Was muss bei der Evaluation berücksichtigt werden?
Was muss bei Evaluationen berücksichtigt werden? Aussagen über die Wirksamkeit von Kompetenzentwicklungsmaßnahmen werden durch Vergleiche gewonnen.

Prä- und Postmessungen

Ein grundlegendes Ziel eines Trainings ist, dass sich die Teilnehmer nach dem Training im weitesten Sinne anders verhalten als vor dem Training. Um eine solche Veränderung festzustellen reicht es aber nicht allein, die Trainingsgruppe nach der Beendigung des Trainings einer Postmessung zu unterziehen. Vorher-Nachher-Messungen sind reinen Nachher-Messungen vorzuziehen. Eine Prämessung vor der Durchführung des Trainings stellt eine Vergleichsbasis her. Mit Bedacht zu wählen ist der Zeitpunkt der Postmessung. So erfasst eine Postmessung direkt im Anschluss an das Training den Lernerfolg, eine Postmessung mehrere Wochen oder Monate später den Transfererfolg.

Interne Validität

Wenn durch Prä- und Postmessungen eine Veränderung in der Trainingsgruppe festgestellt wird, stellt sich anschließend die Frage nach der internen Validität (Gültigkeit) der Untersuchung: geht die Veränderung auf das Training zurück oder hat sie andere Ursachen? Bei niedriger interner Validität ist unklar, worauf die Veränderung der Trainingsgruppe zurückzuführen ist. Dadurch ist die Untersuchung wertlos, weil sie keine Information über die Wirksamkeit des Trainings liefert. Daher trägt ein experimentelles Design mit Kontrollgruppe entscheidend zur internen Validität bei.

Kontrollgruppe

Die Einführung einer Kontroll- oder Vergleichsgruppe ist ein bewährtes Mittel, hohe interne Validität herzustellen. Die Kontrollgruppe wird genauso behandelt wie die Trainingsgruppe, mit der Ausnahme, dass sie das zu evaluierende Training nicht erhält. Die Vergleichsgruppen als Kontrollgruppen sollen es ermöglichen, nicht-programmgebundene Wirkungen zu kontrollieren, um Veränderungen tatsächlich auf das Training zurückführen zu können. Anhand des Vergleichs der Trainingsgruppe mit der Kontrollgruppe lässt sich unter anderem ausschließen, dass eine Veränderung in der Trainingsgruppe allein auf die verstrichene Zeit (Mitarbeiter haben z.B. Zeit gehabt, zu üben und sich dadurch zu verbessern) oder äußere Ereignisse (z.B. Konjunktureinbrüche, Entlassungen, Sommerferien) zurückgeht. Denn solche Ursachen wirken sich in gleichem Maße auf die Kontrollgruppe aus. Ohne Kontrollgruppe kann eine Konfundierung der Interventionswirkung mit der Wirkung anderer auftretender Faktoren nicht ausgeschlossen werden und die Wirkung wird im schlimmsten Fall irrtümlicherweise auf die Intervention attribuiert (Hager et al., 2000).

Hawthrone Effekt

Die Einführung einer Kontroll- oder Vergleichsgruppe ist ein bewährtes Mittel, hohe interne Validität herzustellen. Die Leistung der Trainingsgruppe kann sich schon allein dadurch verändern, dass ihr mehr Aufmerksamkeit geschenkt wird und sie sich beobachtet fühlt. Die Kontrollgruppe sollte daher, wenn praktisch und vor allem finanziell umsetzbar, ebenso viel Aufmerksamkeit und Beobachtung erfahren wie die Trainingsgruppe. D.h. dass die Kontrollgruppe z. B. ein nicht aufwendiges „Placebotraining“ erhält, von dem man sich keine übermäßig große Wirkung, aber auch keinen Schaden verspricht. Der Effekt, dass Versuchspersonen ihr natürliches Verhalten ändern können, wenn sie wissen, dass sie Teilnehmer an einer Untersuchung sind, stellt eine Bedrohung der externen Validität dar. Die Entdeckung des Effektes geht auf die sogenannten **Hawthorne-Experimente** von Roethlisberger und Dickson zurück, die sie zwischen 1924 und 1932 in der Hawthorne-Fabrik der Western Electric Company in Chicago (USA) durchgeführt wurden, um festzustellen, wie man die Leistung von Arbeitern steigern kann. Erstaunerlicherweise steigerte sich die Leistung der Arbeiter unabhängig davon, welche – auch konträre – Interventionen vorgenommen wurden.

Experimentelle und quasiexperimentelle Evaluation

Evaluationen, die neben der Trainingsgruppe eine Kontrollgruppe verwenden und die Personen zufällig diesen Bedingungen zuordnen (Randomisierung), gelten als **experimentelle Evaluationen**. Wenn es praktisch möglich ist, sollte die Zuweisung der Teilnehmer auf die Trainings- und Kontrollgruppe zufällig erfolgen. In der Unternehmenspraxis werden häufig **quasi-experimentelle Evaluationen** eingesetzt, bei denen im Gegensatz zur experimentellen Evaluation keine Randomisierung stattfinden konnte. Ohne zufällige Zuweisung muss man begründen können, warum sich die beiden Gruppen nicht systematisch voneinander unterscheiden. Systematische Unterschiede untergraben den Sinn einer Kontrollgruppe und somit die interne Validität. Auch Zuweisungsverfahren wie die Parallelisierung, bei der darauf geachtet wird, dass die beiden Gruppen auf Merkmale wie Geschlecht oder Alter bezogen identisch sind, bieten keine Garantie gegen systematische Unterschiede. In der Regel wird man durch die Parallelisierung die wirklich wichtigen Merkmale gar nicht berücksichtigen, weil man sie nicht kennt.

Tab. Mögliche Störvariablen bei der experimentellen Evaluation

Reaktion auf Gruppenzuweisung	Schon die Zuweisung in die Trainings- bzw. Kontrollgruppe wirft bei den Teilnehmern Fragen auf: Warum erhält einer das womöglich begehrte neue Training, der andere nicht? Oder bekommen die einen etwa das Training, weil ihre Leistung schwächer ist? Damit Mitarbeiter sich nicht abgewertet fühlen, ist es wichtig, über das logische Prinzip hinter der Gruppenzuweisung aufzuklären. Sonst könnte es z.B. passieren, dass die Kontrollgruppe aus Resignation unter ihrem normalen Leistungsniveau bleibt.
Rivalität zwischen den Gruppen	Wenn die Kontrollgruppe das zu evaluierende Training als besonders „begehrenswert“ sieht, entsteht leicht eine Rivalität zwischen der Kontroll- und Trainingsgruppe. Rivalität kann dazu führen, dass sich die Kontrollgruppe besonders anstrengt, um die Trainingsgruppe zu übertrumpfen und dadurch die Effekte des Trainings, im Vergleich mit der Kontrollgruppe, nicht mehr erkennbar sind.
Reaktionen der Trainingsgruppe	Die Leistung der Trainingsgruppe kann sich schon allein dadurch verändern, dass ihr mehr Aufmerksamkeit geschenkt wird und sie sich beobachtet fühlt. Die Kontrollgruppe sollte daher, wenn praktisch und vor allem finanziell umsetzbar, ebenso viel Aufmerksamkeit und Beobachtung erfahren wie die Trainingsgruppe. Das kann bedeuten, dass die Kontrollgruppe ein nicht aufwendiges „Placebotraining“ erhält, von dem man sich keine übermäßig große Wirkung, aber auch keinen Schaden verspricht.
Veränderungen der Messinstrumente	Manchmal gehen Veränderungen von der Prä- zur Postmessung nicht auf das Training sondern auf die Messinstrumente zurück. Dies kann der Fall sein, wenn z.B. die Leistungsbewertung in der Postmessung weniger streng ist oder Beurteiler sich im Laufe der Untersuchung weniger Mühe geben.
Regression zur Mitte	Teilnehmer in Evaluationsstudien werden oft wegen extrem hoher oder niedriger Testwerte ausgewählt. Beispielsweise erhalten vielleicht nur ausgesprochen leistungsstarke oder -schwache Mitarbeiter ein Training. Mit der Zeit tritt aber ganz von selbst eine Regression zur Mitte auf. Dies bedeutet, dass sowohl extrem hohe als auch extrem niedrige Werte über die Zeit auf einen mittleren Wert hin tendieren. Der Grund dafür ist, dass Tests keine perfekten Messinstrumente sind und jede Messung fehlerbehaftet ist. Die Messwerte variieren um einen „wahren Wert“ der Person herum. Bei extremen „wahren Werten“ kann diese Variation aber nur in Richtung der Skalenmitte gehen, denn zu den Skalenendpunkten hin ist kein Platz mehr für Variation, da

	das Messinstrument gute oder schlechte Leistungen im Extrembereich weniger gut differenzieren kann. Teilnehmer hatten vielleicht am Tag des Prätests außergewöhnlich viel Glück oder Pech, so dass beim Posttest ihre Leistung viel weniger extrem ist. Eine Regression der Werte zur Mitte kann leicht mit einem Trainingseffekt verwechselt werden.
Selektives Drop-out	Drop-out bezeichnet das frühzeitige Ausscheiden von Teilnehmern aus der Untersuchung. Problematisch wird es, wenn das Ausscheiden selektiv ist, d.h. wenn dadurch die relative Zusammensetzung der Kontroll- und Trainingsgruppe verändert wird. In einer Kontrollgruppe könnten z.B. alle unmotivierten Teilnehmer wegfallen, während sich die unmotivierten in der Trainingsgruppe verpflichtet fühlen zu bleiben. Ein unterschiedliches Abschneiden der Gruppen in der Postmessung wäre dann zumindest teilweise auf die höhere Motivation in der Trainingsgruppe zurückzuführen.
Sensitivierung durch die Prämessung	Es kann vorkommen, dass die Teilnehmer durch die Prämessung einem Trainingsinhalt mehr Aufmerksamkeit schenken. Beispielsweise achten Teilnehmer vielleicht besonders auf Ausführungen zur Kundenzufriedenheit, wenn dieses Thema häufig im Prätest auftaucht. Wenn sich dann durch das Training ihre Leistung im Kundenumgang besonders stark verbessert ist dieser Effekt möglicherweise nicht auf Personengruppen generalisierbar, die keinen Prätest erhalten haben.

Externe Validität

Angenommen von der Prä- zur Postmessung konnte eine Veränderung im Verhalten der Trainingsgruppe festgestellt werden. Ist diese Veränderung auf andere Personen und Situationen generalisierbar? Diese Frage betrifft die externe Validität einer Untersuchung. Je mehr sich die Untersuchungsteilnehmer von anderen Personen unterscheiden, desto weniger sind die Ergebnisse der Evaluation auf andere Personen übertragbar. Bestehen die Untersuchungsgruppen beispielsweise nur aus sehr leistungsstarken Mitarbeitern muss man davon ausgehen, dass das Training einen anderen Effekt auf leistungsschwache Mitarbeiter haben könnte. Bestehen die Untersuchungsgruppen nur aus im Team arbeitenden Mitarbeitern hat das Training womöglich andere Effekte auf sie, als auf Mitarbeiter, die traditioneller Einzelarbeit nachgehen. Die Teilnahme an einer Trainingsevaluation unterscheidet sich beträchtlich vom normalen Arbeitsalltag. Es werden Tests durchgeführt, man wird beobachtet (manchmal sogar mit Videokameras), ein Untersuchungsleiter und seine Assistenten sind immer wieder anwesend. Trainingseffekte, die in dieser ganz besonderen Situation auftreten, kommen vielleicht im normalen Arbeitsalltag nicht vor. Notwendige Voraussetzung für die Generalisierbarkeit ist, dass die interne Validität gegeben ist. Denn wenn unklar ist, ob es überhaupt einen Trainingseffekt gab, kann man sich das Nachsinnen über die Anwendung bei anderen Personengruppen und Situationen sparen.

Follow Up Messungen

Follow Up Messungen. Um feststellen zu können, ob der Kompetenzzuwachs über die Zeit bestehen bleibt oder um Langzeiteffekte zu identifizieren, die möglicherweise bei der Nachher-Messung verdeckt geblieben sind, ist es nötig Follow-up-Messungen durchzuführen. Diese erlauben Aussagen darüber, ob es sich nur um einen kurzfristigen Erfolg der Maßnahme gehandelt hat. Vor der Trainingsimplementierung sollte bereits eine Vorstellung darüber bestehen, wie stark die durch das Training ausgelöste Veränderung sein sollte. An diesem Ziel hat sich das Training daraufhin zu bewähren ▸ 4. Die Messung zu mehreren Zeitpunkten nach Beendigung der Interventionsmaßnahme ist äußerst ratsam, da sich Leistungsentwicklungen nicht immer anhand einmaliger Post-Interventions-Messungen feststellen lassen (Hager, W. et al., 2000).

Box

Wirksamkeitsverläufe erfolgreicher Trainingsprogrammen

Anhand der folgenden grafischen Darstellungen lässt sich veranschaulichen, welche Wirksamkeitsverläufe bei einer Intervention auftreten können. Der angewendete Versuchsplan umfasst drei Erhebungszeitpunkte und zusätzlich eine Vergleichsgruppendarstellung.

Der in ▸ **Abb.** dargestellte Typ einer erfolgreichen Intervention ist in der Praxis sehr verbreitet. Vom Vortest zum Nachtest findet ein erkennbarer Leistungszuwachs statt, die Intervention ist also wirksam. Die Leistungsdifferenz zwischen Interventions- und Vergleichsgruppe bleibt im Nachtest bestehen, d.h. der Leistungszuwachs bleibt stabil. Eine weitere Leistungssteigerung tritt jedoch nicht auf.

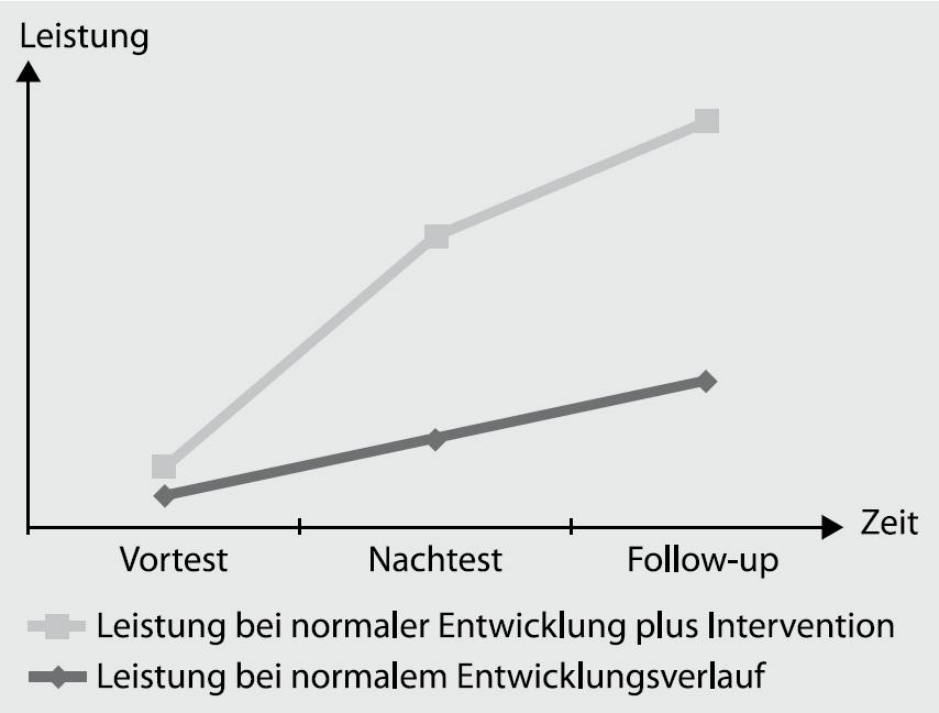

Abb. 1 Typ erfolgreicher Intervention.

Der Idealtypus einer erfolgreichen Intervention, der in ▸ **Abb.** illustriert ist, zeichnet sich durch einen zusätzlichen Anstieg in der Leistungsdifferenz zwischen Nachtest und Follow-up bei der Interventions- und Vergleichsgruppe im Follow Up aus. Dies bedeutet eine zusätzliche Leistungssteigerung und einen längerfristigen Entwicklungsschub für die Trainingsgruppe.

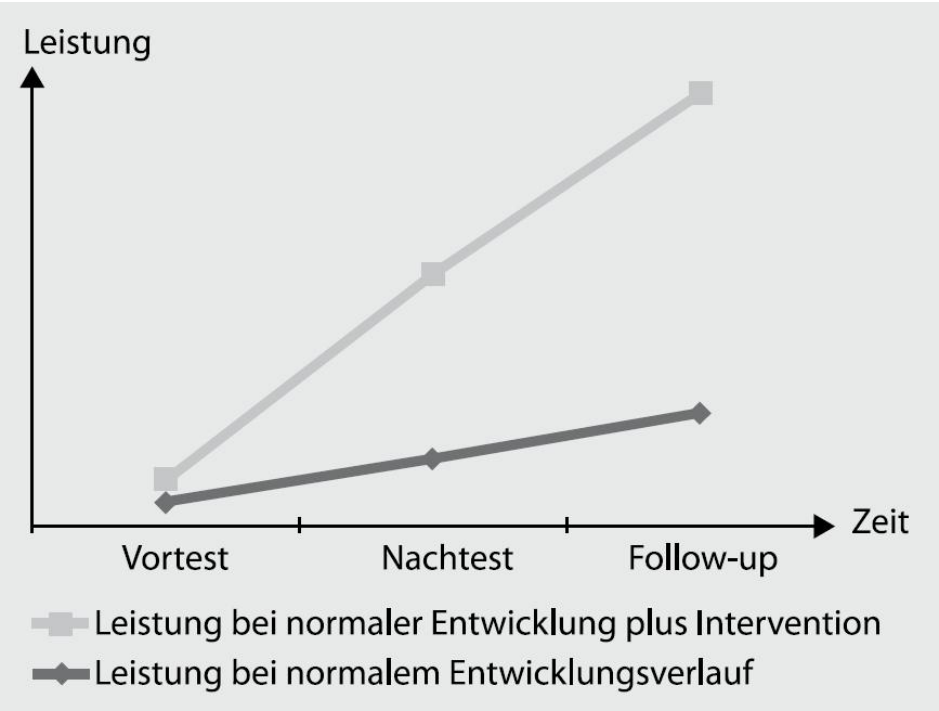

Abb. 2 Idealtypus einer Intervention.

▸ **Abb.** zeigt den oft erwünschten Fall bei einer Präventionsmaßnahme. Hierbei liegt zwar unmittelbar nach der Interventionsdurchführung kein signifikanter Unterschied in der Leistung vor, doch auf lange Sicht gesehen ist ein Entwicklungsschub zu verzeichnen.

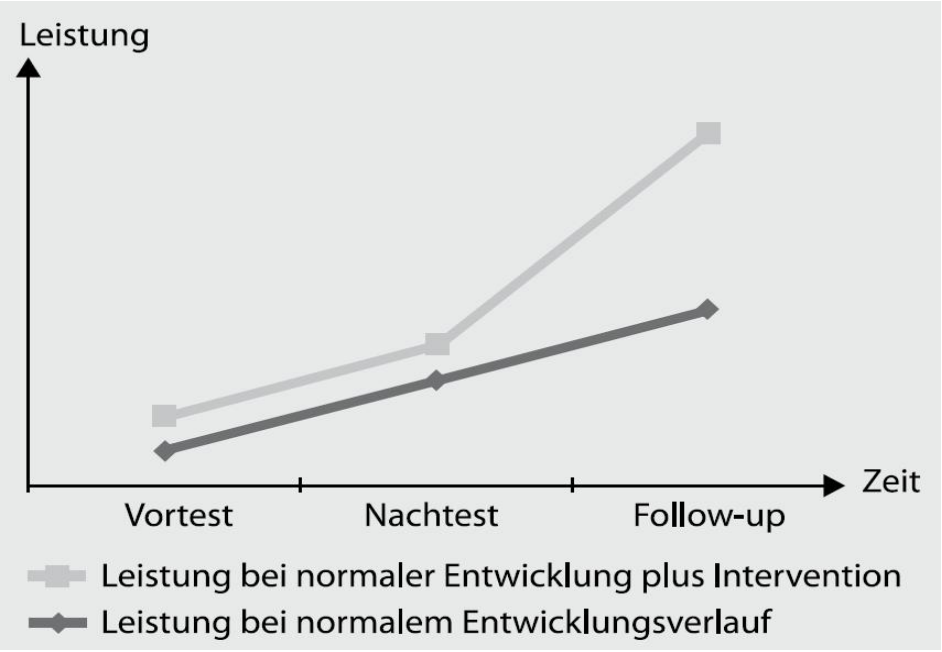

Abb. 3 Interventionstyp mit längerfristigem Entwicklungsschub.

Box

Kasuistische Evaluation

Evaluationen, bei denen keine Kontroll- oder Vergleichsbedingung zur Gegenüberstellung mit der zu evaluierenden Maßnahme herangezogenen werden kann, gelten als kasuistische Evaluationen. Wenn

nur die Bewertung eines einzigen Trainingsprogramms möglich ist, sollten vorab Zielgrößen festgelegt werden, an deren Erreichung das Training zu messen ist. Darüber hinaus kann es hilfreich sein, Vergleichs- oder Benchmarkwerte standardisierter Fragebögen zu nutzen, die sich auf Daten vergleichbarer Trainingsmaßnahmen aus anderen Unternehmen beziehen. Dies setzt jedoch den Einsatz entsprechender Instrumente voraus, die in der Regel nur sehr global und nicht spezifisch auf die einzelne Trainingsmaßnahme zugeschnitten den Transfer messen ▸ **Fehler! Verweisquelle konnte nicht gefunden werden.**; ▸ **Fehler! Verweisquelle konnte nicht gefunden werden.**. Um Veränderungen abzubilden, kann in der einmaligen Postbefragung direkt nach Veränderungen gefragt werden wie z.B. Durch das Training hat sich meine Arbeitsqualität verbessert. Darüber hinaus können retrospektive Einschätzungen vorgenommen werden bei denen die Teilnehmer nach dem Training ihre Kompetenz vor dem Training und nach dem Training anhand der gleichen Aussagen einschätzen. Veränderungen können so aufgezeigt werden. Gegenüber einem klassischen Vortest- Nachtest Design hat dieses Vorgehen ggf. den Vorteil, dass Veränderungen aufgedeckt werden können, die im klassischen Vortest-Nachtest durch die Wissensveränderung zum Trainingsgegenstand verdeckt bleiben: Vor dem Training schätzen die Teilnehmer ihre Fähigkeiten zur Kundenansprache im Vertreib als durchaus gut ein. Im Training werden die Teilnehmer an ihre Grenzen geführt und bekommen eine Idee davon, was sie alles noch nicht wissen. Das Training dauert an, die Teilnehmer lernen dazu und stabilisieren sich in ihrer Kompetenzeinschätzung auf dem Niveau der Vorhermessung.

Erfolgsmaße

Die Maße, mit denen der Erfolg gemessen wird, sollten sich an den Zielen des Trainingsprogramms orientieren und psychometrischen Gütekriterien genügen. Ein Grund für den zögerlichen Einsatz systematischer Evaluationen kann so auch im mangelnden Angebot von standardisierten Evaluationsinstrumenten liegen, die mit einem ökonomisch vertretbaren Aufwand interpretationsfähige Ergebnisse liefern (vgl. z.B. Maßnahmenerfolgsinventar, Lerntransfer-System-Inventar).

Literatur

Hager, W., Patry, J.-L. & Brezing, H. (2000). *Evaluation psychologischer Interventionsmaßnahmen: Standards und Kriterien: ein Handbuch*. Bern: Verlag Hans Huber

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

Kauffeld, S. (2010). *Nachhaltige Weiterbildung*. Heidelberg: Springer.