speziell im Jahr 2011.

¹³ Die Beschreibung der eingesetzten Erhebungsinstrumente erfolgt im Methodenteil in Abschnitt 4.3.1. Anhang C1 und C2 enthalten exemplarisch einen Modul- und einen Gesamtevaluationsbogen.

Tabelle 2.1: Ergänzende Beiträge der formativen Evaluation im ersten Durchführungsjahr

Zeitpunkt – Adressat	Beitrag der Evaluation
Fortlaufend – Steuerungsgruppe	• Executive Summary der Kernpunkte zu jedem Modulevaluationsbericht • Teilnahme an allen Steuerungsgruppentreffen zwischen den Modulen, Moderation des Austauschs über die Modulevaluationsberichte sowie Zusammenfassung der wichtigsten Diskussionspunkte und Lessons Learned
Nach Verfügbarkeit und Bedarf – Steuerungsgruppe	• Zusammenfassungen relevanter Informationen aus teilnehmender Beobachtung, Interviews und Fragebögen (z.B. Zusammenstellung der von Teilnehmern als kritisch benannten Führungssituationen, Dokumentation der individuellen Teilnehmerziele)
August 2011 – Bertelsmann Stiftung	• Zwischenbericht mit vorläufigen Schlussfolgerungen und Empfehlungen bezüglich der von der Bertelsmann Stiftung formulierten Fragen, (a) ob die Struktur der vier Module beibehalten werden sollte, (b) ob die Teilnehmerzusammensetzung passend war und in den Folgejahrgängen ähnlich sein sollte und (c) welche Stärken und Schwächen das Führungstraining aufwies im Hinblick auf das Gesamtdesign, Inhalte und theoretische Konzepte, Didaktik und gruppendynamisches Arbeiten
November 2011 – Steuerungsgruppe	• Strukturierte Zusammenfassung der Teilnehmerrückmeldungen und Verbesserungsvorschläge aus der Feedbackrunde zum Programmabschluss
November 2011 – Bertelsmann Stiftung	• Detaillierte, kommentierte Dokumentation der tatsächlichen Implementierung des Programms mit Schwerpunkt auf dem Führungstraining • Schriftliche Zusammenfassung der Empfehlungen für die Gestaltung der folgenden Durchführungsjahrgänge

Im zweiten Durchführungsjahr 2012 brachen fünf Teilnehmer das Programm nach dem ersten Modul ab. Daher wurden mit drei der fünf „Abbrecher“, die sich dazu bereit erklärten, Interviews geführt, um sie zu ihrer Einschätzung des Programms zu befragen und die Gründe für den Ausstieg aus dem Programm zu erfassen. Daraus wurde im Auftrag der Bertelsmann Stiftung ein Bericht erstellt und der Steuerungsgruppe zur Verfügung gestellt.

2.2.2.2 Summative Evaluation: Ableitung von Forschungsfragen

Auf der Basis der bisherigen Ausführungen zu grundlegenden Aspekten der Evaluationsforschung sowie der Beschreibung des Untersuchungsgegenstandes wird im

Folgendes das summative Evaluationsvorhaben der vorliegenden Arbeit konkretisiert. Dazu werden die Forschungsfragen für die Evaluationsstudie abgeleitet.

Als Evaluationsstrategie wurde eine sowohl ergebnis- als auch prozessbezogene Vorgehensweise gewählt. Neben einer Einschätzung des Erfolgs des X-Change for Competence Programms soll auch Ursachenforschung für erzielte – oder ausgebliebene – Ergebnisse betrieben werden. Für die Konzeption zukünftiger Programme können insbesondere prozessbezogene Evaluationsergebnisse wichtige Hinweise auf möglichen Veränderungsbedarf liefern.

Die *ergebnisbezogene Evaluation* lehnt sich an die in Abschnitt 2.1.2.1 ausgeführte Unterscheidung von Wirkebenen an. Ergebnisse des Programms sollen auf mehreren unterschiedlichen Ebenen erfasst werden. Dazu werden die Überlegungen von Martineau et al. (2007) sowie die Vier-Ebenen-Taxonomie der Trainingsevaluation nach Kirkpatrick (1994) integriert. Auf individueller Ebene werden in Anlehnung an Kirkpatrick (1994) Ergebnisse auf den Ebenen 2 und 3, das heißt auf der Lern- und Verhaltensebene erfasst. Die vierte Ebene der Taxonomie bezeichnet Resultate auf der organisationalen Ebene. Da das X-Change for Competence Programm ein organisationsübergreifendes Programm ist, das als langfristiges Ziel die Entwicklung eines über die Programmlaufzeit hinaus aktiven Netzwerks anstrebt und weniger auf Veränderungen innerhalb einzelner Organisationen ausgerichtet ist, wird diese vierte Ebene etwas weiter gefasst. In Anlehnung an Martineau et al. (2007) ist damit in der vorliegenden Evaluationsstudie die Bildung von persönlichen und interorganisationalen Netzwerken gemeint.

Im Rahmen der *ergebnisbezogenen* Evaluation wird in der Arbeit der folgenden übergeordneten Fragestellung nachgegangen:

Welche Wirkungen konnten durch das X-Change for Competence Programm erzielt werden?

Diese übergreifende Fragestellung lässt sich in mehrere Teilfragen aufgliedern. Die Forschungsfragen, die sich auf die individuelle Ebene beziehen, umfassen sowohl die Lern- als auch die Verhaltensebene (Kirkpatrick, 1994):

Hat sich die Führungskompetenz der Programmteilnehmer über die Zeit verbessert?
Hat sich die Wahrnehmung und Gestaltung deutsch-israelischer Beziehungen über die Zeit positiv verändert?

Außerdem sollen die Ergebnisse zur Vernetzung der Teilnehmer, als Erweiterung der Ebene organisationaler Resultate bei Kirkpatrick (1994) in Anlehnung an Martineau et al. (2007) untersucht werden:

BLEIBEN DIE PROGRAMMTEILNEHMER ÜBER DAS PROGRAMM ENDE HINAUS MITEINANDER IN KONTAKT UND WIE GESTALTET SICH DIESER KONTAKT?

Neben der Betrachtung der Wirkungen des X-Change for Competence Programms interessiert im Rahmen der *prozessbezogenen Evaluation*, welche Einflussfaktoren diese Wirkungen ermöglicht und welche Einflussfaktoren sie behindert haben. Dazu wird auf die in Abschnitt 2.1.2.2 eingeführte Unterscheidung zwischen teilnehmer-, programm- und kontextbezogenen Einflussfaktoren Bezug genommen. Programmbezogene Einflussfaktoren stehen im Fokus der vorliegenden Untersuchung. Teilnehmer- und kontextbezogene Faktoren werden nur ergänzend untersucht, um differenzielle Wirkungen des Programms etwa in Abhängigkeit von der Nationalität der Teilnehmer zu erfassen. Die Forschungsfrage zum Gruppenprozess, der im Model of Change unter den Programmzielen aufgeführt wird („Erreichen von Arbeitsfähigkeit innerhalb der Teilnehmergruppe“), wird in die prozessbezogene Evaluation eingeordnet, da die Arbeitsfähigkeit der Gruppe nicht am Ende festgestellt werden kann, sondern eng mit dem Gruppenprozess über den Verlauf des Programms hinweg zusammenhängt. Die folgende übergeordnete Fragestellung leitet die prozessbezogene Evaluation:

WELCHE EINFLUSSFAKTOREN TRAGEN ZU DEN WIRKUNGEN DES X-CHANGE FOR COMPETENCE PROGRAMMS BEI?

Auch hier lässt sich die übergeordnete Fragestellung in mehrere Teilfragen aufgliedern und somit konkretisieren. Der Schwerpunkt der prozessbezogenen Evaluation liegt auf Einflussfaktoren, die mit der konkreten Ausgestaltung des X-Change for Competence Programms zusammenhängen:

INWIEFERN TRÄGT DIE SPEZIFISCHE PROGRAMMGESTALTUNG DES X-CHANGE FOR COMPETENCE PROGRAMMS ZU DEN WIRKUNGEN BEI?

- Welche Aspekte des Programmdesigns (Programmstruktur, generelle methodisch-didaktische Ausrichtung, Gestaltung der einzelnen Programmkomponenten) sind insgesamt positiv, welche kritisch zu bewerten?
- Wie ist die Gruppenzusammensetzung zu bewerten und wie gestaltet sich der Gruppenprozess innerhalb der Teilnehmergruppe?
- Wie ist die Gewichtung und Kombination der verschiedenen Programmkomponenten zu bewerten?

Teilnehmer- und kontextbezogene Faktoren werden nur ergänzend untersucht, um differenzielle Wirkungen und den Einfluss des institutionellen Rahmens zu erfassen. Die zugehörigen Fragen lauten:

Inwiefern tragen Merkmale der Teilnehmer und des institutionellen Rahmens (Kooperationsbeziehungen zwischen den Stiftungen) zu den Wirkungen des Programms bei?

Deutsch-israelisches Leadership Development

Eine Evaluationsstudie

Kotte, S.

2017, XXIII, 391 S. 27 Abb., Softcover

ISBN: 978-3-658-16246-7