Glossar

Abkürzungen
AV: abhängige Variable
UV: unabhängige Variable

Adversary Evaluation → gegnerschaftsorientierte Evaluation

Änderungssensitivität. Damit ein Messinstrument in der Lage ist, Veränderungen über die Zeit hinweg erfassen zu können (→ Veränderungsmessung), muss es für solche Veränderungen sensitiv genug sein. Die Änderungssensitivität ist eine Funktion der Stabilität des zu messenden Merkmals (je stabiler, d. h. zeitinvarianter, desto weniger Veränderung gibt es) und der → Schwierigkeit des Messinstruments: Zu schwierige oder zu leichte Messinstrumente können per se nicht änderungssensitiv sein.

Aufforderungscharakter/Demand Characteristics. In jeder diagnostischen Situation generieren die Befragten Hypothesen darüber, welches Verhalten von ihnen erwartet wird. Solche Vermutungen können die Evaluationsdaten verfälschen, denn die Personen können motiviert sein, der Erwartung Folge zu leisten oder nicht. Ein hoher Aufforderungscharakter kann die Messwerte also verzerren, und das entweder unsystematisch (das erhöht die Fehlervarianz und verringert die → Teststärke) oder systematisch (das verringert die → interne Validität).

Ausbalancierung. Hierbei handelt es sich um eine Möglichkeit, den Einfluss von Störvariablen zu kontrollieren und damit die → interne Validität einer Evaluationsuntersuchung zu erhöhen. Eine Störvariable ist ausbalanciert, wenn es gelingt, die Verteilung von Ausprägungen einer Störvariablen in allen Stufen der UV gleich zu halten.

Ausführungsintegrität/Treatment Fidelity. Für viele klinisch- und pädagogisch-psychologische Interventionsmaßnahmen existieren Manuale, in denen Inhalte, Abläufe, didaktische Methoden und möglicherweise sogar Instruktionen detailliert vorgegeben werden. Eine Aufgabe der → formativen Evaluation kann darin bestehen, empirisch zu ermitteln, ob die Maßnahme tatsächlich wie vorgesehen durchgeführt wird oder ob es Abweichungen vom Manual gibt (→ Implementationskontrolle). Nur wenn Ausführungsintegrität gegeben ist, kann die → Nettowirkung einer Intervention ermittelt werden. Außerdem erhöht die Sicherung der Ausführungsintegrität die → interne Validität der Evaluationsuntersuchung.

Auspartialisierung. Eine Möglichkeit, den Einfluss von Störvariablen zu kontrollieren, besteht darin, die Varianz der AV um jenen Anteil, den diese mit der Störvariablen gemeinsam hat, zu bereinigen. Die resultierende Variable ist dann mit der Störvariablen unkorreliert. Als Alternativerklärung für einen Effekt der UV auf die AV kommt die Störvariable also nicht mehr in Frage. Die Auspartialisierung einer Störvariablen erhöht also die → interne Validität einer Untersuchung. In der varianzanalytischen Terminologie werden Analysen mit Kovariaten als Kovarianzanalysen bezeichnet.

Ausschöpfungsquote. Interventionsmaßnahmen richten sich an mehr oder weniger spezifische Zielgruppen: Es gibt Personen, die tatsächlich Bedarf an der Intervention haben und solche, die keinen Bedarf haben. Je mehr bedürftige und je weniger nicht-bedürftige Personen an der Intervention teilnehmen, desto größer ist die Ausschöpfungsquote der Intervention. Gelingt es nicht, die bedürftigen Personen zur Teilnahme an der Intervention zu bewegen, so leidet die Interventionsstichprobe unter Underinclusion. Besteht die Interventionsstichprobe hingegen aus vielen Personen, die eigentlich gar keinen Bedarf haben, spricht man von Overinclusion.

Autoregressor-Modelle. In solchen Modellen werden die Pre-Testwerte aus den Post-Testwerten → auspartialisiert (→ auto-residualisierte Werte). Dadurch wird die AV um Vortestunterschiede bereinigt: Man tut so, als hätten alle Personen den gleichen Wert im Pre-Test gehabt. Ein solches Vorgehen ist nichts anderes als eine Kovarianzanalyse, bei der der Pre-Test eine Kovariate (bzw. ein Autoregressor) ist. In Autoregressor-Modellen wird eine AV (gemessen zu einem »späteren« Zeitpunkt) gleichsam »um sich selbst« (gemessen zu einem früheren Zeitpunkt) bereinigt.

auto-residualisierte Variable/Residual Change Score. Werden im Rahmen von → Autoregressor-Modellen Post-Testwerte um Pre-Testwerte bereinigt, resultieren sog. auto-residualisierte Werte (engl. Residual Change Scores). Residual Change Scores geben an, ob der tatsächliche Post-Testwert größer oder kleiner als erwartet ist. Mit »erwartet« ist gemeint, welchen Wert die Person gehabt hätte, wenn es lediglich einen → Re-

gressionseffekt, nicht aber »wahre« Veränderung gegeben hätte.

Bedarfsanalyse/Need Assessment. Mit Hilfe einer Bedarfsanalyse soll geklärt werden, ob und wie groß in einem konkreten Fall der Bedarf an einer Intervention ist. Hierzu müssen insbesondere drei Dinge abgeklärt werden: (1) Gibt es überhaupt ein Problem und wie ist das Problem beschaffen, d. h. wie groß ist es, wo ist das Problem lokalisiert, wie lange existiert es schon, wie verteilt es sich, wo sind die Ursachen zu suchen? (2) Was ist die Zielgruppe, wodurch ist sie definiert und wie kann man die Zugehörigkeit einer Person zur Zielgruppe empirisch ermitteln? (3) Worin besteht das Ziel der Intervention? Der Bedarf an der Intervention ist gegeben, wenn plausibel erwartet werden kann, dass das Problem der Zielgruppe mit Hilfe der Intervention auch tatsächlich gelöst werden wird.

Benchmark. Benchmark bezeichnet ein strukturiertes und formalisiertes Konzept, das dazu eingesetzt wird, Optimierungen durch den Vergleich von Leistungsindikatoren vergleichbarer Objekte, Prozesse oder Programme vorzubereiten. Die Vergleichbarkeit wird durch ähnliche Randbedingungen gewährleistet.

Boden- vs. Deckeneffekt. Von einem Bodeneffekt spricht man, wenn die Werte einer Person (oder einer Gruppe von Personen) so niedrig sind, dass sie nach unten keinen weiteren Veränderungsspielraum haben. Bodeneffekte können resultieren, wenn die tatsächliche Merkmalsausprägung wirklich so gering ist (z. B. ein schwach ausgeprägtes Problem, sehr geringe soziale Kompetenzen) oder wenn das Messinstrument eine zu hohe → Schwierigkeit hat. Der Deckeneffekt ist das Gegenteil des Bodeneffekts: Hier sind die Werte so hoch, dass sie nach oben hin keinen Veränderungsspielraum mehr haben. Die hohen Werte können daraus resultieren, dass die Merkmalsausprägung tatsächlich so hoch ist, oder dadurch, dass das Messinstrument eine zu geringe → Schwierigkeit hat.

Bruttowirkung → Nettowirkung

Complianceevaluation. Im Rahmen einer Compliance- evaluation wird das Einverständnis, die Einwilligung, die Zustimmung oder die aktive Mitarbeit der von der Interventionsmaßnahme betroffenen Personen bewertet. Die Compliance (dt.: Befolgung, Einhaltung, Übereinstimmung, Zustimmung) ist eine wichtige Bedingung für die Wirksamkeit einer Maßnahme. Eine Maßnahme, die von den Betroffenen nicht akzeptiert wird, wird mit großer Wahrscheinlichkeit auch nicht wirksam sein. Insofern sollte Complianceevaluation im Idealfall bereits im Rahmen einer → prospektiven Evaluation stattfinden. Auch im Rahmen einer → formativen Evaluation kann die Compliance Evaluationsgegenstand sein.

Confirmation Bias → Konfirmationseffekt

Demand Characteristics → Aufforderungscharakter

Drop-out → experimentelle Mortalität

Effektstärke. Standardisierte, d. h. über Untersuchungen, Variablen, Skalierungen usw. hinweg vergleichbare Quantifizierung eines empirischen Effekts. Während ein statistischer Test (Signifikanztest) lediglich eine Entscheidung über die Ablehnung der Null- oder der Alternativhypothese nahe legen kann (→ Irrtumswahrscheinlichkeiten), erlaubt die Effektstärke eine Aussage darüber, wie groß der gefundene Effekt ist. Daher wird die Effektstärke auch als Maß der praktischen Signifikanz bezeichnet.

Effizienzanalyse. Bei der Effizienzanalyse geht es um den Nutzen und die Kosten-Nutzen-Bilanz einer Maßnahme. Effizienzanalysen helfen, Entscheidungen für eine konkrete Maßnahme (und gegen eine andere Maßnahme) im Vorhinein zu begründen (**a priori** Effizienzanalysen). Zum anderen kann man im Nachhinein beurteilen, ob sich die Maßnahme tatsächlich gelohnt hat (**a posteriori** Effizienzanalysen). Grundsätzlich lassen sich zwei effizienzanalytische Modelle unterscheiden: Wenn sich neben den --- Kosten auch die Wirksamkeit einer Maßnahme sinnvoll in Geldwerte umrechnen lässt, erlaubt dies eine Kosten-Nutzen-Analyse. Wenn sich lediglich die Kosten in Geldeinheiten überführen lassen, nicht aber die Wirksamkeit, kann man eine Kosten-Effektivitäts-Analyse durchführen.

Eichstichprobe. Die Eichstichprobe ist diejenige → Stichprobe, anhand derer die psychometrischen → Gütekriterien eines Messinstruments ermittelt sowie die Populationsverteilung der Messwerte geschätzt wird. Die Eichstichprobe sollte daher ausreichend groß, → repräsentativ und → probabilistisch sein.

Evaluationsdesign. Versuchsanordnung oder Untersuchungsplan, die/der erlaubt, die → Nettowirkung einer Maßnahme zu ermitteln und Alternativhypothesen für das Zustandekommen eines Effekts auszuschließen. Das häufigste Evaluationsdesign ist das → Split-Plot-Design. In diesem gibt es eine Interventionsgruppe sowie eine (oder mehrere) Kontrollgruppen, außerdem werden Werte zu mehreren Messzeitpunkten erhoben: Pre-Test, Post-Test, → Follow-up-Messung, usw.

experimentelle Mortalität/Drop-out. Viele Längsschnittuntersuchungen mit echter Messwiederholung leiden daran, dass Personen über die Messzeitpunkte hinweg verloren gehen, etwa weil sie bei einem Messzeitpunkt fehlen oder weil sie im Laufe der Zeit die Stichprobe verlassen haben. Solange ein solcher Drop-out unsystematisch ist, d. h. unabhängig von irgendwelchen Faktoren, die die Ergebnisse der Untersuchung beeinflussen können, reduziert sich lediglich die → Teststärke. Ist der Drop-out jedoch mit anderen Variablen korreliert (z. B. mit der Merkmalsausprägung im Pre-Test), leidet auch die → interne Validität der Untersuchung. In diesem Fall spricht man von selektivem Drop-out.

externe Evaluation/Fremdevaluation → interne Evaluation

externe Validität/ökologische Validität. Externe oder ökologische Validität bedeutet, dass ein Zusammenhang oder Effekt, der in einem Kontext besteht und dort ermittelt wurde (z. B. in einem Labor), auch in einem anderen Kontext (z. B. außerhalb des Labors) besteht und folglich auf diesen generalisiert werden kann.

Fehlerkumulierung. Wird eine inhaltliche Hypothese über ein Set inferenzstatistischer Tests geprüft, so stellt sich das Problem einer Kumulierung der → Irrtumswahrscheinlichkeiten. Mit jedem Test steigt die Wahrscheinlichkeit, sich mindestens einmal falsch zu entscheiden, also einen Fehler 1. oder 2. Art zu begehen. Die Kumulierung von Irrtumswahrscheinlichkeiten erhöht also das »tatsächliche« multiple α- bzw. β-Fehlerniveau und schwächt somit die Aussagekraft der statistischen Hypothesentestung. Daher sollte man die Anzahl statistischer Tests pro Hypothese wenn möglich reduzieren und ggf. die → Irrtumswahrscheinlichkeiten (α und β) strenger festsetzen.

Follow-up-Messung. Während der Pre-Test unmittelbar vor Beginn einer Maßnahme und der Post-Test unmittelbar nach Beendigung einer Maßnahme angesiedelt sind und damit eine Abschätzung der kurzfristigen Wirksamkeit erlauben, werden Follow-up-Erhebungen einige Zeit nach Beendigung der Maßnahme angesetzt. Im Idealfall geht aus dem Wirkmodell für eine Maßnahme hervor, welches sinnvolle Follow-up-Messzeitpunkte sein könnten. Follow-up-Erhebungen erlauben eine Abschätzung der mittel- und langfristigen Wirksamkeit einer Maßnahme (Persistenz).

formative Evaluation → summative Evaluation

Fremdevaluation/externe Evaluation → interne Evaluation

gegnerschaftsorientierte Evaluation/Adversary Evaluation. Zwei unabhängige Evaluatoren (oder auch Teams von Evaluatoren) vertreten unterschiedliche Positionen und tauschen Pro- bzw. Kontraargumente aus. Am Ende entscheidet eine Jury über das endgültige Vorgehen bei der Evaluation.

geschlossene Evaluation → offene Evaluation

Gütekriterien. Gütekriterien sind Maße oder Kennwerte, mit denen die Qualität eines Produkts erfasst wird. Im Kontext sozialwissenschaftlich-empirischer Untersuchungen unterscheidet man vier allgemeine Gütekriterien: (1) Komplexität, (2) Gültigkeitsbereich, (3) Objektivierbarkeit, (4) Transparenz. In der Psychometrie beziehen sich Gütekriterien vor allem auf die Qualität eines Messinstruments. Die psychometrischen Hauptgütekriterien sind die → Reliabilität, → Validität und → Objektivität eines Instruments. Die wichtigsten Gütekriterien für → Evaluationsdesigns sind die → interne und die → externe Validität.

Hawthorne-Effekt. Der Hawthorne-Effekt besagt, dass die empirisch nachgewiesene → Wirksamkeit einer Interventionsmaßnahme möglicherweise gar nichts mit dem Inhalt der Intervention zu tun hatte, sondern lediglich darauf zurückzuführen ist, dass mit den Personen überhaupt irgendetwas durchgeführt wurde bzw. dass eine Messung stattgefunden hat. Allein die Tatsache, dass etwas passiert, erhöht die Motivation der Teilnehmer, gute Leistungen zu zeigen, positive Werte abzugeben, sich besonders positiv und → »sozial erwünscht« zu verhalten.

Implementation und Implementationskontrolle. Mit Implementation ist die konkrete Vorbereitung und Durchführung einer Interventionsmaßnahme gemeint. Im Rahmen einer Implementationskontrolle wird geprüft, ob die Maßnahme auch tatsächlich so durchgeführt wird wie vorgesehen (Prüfung der → Ausführungsintegrität). Hierzu können Beobachtungsmethoden, Videoanalysen, aber auch Teilnehmerbefragungen und Supervisionsgespräche mit den Durchführenden angewendet werden.

Inputevaluation. Bei der Inputevaluation steht eine Bewertung der Rahmenbedingungen für die Durchführung einer Maßnahme im Vordergrund. Beispiele für mögliche Fragestellungen: Welche Ressourcen stehen zur Verfügung? Welche Qualität besitzen bspw. die verwendeten Unterrichtsmaterialien? Welche Zeitspanne steht zur Verfügung? Inputevaluationen werden typischerweise im Rahmen einer → prospektiven Evaluation durchgeführt.

interne Evaluation/Selbstevaluation vs. externe Evaluation/Fremdevaluation. Bei der internen Evaluation, auch Selbstevaluation genannt, führen diejenigen Personen, welche die Maßnahme umsetzen, auch die Evaluationsuntersuchung durch. Dies ist nicht unproblematisch, weil die Evaluatoren möglicherweise parteiisch sind und Eigeninteressen verfolgen. Im Falle der externen Evaluation, auch als Fremdevaluation bezeichnet, wird die Durchführung der Evaluation an eine andere Person oder Institution übergeben.

interne Validität. Ein → Evaluationsdesign wird als intern valide bezeichnet, wenn ein beobachteter Zusammenhang zwischen UV und AV auf eine kausale Wirkung der UV auf die AV zurückgeführt werden kann, und nicht etwa auf systematische Störvariablen, die mit der UV → konfundiert sind. Im Rahmen nicht-randomisierter, also → quasi-experimenteller Evaluationsdesigns stellen systematische Unterschiede zwischen einer behandelten und einer unbehandelten Gruppe (z. B. in der Motivation, an der Maßnahme teilzunehmen) eine substanzielle Gefahr für die interne Validität dar.

intraindividuelles Design. Von einem intraindividuellen Design spricht man, wenn die gleichen Personen zu unterschiedlichen Messzeitpunkten oder in unterschiedlichen experimentellen Bedingungen wiederholt getestet bzw. beobachtet werden (→Veränderungsmessung). Der Vorteil eines messwiederholten Designs ist, dass stabile Unterschiede zwischen Personen kontrolliert bzw. → aus- partialisiert werden können. Dies reduziert die Fehlervarianz und erhöht die → Teststärke. Der Nachteil ist, dass die → Veränderungsmessung aufgrund von Artefakten (z. B. → Konfirmationseffekt, Erinnerungseffekte, → Regressionseffekt) verzerrt sein kann.

intrinsische vs. extrinsische Evaluation. Unter einer intrinsischen Evaluation versteht man eine Bewertung der inneren Struktur der Maßnahme, z. B. inwiefern das Wirkmodell einer Interventionsmaßnahme theoretisch sinnvoll begründet ist. Intrinsische Evaluationen sind damit quasi →Konzeptionsanalysen; sie finden im Rahmen einer → prospektiven Evaluation statt. Unter extrinsischer Evaluation versteht man die Analyse der Effekte einer Maßnahme (Wirksamkeitsevaluation).

Inzidenz. Inzidenz ist ein Begriff aus der Epidemiologie. Die Inzidenz gibt die Anzahl der Neuerkrankungen innerhalb eines bestimmten Zeitraumes (meist eines Jahres) an. Die **kumulative** Inzidenz gibt die Anzahl der innerhalb einer Periode neu Erkrankten relativiert an der Menge aller zu Beginn der Periode Untersuchten an. **Spezifische** Inzidenzen kann man dann berechnen, wenn sich die Inzidenzraten in Abhängigkeit von bestimmten demografischen Merkmalen (z. B. Alter, Herkunft, Geschlecht) unterscheiden.

Irrtumswahrscheinlichkeit. Im Rahmen eines inferenzstatistischen Tests wird eine Entscheidung über die Ablehnung der statistischen Nullhypothese bzw. der statistischen Alternativhypothese getroffen. Diese Entscheidungen sind jedoch immer mit einem spezifischen Fehler behaftet. Die Wahrscheinlichkeit, sich fälschlicherweise gegen die Nullhypothese zu entscheiden, wird als **α-Fehler** oder **Fehler 1. Art** bezeichnet. Die Wahrscheinlichkeit, sich fälschlicherweise gegen die Alternativhypothese zu entscheiden, wird als **β-Fehler** oder **Fehler 2. Art** bezeichnet. Üblicherweise sollen beide Fehlerwahrscheinlichkeiten nicht mehr als 5 % betragen. Ist die Wahrscheinlichkeit, ein empirisch beobachtetes oder jedes noch stärker gegen die Nullhypothese sprechendes Ergebnis zu erhalten, unter der Null- hypothese (bzw. unter der Alternativhypothese) kleiner als 5 %, so wird die entsprechende Hypothese abgelehnt; das Ergebnis ist statistisch bedeutsam, d. h. signifikant.

Konfirmationseffekt/Confirmation Bias. Subjektive Einschätzungen sind stets fehlerbehaftet bzw. anfällig für Verzerrungen. Eine dieser Verzerrungen ist der Konfirmationseffekt. Ein bestimmter Gegenstand wird subjektiv so wahrgenommen, wie man es erwartet hat, weil man diejenigen Informationen stärker gewichtet, die die Erwartung bestätigen. Mit einem Konfirmationseffekt ist bspw. im Rahmen einer direkten → Veränderungsmessung zu rechnen: Man nimmt die Veränderung so wahr, wie man sie erwartet hat oder wie man sie gerne hätte.

Konfundierung. Ist eine Störvariable sowohl mit der AV als auch mit der UV korreliert, so liegt eine Konfundierung vor. Die → interne Validität der Untersuchung ist bedroht; ein beobachteter Zusammenhang zwischen UV und AV kann nicht mehr kausal interpretiert werden. Störvariablen können dazu führen, dass die Hypothese artifiziell bestätigt wird (gleichsinnige Konfundierung) oder dass sie artifiziell widerlegt wird (gegensinnige Konfundierung).

Konzeptionsanalyse. Die Konzeptionsanalyse ist im Rahmen einer → prospektiven Evaluation angesiedelt. Sie fragt danach, ob die Maßnahme auch tatsächlich dem Bedarf angepasst ist, ob das der Maßnahme zugrunde liegende Wirkmodell plausibel ist (→ Inputevaluation) und ob die Rahmenbedingungen für eine wirksamkeitsförderliche Durchführung der Maßnahme gegeben sind bzw. geschaffen oder verbessert werden müssen.

Kosten. Kostenabschätzungen werden im Rahmen von → Effizienzanalysen relevant. Die Kosten einer Maßnahme können manifest oder latent sein. **Manifeste** Kosten werden zwischen dem Auftraggeber und dem Auftragnehmer ausgetauscht; es handelt sich um Geld, das tatsächlich bezahlt wird, z. B. für Personal, für Material, für Raummieten. **Latente** Kosten können zumeist nicht direkt belegt werden. Hierzu gehören staatliche Versorgungsmaßnahmen, die über Steuereinnahmen gegenfinanziert werden, oder Kosten, die von anderen Kostenträgern beigesteuert werden, ohne dass dies nach außen sichtbar wird.

Management Information System (MIS). Im Rahmen von → formativen Evaluationen werden u. a. begleitende Prozessevaluationen durchgeführt (Programm-Monitoring). Die Daten solcher Prozessevaluationen können dazu verwendet werden, den Auftraggebern in regelmäßigen Abständen Rückmeldung über neue Entwicklungen, zu beobachtende Veränderungen, strukturelle Hindernisse usw. zu geben und konkrete Maßnahmen zur Programmoptimierung vorzuschlagen. In einigen Bereichen hat sich für ein solches bedarfsorientiertes Monitoring-System der Begriff MIS eingebürgert.

MAUT-Technik. Die MAUT-Technik ist eine Linearkombination zur rationalen Bestimmung des Nutzens einer Maßnahme im Rahmen einer → Effizienzanalyse. Der Gesamtnutzen einer Maßnahme (N) entspricht der Summe aller mit einem (subjektiven oder objektiven) Wert W gewichteten Effekte E. Dieser Wert lässt sich heranziehen, um sowohl die Nutzenerwartungen unterschiedlicher Beteiligtengruppen als auch den Gesamtnutzen unterschiedlicher Maßnahmen direkt miteinander zu vergleichen.

Metaevaluation. Sie zielt darauf ab, auf der Basis von Einzelevaluationen über die gleiche Maßnahme eine Generalisierung der Aussagen über die →Wirksamkeit oder über die → Effizienz der Maßnahme zu erreichen. Dies kann entweder über die statistische Methode der Metaanalyse (summative Metaevaluation) oder über eine Programm-Design-Evaluation geschehen.

Makro- vs. Mikroevaluation. Bei der Makroevaluation wird der gesamte Evaluationsgegenstand (z. B. ein Programm oder eine Interventionsmaßnahme) umfassend bewertet. Bei einer Mikroevaluation werden lediglich einzelne Aspekte des Evaluationsgegenstands evaluiert.

Moderatorvariable. Dies sind Variablen, die die Stärke bzw. die Richtung des Effekts einer UV auf die AV beeinflussen. Moderatoreffekte sind Interaktionseffekte zwischen der Moderatorvariable und einer UV auf eine AV.

multimodale und multimethodale Diagnostik. Evaluationskriterien können, insbesondere wenn sie nicht direkt beobachtbar sind, auf verschiedene Arten → operationalisiert werden. Mit multimodaler Diagnostik ist gemeint, dass alle Modalitäten eines Konstrukts erfasst werden sollen, d. h. Bereiche, in denen sich das Merkmal manifestiert, z. B. Kognitionen, Emotionen, Ausdrucksverhal- ten, physiologische Reaktionen. Mit multimethodaler Diagnostik ist gemeint, dass zur Messung unterschiedliche Methoden zur Anwendung kommen sollen, z. B. Selbstbeschreibung, Fremdbeschreibung, indirekte Messung, Beobachtung. Aggregiert man Maße über Modalitäten und Methoden

hinweg, so kann man das latente Merkmal messfehlerfreier (reliabler; → Reliabilität) und valider (→ Validität) erfassen, als es mit einem einzigen Indikator jemals möglich wäre.

Netto- und Bruttowirkung. Interventionsmaßnahmen können Effekte haben, die auf eine Wirkung des »Hauptwirkstoffs« einer Maßnahme zurückgehen (maßnahmenspezifische Wirkung). Es kann aber auch sein, dass beobachtete Wirksamkeitseffekte auf Faktoren zurückgehen, die gar nichts mit der eigentlichen Maßnahme zu tun haben. Dazu gehören unbeabsichtigte Neben- und Folgewirkungen der Maßnahme, unspezifische Effekte sowie externe Effekte aufgrund von → Konfundierungen. Die Gesamtheit aller Wirkungen wird als Bruttowirkung, die maßnahmenspezifischen Haupt-, Neben- und Folgewirkungen werden als Nettowirkung einer Intervention bezeichnet.

Norm bzw. Normierung. Unter Normen versteht man soziale Erwartungen oder empirisch feststellbare Gesetzmäßigkeiten. In der Psychometrie sind mit Normen Vergleichsdaten zur Beurteilung von Einzeldaten gemeint. Um die Werte einer Person besser interpretieren zu können, kann man den Wert einer Person mit Normwerten vergleichen, d. h. mit der Verteilung der Werte aus der → Eichstichprobe, z. B. indem man die Einzelwerte in → Standardwerte transformiert oder indem man einen → Prozentrang angibt.

Objektivität. Messinstrumente sind objektiv, wenn das Ergebnis einer Messung unabhängig davon ist, wer sie vorgenommen hat und unter welchen Bedingungen die Messung stattfand (**Durchführungsobjektivität**), wie und von wem die Testwerte ausgewertet wurden (**Auswertungsobjektivität**) und wer die Testwerte interpretiert (**Interpretationsobjektivität**).

offene Evaluation vs. geschlossene Evaluation. Das Begriffspaar wird in zwei Bedeutungen verwendet: (1) Im Falle einer geschlossenen Evaluation ist die Fragestellung des Evaluationsvorhabens bereits im Vorhinein genau definiert. Bei einer offenen Evaluation sind die Bestimmung der Fragestellung, der Methoden und der Hypothesen selbst Gegenstand des Evaluationsprozesses. (2) Wird eine Evaluation offen durchgeführt, so ist sie für die Beteiligten, Betroffenen sowie die Öffentlichkeit transparent. Bei einer geschlossenen Evaluation sind die Ergebnisse nur für den Auftraggeber bestimmt.

Operationalisierung. Der Begriff wird in zwei verwandten Bedeutungen verwendet. (1) Definition eines psychologischen Merkmals anhand von objektiv beobachtbaren Anzeichen einschließlich der Festlegung von Regeln, nach denen die Beobachtung geschieht. (2) Definition einer Versuchsbedingung anhand von objektiv beschreibbaren Merkmalen, die nach festgelegten Regeln zwischen den Bedingungen variiert werden.

Parallelisierung (»Matching«). Die Parallelisierung ist eine Möglichkeit, den Einfluss von Störvariablen zu kontrollieren und damit die → interne Validität einer Evaluationsuntersuchung zu erhöhen. Wenn die Ausprägungen der Störvariablen in allen Stufen der UV gleich verteilt sind, ist sie keine plausible Alternativerklärung mehr für das Zustandekommen eines Effekts (→ Ausbalancierung). Möglichkeiten der Parallelisierung können danach unterschieden werden, ob es sich um (a) eine univariate oder multivariate, (b) eine Durchschnitts- oder eine individuelle, (c) eine exakte oder eine näherungsweise und (d) eine »Eins-zu-eins« oder eine »Eins-zu-*n*«-Parallelisierung handelt. Eine Form der multivariaten Parallelisierung, die sich besonders bei quasi-experimentellen Designs mit Selbstselektion eignet, ist das Propensity Score Matching.

Perzentil/Prozentrang. Das Perzentil/der Prozentrang eröffnet die Möglichkeit, einen Einzelwert unter der Verteilung aller möglichen Werte vergleichend zu bewerten (→ Normierung). Wenn das Merkmal in der Population normalverteilt ist, ist die Bewertung eines Einzelwertes im Vergleich zu Mittelwert und Standardabweichung einer Eichstichprobe (z. B. über einen → *z*-Wert) sinnvoll. Bei nicht normalverteilten Merkmalen bietet es sich stattdessen an, den Prozentrang/das Perzentil eines Einzelwertes anzugeben. Mit dem Prozentrang kann man für einen beliebigen Wert (unter allen möglichen Werten) angeben, wie viel Prozent aller anderen Fälle den gleichen oder einen kleineren Wert haben.

Power → Teststärke

Prävalenz. Die Prävalenz gibt die Anzahl all jener Indi- viduen in einer Population (oder Stichprobe) an, die zu einem bestimmten Zeitpunkt (**Punktprävalenz**) oder während einer bestimmten Zeitperiode (**Periodenprävalenz**) als positiv (z. B. »krank«) diagnostiziert wurden. Relativiert an der Gesamtzahl aller erfassten Individuen spricht man von der **Prävalenzrate**. Die **Lebenszeitprävalenz**

ist eine Sonderform der Periodenprävalenz, da die Periode eine gesamte Lebensspanne umfasst. Sie gibt die Wahrscheinlichkeit an, mit der ein Mensch mindestens einmal im Leben positiv diagnostiziert wird.

Prävention. Eine Maßnahme wird als präventiv bezeichnet, wenn sie dazu beitragen soll, ein möglicherweise auftretendes Problem im Vorhinein zu verhindern (**Primärprävention**), die Belastung eines bereits existierenden Problems zu verringern (Sekundärprävention) oder infolge des Problems bereits entstandene Schäden zu beheben (**Tertiärprävention**). Ursprünglich stammt diese Trichotomie von Caplan (1964). Heute wird der Ansatz einer Präventionsmaßnahme anstatt über ihr Ziel eher über die Zielgruppe definiert, an die sich die Maßnahme richtet (Gordon, 1983). Primärpräventive Maßnahmen richten sich an die gesamte Population (universeller Ansatz) oder an Risikogruppen (selektiver Ansatz). Sekundärpräventive maßnahmen richten sich an diejenigen personen, die ein problematisches Verhalten bereits ausgebildet haben (indizierter Ansatz).

probabilistische Stichprobe vs. nicht-probabilistische Stichprobe. Eine → Stichprobe wird als probabilistisch bezeichnet, wenn die Wahrscheinlichkeit, mit der ein beliebiges Element der Population in die Stichprobe »gezogen« wird, bekannt oder zumindest prinzipiell kontrollierbar ist. Für die Theorie des statistischen Hypothesentestens spielt es eine entscheidende Rolle, ob die Stichprobe tatsächlich probabilistisch ist. In der Praxis ist es so, dass nicht-probabilistische Stichproben die Generalisierung eines empirischen Befundes auf die Population lediglich erschweren (→ externe Validität).

prospektive Evaluation. Als prospektive Evaluation bezeichnet man die Prüfung bzw. Schaffung von Voraussetzungen und Rahmenbedingungen, welche die Wirksamkeit einer geplanten Maßnahme positiv beeinflussen bzw. sichern sollen. Es geht also darum, die Maßnahme unter den gegebenen Bedingungen zu bewerten, bevor sie implementiert wird. Im Allgemeinen sollen im Rahmen einer prospektiven Evaluation zwei Fragen geklärt werden: (1) Besteht Bedarf an einer Maßnahme (→Bedarfsanalyse)? (2) Wird eine konkrete Intervention diesem Bedarf gerecht (→ Konzeptionsanalyse)?

Prozentrang → Perzentil

Pygmalioneffekt → Rosenthaleffekt

quasi-experimentelles Design. Ist die Zuweisung von Personen zu experimentellen Bedingungen (z. B. Interventions- oder Kontrollgruppe) weder zufällig (→ randomisiert) noch vom Evaluator kontrolliert, so sind die Voraussetzungen für ein echtes Experiment nicht erfüllt. Das Problem besteht darin, dass bei quasi-experimentellen Designs die Gefahr möglicher Konfundierungen nicht restlos ausgeschlossen werden kann: Man kann nicht ausschließen, dass es noch weitere systematische Störvariablen gibt, die ungleich über die Bedingungen hinweg verteilt sind. Deshalb sind quasi-experimentelle Designs echten experimentellen Designs mit → Randomisierung meist methodisch unterlegen.

quasi-indirekte Veränderungsmessung/Retrospective Pretest. Hierbei werden die Daten für den ersten Messzeitpunkt (Pre-Test) retrospektiv zum zweiten Messzeitpunkt (Post-Test) erhoben. Die Probanden werden also beim Post-Test gebeten, sich noch einmal in ihre Situation zum Zeitpunkt t_1 zu versetzen und die Merkmalsausprägung zu diesem Zeitpunkt zu beurteilen. Anschließend berechnet man die Differenz zwischen der (zu t_2 retrospektiv eingeschätzten) Ausprägung des Merkmals zu t_1 und der aktuellen Merkmalsausprägung zu t_2. Man nutzt damit quasi die Vorteile eines intraindividuellen Designs und kann einige Probleme der indirekten sowie der direkten →Veränderungsmessung umgehen.

Randomisierung. Als Randomisierung bezeichnet man die zufällige Zuweisung von Versuchspersonen zu Versuchsbedingungen in einem Experiment. Dadurch soll sichergestellt werden, dass sich die Versuchsgruppen nur in den für die Fragestellung relevanten Merkmalen unterscheiden, nicht aber in irrelevanten Merkmalen. Würden sich die Versuchsgruppen auch in irrelevanten Merkmalen (Störvariablen) unterscheiden, könnten die Verhaltensunterschiede zwischen den Gruppen nicht mehr zweifelsfrei auf die Versuchsbedingungen zurückgeführt werden. Der Versuch wäre dann nicht mehr intern valide (→ interne Validität).

Regressionseffekt/Regression zur Mitte. Im Rahmen einer indirekten → Veränderungsmessung mit Hilfe eines → intraindividuellen Designs kann es sein, dass Werte, die zum Messzeitpunkt t_1 weit vom Mittelwert abweichen, zum Messzeitpunkt t_2 weniger weit vom Mittelwert abweichen. Das glei-

che gilt für Werte, die zum Messzeitpunkt $t_2$weit vom Mittelwert abweichen: Sie weichen zu t_1 weniger weit vom Mittelwert ab. Insofern sieht es aus, als habe eine Veränderung stattgefunden; in Wirklichkeit geht der Effekt jedoch lediglich auf ein mathematisches Artefakt zurück, das entsteht, wenn (1) die Messung zu beiden Zeitpunkten jeweils unreliabel ist (→ Reliabilität), (2) die Messwerte zu beiden Zeitpunkten gleiche Varianzen haben und (3) die Messwerte zu beiden Zeitpunkten hoch miteinander korreliert sind.

Regressions-Diskontinuitäts-Design (RD-Design). Design, das eine hohe interne Validität besitzt, obwohl die Zuweisung zu einer Therapiebedingung (z. B. alte vs. neue Therapie) nicht per Zufall (→ Randomisierung), sondern auf der Basis einer Kontrollvariablen vorgenommen wird. Dabei handelt es sich um einen Indikator der Behandlungsbedürftigkeit (z. B. Symptomschwere zu Beginn einer Therapie). Auf dieser Kontrollvariablen wird ein Cut-off-Wert definiert: Alle Personen oberhalb des Cut-offs erhalten die Therapie, Personen unterhalb des »Cut-offs« werden bspw. einer Wartekontrollgruppe zugewiesen. Wenn die Therapie wirksam ist, sollte es am Cut-off eine Diskontinuität (einen »Sprung«) in der Regressionsfunktion geben, welche den Zusammenhang zwischen der Kontrollvariablen und der AV beschreibt. Das Ausmaß der Regressionsdiskontinuität beschreibt den Effekt der neuen Therapie.

Reliabilität. Unter Reliabilität versteht man die Zuverlässigkeit (im Sinne von »Messfehlerfreiheit«) eines Messinstruments. Nach der Klassischen Testtheorie ist die Reliabilität eines Messinstruments dadurch definiert, inwiefern die Werte – neben der Tatsache, dass sie das zu messende (latente) Merkmal indizieren – durch unsystematische Messfehler beeinflusst sind. Reliabilität ist die Voraussetzung für die → Validität eines Messinstruments. Eine Möglichkeit der empirischen Quantifizierung der Reliabilität ist die Re-Test-Methode: Ein Messinstrument misst ein stabiles Merkmal zuverlässig, wenn es bei wiederholten Messungen den gleichen Messwert liefert.

Repräsentativität. Eine Interventions- oder Evaluations- stichprobe kann für die Population, über die Aussagen getroffen werden soll, mehr oder weniger repräsentativ sein. Dabei ist Repräsentativität nicht gleichbedeutend mit einer → probabilistischen Stichprobe! Ist die Stichprobe hinsichtlich aller Merkmale repräsentativ für die Population, spricht man von globaler Repräsentativität. Ist die Stichprobe hinsichtlich bestimmter Merkmale repräsentativ für die Population, spricht man von spezifischer Repräsentativität.

Residual Change Scores → auto-residualisierte Variable

Response Shift. Mit Response Shift ist gemeint, dass sich zwischen den Messzeitpunkten das Verständnis des Messinstruments bzw. die subjektive Repräsentation des zu messenden Merkmals verändert hat, nicht jedoch die Merkmalsausprägung selbst. Beispiel: Ein hochbegabtes Kind, das gerade auf eine spezielle Förderschule versetzt wurde, hält sich zu Beginn des Schuljahres für extrem intelligent, am Ende des Schuljahres für weniger intelligent. Diese Veränderung ist nicht auf eine »wahre« Verringerung der Intelligenz, sondern eher auf den Wechsel des sozialen Vergleichsstandards zurückzuführen: Nach einem Jahr auf der Förderschule vergleicht sich der Schüler mit anderen Hochbegabten; das beeinflusst seine Selbsteinschätzung – nicht aber seine Intelligenz!

Retrospective Pretest → quasi-indirekte Veränderungsmessung

Rosenthal- oder Pygmalioneffekt. Als Rosenthal- oder Pygmalioneffekt bezeichnet man die unbewusste Beeinflussung der Ergebnisse einer Untersuchung durch Erwartungen des Versuchs- bzw. Testleiters. Ohne es zu wollen und zu wissen, verhalten sich Versuchsleiter in psychologischen Untersuchungen häufig so, dass das Verhalten der Versuchspersonen den Hypothese oder Erwartungen des Versuchsleiters entspricht. Robert Rosenthal, der dieses Phänomen systematisch untersuchte, bezeichnete es als Pygmalioneffekt – nach dem sagenhaften griechischen Bildhauer, der sich in eine von ihm geschaffene Mädchengestalt verliebte.

Schwierigkeit. Schwierigkeit ist ein psychometrisches → Gütekriterium eines Messinstruments, das die Wahrscheinlichkeit angibt, mit der eine beliebige, zufällig aus der Population gezogene Person einen bestimmten Wert erzielt. Ein Messinstrument ist zu schwierig, wenn es so konstruiert ist, dass nur sehr wenige Personen hohe Werte erreichen können, z. B. eine unlösbare Aufgabe in einem Intelligenztest. Ein Messinstrument ist zu leicht, wenn es so konstruiert ist, dass nahezu alle Personen hohe Werte erreichen können, z. B. eine sehr einfache Aufgabe in einem

Intelligenztest. Extrem hohe oder extrem niedrige Schwierigkeiten gehen mit schiefen Häufigkeitsverteilungen einher; die Merkmale haben in der Stichprobe meist eine geringe Varianz. Das beeinflusst auch die → Reliabilität, die → Validität und die →Änderungssensitivität der Messung.

Selbstevaluation → Interne Evaluation

Selbstselektion. Bei der Selbstselektion entscheiden die Personen in der Interventions- bzw. Evaluationsstichprobe selbst, welcher experimentellen Bedingung sie zugeteilt werden. Dies ist oft gegeben, wenn die Teilnahme an einer Maßnahme, deren Wirkung evaluiert werden soll, freiwillig war. Das Problem besteht darin, dass die Selbstzuweisung zu Bedingungen nicht unabhängig von Störvariablen bzw. von → Konfundierungen ist. Aufgrund von Selbstselektion resultiert ein → quasi-experimentelles Design, dessen → interne Validität äußerst gering ist.

Simpson-Paradox. Das Simpson-Paradox beschreibt einen sehr speziellen Fall der → Konfundierung. Das Paradox besteht darin, dass ein empirisches Ergebnis innerhalb von bestimmten Subpopulationen (z. B. Ausprägungen einer Störvariablen) hypothesenkonform ist, über alle Subpopulationen hinweg jedoch sein Vorzeichen umkehrt und hypothesenwidrig ist.

soziale Erwünschtheit. Wenn man das eigene Verhalten oder die Beschreibung der eigenen Person an soziale Normen anpasst, um in den Augen anderer möglichst unauffällig, angepasst oder attraktiv zu erscheinen, handelt man sozial erwünscht. Sozial erwünschtes Verhalten und sozial erwünschte Antworten in Fragebögen resultieren aus dem Anerkennungsbedürfnis, das in seiner Ausprägung interindividuell variiert. Soziale Erwünschtheit ist somit eine Persönlichkeitseigenschaft, die bei starker Ausprägung die Aussagekraft von Selbstbeschreibungen schmälert. Sie verringert die → Validität einer Messung.

Sphärizitäts- bzw. Zirkularitätsannahme. Bei Varianzanalysen mit Messwiederholung wird angenommen, dass die Varianzen aller paarweisen Differenzen zwischen zwei Messzeitpunkten auf Populationsebene homogen sind. Sind sie das nicht, so wird der *F*-Test für den Haupteffekt des Messzeitpunkts zu liberal, d. h. sein α-Fehler ist höher als das festgesetzte Signifikanzniveau. In diesem Fall können die Freiheitsgrade des *F*-Test für den Haupteffekt des Messzeitpunkts mit einem Parameter namens Greenhouse-Geisser-Epsilon gewichtet werden. Diese Korrektur macht den *F*-Test wieder konservativer und wirkt damit der künstlichen Erhöhung des α-Fehlers tendenziell entgegen.

Split-Plot-Design. Das Split-Plot-Design ist der am häufigsten verwendete Versuchsplan im Rahmen von Wirksamkeitsevaluationen. Der Plan besteht im einfachsten Fall aus zwei Messzeitpunkten (Pre-Test und Post-Test) und zwei Interventionsbedingungen (Interventions- und Kontrollgruppe). Die Daten werden – falls die entsprechenden Voraussetzungen erfüllt sind – mit Hilfe einer zweifaktoriellen Varianzanalyse mit Messwiederholung ausgewertet. Hypothesenrelevant ist der *F*-Test für die Wechselwirkung zwischen Messzeitpunkt und Interventionsbedingung: In der Interventionsgruppe sollten sich die Werte in der erwarteten Richtung verändern, in der Kontrollgruppe nicht.

Standards. Ein Standard ist eine allgemein als richtig anerkannte Regel in Bezug auf eine Tätigkeit, ein Vorgehen, eine Handlung, eine Kompetenz. Für die Evaluationsforschung wurden in den USA bereits 1982 erste Standards formuliert. Die Deutsche Gesellschaft für Evaluation (DeGEval) hat Evaluationsstandards veröffentlicht, die sich an den Vorgaben aus den USA orientieren. Dabei werden vier Kategorien von Evaluationsstandards unterschieden; sie beziehen sich auf (1) die Nützlichkeit, (2) die Durchführbarkeit, (3) die Fairness und (4) die Genauigkeit einer Evaluationsuntersuchung.

Standardwert/*z*-Wert. Zieht man von einem Messwert den Mittelwert aller Werte ab und teilt diese Differenz durch die Standardabweichung des Merkmals, so erhält man einen Standardwert, auch *z*-Wert genannt. Die *z*-Standardisierung erlaubt es, einen Wert unter der Verteilung aller möglichen Werte zu beurteilen (→ Normierung). Ist das Merkmal zudem normalverteilt, kann die Wahrscheinlichkeit berechnet werden, mit der eine zufällig aus der Population gezogene Person diesen (oder einen größeren bzw. einen kleineren) Wert hat. Je kleiner diese Wahrscheinlichkeit, desto eher handelt es sich um eine »auffällige« oder extreme Merkmalsausprägung. Ist das Merkmal nicht normalverteilt, bietet sich stattdessen die Angabe eines → Prozentranges an.

Stichprobe. Eine Stichprobe ist eine endliche Menge von Elementen, die aus einer Population gezogen werden. Unter bestimmten Voraussetzungen (→ Repräsentativität) kann man empirische Ergebnisse, die anhand einer Stichprobe gewonnen wurden, auf die gesamte Population beziehen. Man unterscheidet zwischen → probabilistischen und nicht-probabilistischen Stichproben. Bei letzteren ist eine Schlussfolgerung auf die Population nur eingeschränkt möglich.

summative vs. formative Evaluation. Summative Evaluation verfolgt das Ziel, die → Wirksamkeit eines Programms zu beurteilen. Formative Evaluation verfolgt das Ziel, die Programmdurchführung zu optimieren und die Programmkonzeption zu verbessern. Sie setzt in der Phase der Vorbereitung und der Durchführung eines Programms an und richtet sich an diejenigen Personen, die mit der Programmkonzeption und -durchführung befasst sind, z. B. Autoren, Trainer, Therapeuten, Supervisoren.

Teststärke/Power. Die Begriffe Teststärke bzw. Power stammen aus der Theorie des statistischen Hypothesentestens von Neyman und Pearson (1933). Die Teststärke eines statistischen Tests ist definiert als die Wahrscheinlichkeit, mit der dieser Test ein signifikantes Ergebnis produziert, falls in der Population ein Effekt einer spezifizierten Größe tatsächlich existiert. Formal ist sie definiert als $1 - \beta$ (→ Irrtumswahrscheinlichkeiten). Die Teststärke hängt ab von (1) dem Stichprobenumfang (n), (2) dem festgelegten α-Fehlerniveau und (3) der Größe des spezifizierten Effekts (→ Effekstärke). Auch die → Reliabilität der AV wirkt sich auf die Teststärke aus.

Transfer. Von einer Interventionsmaßnahme wird im Allgemeinen erwartet, dass sie nicht nur kurzfristig und sehr spezifisch wirkt; vielmehr sollen die Effekte nachhaltig sein (Persistenz) und sich auch auf andere Fähigkeitsbereiche generalisieren. Eine solche Generalisierung wird als Transfer bezeichnet. Hager und Hasselhorn (2000) unterscheiden zwei Arten von Transfer. Mit Anforderungstransfer ist die Generalisierung der erworbenen Fähigkeiten auf andere Aufgabenanforderungen gemeint: Wer die Prinzipien des Einmaleins für einstellige Zahlen verstanden hat, der kann diese auch auf mehrstellige Zahlen anwenden. Mit Situationstransfer ist die Generalisierung der erworbenen Fähigkeiten auf gleichartige Situationen außerhalb der Intervention gemeint: Wer das Einmaleins in der Schule gelernt hat, der soll es auch an der Supermarktkasse anwenden können.

Treatment Fidelity → Ausführungsintegrität

Trendanalyse. Die Form der Veränderung über die Zeit hinweg kann über Trends (oder Wachstumsgradienten) bestimmt werden. Dabei handelt es sich um eine Regressionsgleichung, bei der die Zeit (operationalisiert über Messzeitpunkte $t_0 \ldots t_i$) die Prädiktorvariable (t) und die Merkmalsausprägung die Kriteriumsvariable (y) darstellt. Somit können bestimmte Trendformen vergleichend getestet werden: Ein **linearer Trend** kann über die allgemeine Gleichung $\hat{y} = b_0 + b_1 \cdot t$ ermittelt werden. Ein **quadratischer Trend** kann über die allgemeine Gleichung $\hat{y} = b_0 + b_1 \cdot t + b_2 \cdot t_2$ ermittelt werden usw. Trends lassen sich entweder für die Werte einer einzelnen Person, für Mittelwerte, aber auch für Häufigkeiten berechnen. Im Falle eines echten → intraindividuellen Designs enthält die Regressionsgleichung zusätzlich eine personspezifische Kovariate (Personeffekt), die nichts anderes ist als der Mittelwert einer Person über alle Messzeitpunkte hinweg.

Validität (von Messinstrumenten). Im psychometrischen Kontext betrifft die Validität die Frage, inwiefern ein psychologisches Messverfahren tatsächlich das misst, was es messen soll (→ Gütekriterien). Die Validität kann über den Zusammenhang mit so genannten Validierungskorrelaten empirisch bestimmt werden. Dabei kann es sich um zukünftiges Verhalten (**prädiktive** Validität), andere Messinstrumente für das gleiche Merkmal (**konkurrente** Validität) oder externe Kriterien für dieses Merkmal (**Kriteriums**validität) handeln. Eine spezielle Form der Validität, die **Konstrukt**validität, basiert auf zwei Voraussetzungen, die gleichzeitig gegeben sein müssen: Das Messinstrument soll mit anderen Instrumenten, die mit Hilfe einer anderen Methode das gleiche Merkmal messen, hoch korrelieren (**konvergente** Validität) und es soll mit anderen Instrumenten, die mit Hilfe der gleichen oder einer anderen Methode ein anderes Merkmal messen, niedrig korrelieren (**diskriminante** Validität).

Varianzhomogenität. Viele statistische Testverfahren basieren auf der Voraussetzung, dass die »Innerhalb-Varianzen« in allen experimentellen Bedingungen auf Populationsebene gleich sind. Ist dies nicht der Fall, d. h. sind die Varianzen zwischen den Bedingungen heterogen, so ist das Testergebnis nicht mehr robust. Eine Möglichkeit,

die Varianzhomogenitätsannahme in der Stichprobe zu überprüfen, besteht in der Anwendung des Levene-Tests. Zeigt der Levene-Test an, dass die Nullhypothese (gleiche Varianzen) verworfen werden muss, so ist die Varianzhomogenitätsannahme verletzt. Dies wirkt sich auf die Robustheit des Tests vor allem dann aus, wenn die Stichprobengrößen klein und zwischen den experimentellen Bedingungen nicht gleich sind.

Veränderungsmessung. Veränderungsmessungen bezeichnen empirische Ermittlungen von Veränderungen in Merkmalsausprägungen über die Zeit hinweg. Bei der **direkten** Veränderungsmessung sollen die Befragten selbst die Veränderung einschätzen. Bei der **indirekten** Veränderungsmessung wird die Veränderung über die Differenz zwischen der Merkmalsausprägung zu zwei (oder mehreren) Messzeitpunkten erfasst (→ intraindividuelles Design). Eine dritte Form der Veränderungsmessung wird als quasi-indirekt bezeichnet (Retrospective Pretest, → quasi-indirekte Veränderungsmessung).

Wirkung und Wirksamkeit. Die Wirksamkeit einer Interventionsmaßnahme gilt als empirisch gesichert, wenn die Effekte, die mit der Maßnahme intendiert waren, auch empirisch beobachtbar sind. Die Wirkung einer Interventionsmaßnahme bezieht sich auf die spezifischen Wirkmechanismen, die zu den beobachtbaren Effekten geführt haben. Die Prozesse, die kausalen Bedingungen, die moderierenden Bedingungen oder die Konsequenzen der Wirksamkeit stehen bei bloßen Wirksamkeitsanalysen im Hintergrund. Eine Analyse der Wirkung einer Maßnahme hingegen erfordert ein Wirkmodell, wenn sie hypothesengeleitet durchgeführt werden soll.

Zirkularitätsannahme → Sphärizitätsannahme

z-Wert → Standardwert