Interventionsdesign, Evaluationsverfahren und Wirksamkeit von Interventionen

Dorota Reis und Laurenz L. Meier

1 Einleitung

Wenn eine Intervention im organisationalen Kontext geplant wird, müssen zwei Welten vereint werden: der Anspruch einer methodisch einwandfreien Umsetzung der Maßnahme und pragmatische Überlegungen. Ausgehend von der umfangreichen Literatur der Interventionsforschung könnte zunächst ein bewährter „Goldstandard" umgesetzt werden (Wadhwa und Cook 2019). Das würde bedeuten, dass die Zielgruppe, die Dauer, die Dosis und die Implementierung einer solchen Maßnahme sich an evidenzbasierten Ergebnissen aus der Literatur orientieren müssten. Die Teilnehmenden müssten zu mindestens zwei Gruppen randomisiert oder parallelisiert werden. Die Teilnehmenden in den Gruppen würden sich darin unterscheiden, ob sie sofort mit der Umsetzung der Maßnahme starten (Interventionsgruppe) oder als Kontrollgruppe entweder eine andere Aktivität implementieren (aktive Kontrollgruppe), oder aber zu einem späteren Zeitpunkt am Training teilnehmen (Wartekontrollgruppe). Einer Vor- und Nachbefragung müsste im Rahmen des „Goldstandards" eine Follow-up Befragung folgen, um die Stabilität der erzielten Veränderungen zu bewerten. Weitere beachtenswerte Aspekte betreffen beispielsweise die Kontrolle von Störvariablen oder die „blinde" (d. h. ohne Kenntnis der Gruppenzugehörigkeit) Auswertung der Daten („triple blind", Miller und Stewart 2011).

Zahlreiche Aspekte der Umsetzung von gesundheitsfördernden Interventionen müssen sich jedoch an pragmatischen Gesichtspunkten orientieren – bzw. erfordern eine Anpassung des „Goldstandards" an komplexe Systeme im organisationalen Kontext (Nielsen und Miraglia 2017): zeitliche Aspekte wie Dauer der Trainingsmaßnahme, mögliche Zeiträume, wie kurzfristig muss die Maßnahme umgesetzt werden; aber auch Aspekte wie: wer kann bzw. darf teilnehmen, wer kann das Training leiten, welcher Kostenrahmen ist vorgegeben. Manchmal ist auch eine Umsetzung einer Kontrollgruppe aus ethischen oder finanziellen Gründen nicht möglich – oder aber es existiert schlichtweg keine äquivalente Vergleichsgruppe (Pawson 2013; Pawson und Tilley 1997). Im Normalfall geht aber das Aufgeben oder Aufweichen eines Aspekts der bewährten Interventionsstandards mit einem Verlust an interner oder externer Validität einher. Weil in der Praxis dennoch häufiger Kompromisse geschlossen werden müssen, – Abweichungen vom so genannten „Goldstandard" also status quo sind,

D. Reis (✉)
FR Psychologie, Universität des Saarlandes, Saarbrücken, Deutschland
E-Mail: dorota.reis@uni-saarland.de

L. L. Meier
Institute for Work and Organizational Psychology (IPTO), University of Neuchâtel, Neuchâtel, Schweiz
E-Mail: laurenz.meier@unine.ch

A. Michel, A. Hoppe (Hrsg.), *Handbuch Gesundheitsförderung bei der Arbeit*,
https://doi.org/10.1007/978-3-658-28651-4_24

– wollen wir auf den folgenden Seiten damit einhergehende Implikationen für die Interpretation der Ergebnisse aufzeigen. Das erste Ziel des vorliegenden Beitrags ist somit, dafür zu sensibilisieren, welche Entscheidungen bei der Planung einer Intervention besonders relevant für dessen spätere Evaluation sind und welche Aspekte unverzichtbar erscheinen, damit das Design sensitiv bleibt – also die Chance einer akkuraten Beurteilung der Wirksamkeit unter Berücksichtigung pragmatischer Einschränkungen maximiert werden kann. Des Weiteren wollen wir ausgewählte relevante Aspekte von Interventionen diskutieren und Empfehlungen aussprechen, wie (gesundheitsfördernde) Maßnahmen im Arbeitskontext berichtet werden sollten, zum einen damit sie im Rahmen zukünftiger Forschungsbemühungen repliziert werden können und zum anderen damit sie für die Bedarfe von PraktikerInnen in den Organisationen nützlich sind.

2 Wie sollten Interventionseffekte definiert werden?

In der Literatur gibt es eine Vielzahl von Begriffen, mit denen ein Interventionseffekt umschrieben wird, beispielsweise *treatment effect*, *treatment outcome* oder *treatment response*. Wird eine gesundheitsfördernde Maßnahme geplant, entscheidet die Umsetzung einer Kontrollgruppe darüber, mit welcher Stringenz ein potenzieller Interventionseffekt interpretiert werden kann. Im Rahmen der Modelle zur kausalen Inferenz (Rubin 1974) wurde ein Interventionseffekt für ein einzelnes Individuum definiert als die Differenz zwischen dem potentiellen Wert der Outcome-Variable, wenn ein Individuum an einer Intervention teilnimmt minus der potentielle Wert der Outcome-Variable (z. B. Erschöpfung), wenn ein Individuum nicht an einer Intervention (z. B. Stress-Workshop) teilnimmt. Folglich ist der (hypothetische) Interventionseffekt ein Abgleich zwischen dem Resultat einer Intervention und einem rein hypothetischen Outcome, das man messen würde, wenn die gleiche Person sich gegen die Teilnahme entscheiden würde (*counterfactual outcome*). Weil es unmöglich ist, beide Ergebnisse auf individueller Ebene zu beobachten bzw. zu messen, wird stattdessen zur Schätzung eines durchschnittlichen Effekts auf Gruppenebene der Vergleich eines über alle Personen der Interventionsgruppe aggregierten Effekts mit dem aggregierten Effekt in einer randomisierten Kontrollgruppe vorgenommen (für eine detailliertere Darstellung des *potential outcome models*, siehe Morgan und Winship 2015). Folgt man diesem kausalen Verständnis nach Rubin (1974), so wird deutlich, warum der randomisierten Kontrollgruppe eine so zentrale Rolle zukommt. Des Weiteren folgt daraus, dass unabhängig davon, ob eine Kontrollgruppe realisiert wurde oder nicht, eine mittlere Veränderung über die Zeit (von Vortest zu Posttest) in der Interventionsgruppe nicht kausal als Effekt des Treatments interpretiert werden kann (Bland und Altman 2015). In der Definition des Interventionseffekts als Differenz zwischen der Befindlichkeit, die nach dem Training gemessen wurde und der Befindlichkeit, die gemessen worden wäre, wenn kein Training statt gefunden hätte, werden Effekte wie beispielsweise Spontanremission, natürliche Fluktuationen, saisonale Effekte oder Regression zu Mitte berücksichtigt und kontrolliert. Eine solche Kontrolle findet jedoch nicht statt, wenn bei Vorliegen oder in Abwesenheit von Kontrollgruppe(n), die Effekte in der Interventionsgruppe per se ausgewertet werden. So führt allein das Problem der Regression zu Mitte zu künstlich nach oben verzerrten Effektstärken, insbesondere wenn Individuen anhand von Einschlusskriterien zur Intervention zugelassen wurden (z. B. aufgrund erhöhter Burnoutwerte) und damit gewissermaßen eine Extremgruppe repräsentieren (Fava et al. 2003).

Das klassische Paradigma der Schätzung von Effekten auf Gruppenebene und damit einhergehende Definition von Interventionseffekten ist jedoch nicht ohne Kritik geblieben. Bereits seit den 70er-Jahren des letzten Jahrhunderts wird diesem nomothetischen Ansatz ein idiographischer gegenübergestellt (siehe z. B. Barlow und Nock 2009; Jayaratne 1977; Molenaar 2004). Die Diskrepanz wurde lange zwischen Forschenden, die beinahe ausschließlich Gruppeneffekte untersucht haben (nomothetischer Ansatz) und Praktikern, deren Arbeit auf die Individualebene fokussiert (idiographischer

Ansatz), diskutiert. Die grundlegende Frage, ob und wenn ja unter welchen Bedingungen Ergebnisse der Intervention auf Gruppenebene auf die Veränderung auf Individualebene übertragen werden dürfen (so genannte Ergodizitätsannahme, siehe z. B. Fisher et al. 2018; Hamaker 2012), ist mit zunehmender Verfügbarkeit von idiographischen Datenstrukturen (z. B. Experience Sampling Studien) auch in der Forschung stärker in den Fokus gerückt (Piccirillo et al. 2019). Im Rahmen der Forschung zu gesundheitsfördernden und -behindernden Faktoren im Arbeitskontext existieren zahlreiche Arbeiten, welche die Unterscheidung von Prozessen innerhalb versus zwischen Individuen klar berücksichtigen. Wir werden deshalb in den Abschnitten „Evaluationsdesigns“ und „Generalisierbarkeit“ noch einmal auf das Thema zurückkommen.

3 Generelle Effekte, spezifische Effekte und Mechanismen

Eine wichtige Unterscheidung hinsichtlich der Art der Interventionseffekte besteht darin, ob eine Maßnahme generelle Effekte (*common factors*), spezifische Effekte (*specific factors*) oder Mechanismen adressiert (Cuijpers et al. 2018). Die basale Frage zielt hier also darauf ab, woran wir als Forschende oder PraktikerInnen merken würden, dass die implementierte Maßnahme gewünschte Effekte zeigt. Aus versuchsplanerischer Sicht ist diese Frage nicht so einfach zu beantworten. Das medizinische Ideal der double-blind randomisierten und kontrollierten Studien in denen die Kontrollgruppe ein Placebo erhält, ist im Rahmen der Gesundheitsförderung, die auf (psychische) Gesundheit und Wohlbefinden abzielt, nicht umsetzbar. Inwieweit sollen oder wollen wir explizit machen, wie sehr der gefundene Effekt durch generelle Faktoren wie beispielsweise Zuwendung von TrainingsleiterInnen, Aufforderungscharakter der Situation oder Erwartungseffekte erklärt werden kann und wie sehr durch spezifische Faktoren, die in der Intervention intendiert sind und gezielt zur Lösung des Problems beitragen sollen? Die generellen Faktoren, die zu einer Änderung der Befindlichkeit oder des Verhaltens beitragen können, gelten in der experimentellen Forschung als Störvariablen (bedingungs-, situations- und personengebunden), die es mit Hilfe experimenteller Kontrolltechniken zu eliminieren, auszubalancieren, oder konstant zu halten gilt. Im Gegensatz dazu argumentieren Vertreter der *common factors theory*, dass die generellen Faktoren (z. B. das Bilden von Erwartungen auf Seiten der Teilnehmenden) explizit in einem theoretischen Modell als Veränderungsmechanismen anzusehen sind und dass ihre Wirkung zu maximieren sei, um den bestmöglichen Effekt zugunsten der TeilnehmerInnen zu generieren. Im Unterschied dazu sehen Proponenten der spezifischen Faktoren die Notwendigkeit, das zugrundeliegende Problem zu identifizieren (z. B. Rumination über negative Ereignisse am Arbeitsplatz als beeinträchtigender Faktor für die Schlafqualität), um eine Strategie zu implementieren, die gezielt das problematische Verhalten reduziert (z. B. Erlernen von Entspannungstechniken). In diesem Ansatz wird also der Versuch unternommen, den beeinträchtigenden oder aber aufrechterhaltenden Prozess zu identifizieren und spezifisch zu behandeln (siehe z. B. Barlow 2004; Kazdin 2007). Das Verständnis kritischer Mechanismen ist auch immer dann entscheidend, wenn bestehende Interventionsprogramme in ein neues Medium übertragen werden sollen, z. B. im Rahmen von eHealth oder mHealth (Magnusson 2019; Kazdin 2011). Durch den Wechsel von einem Einzel- oder Gruppenformat zu einem self-help (unguided) Format, gehen mit einer solchen Anpassung häufig auch Änderungen bezüglich der Länge und/oder Dosis einher, so dass die Frage nach entscheidenden Wirkkomponenten zentral wird. Eine der wichtigsten Herausforderungen bei der Planung von Interventionen, die auf spezifische Faktoren oder aber Mechanismen/Prozesse wirken sollen, ist die konkrete Ausgestaltung der Kontrollgruppe(n) – insbesondere weil diese sich ausschließlich hinsichtlich des angenommenen Wirkmechanismus von der Interventionsgruppe unterscheiden darf, aber nicht hinsichtlich weiter möglicher Prozesse (beispielsweise Erwartungseffekte). Während im Ansatz der „komplexen“ oder „realistischen“ Evaluationen (Nielsen und Miraglia 2017; Pawson

2013) die Möglichkeit, eine solche Kontrollgruppe in organisationalen Kontexten umzusetzen, kritisch gesehen wird, wird die Idee einer Wartekontrollgruppe auch in klinischen Studien zunehmend hinterfragt. Grundlage dafür liefern sowohl Primärstudien als auch metaanalytische Reviews, die zeigen konnten, dass TeilnehmerInnen in Wartekontrollgruppen über eine weniger günstige Entwicklung berichten, als dies ohne eine Teilnahme an der Studie zu erwarten wäre. Möglicherweise wird dieser Effekt dadurch ausgelöst, dass die auf die Intervention wartenden TeilnehmerInnen nichts (oder weniger) unternehmen, um die Problematik selbst in den Griff zu bekommen (Mohr et al. 2014), wodurch natürliche Fluktuationen in den relevanten Outcomes oder Spontan-Remissionen unterbunden werden. In der Konsequenz kann die durchschnittliche Befindlichkeit in einer Wartekontrollgruppe schlechter sein als bei nicht teilnehmenden Individuen (so genannter Nocebo-Effekt), wodurch die Effekte der Interventionsgruppen nach oben verzerrt werden (siehe z. B. Cristea 2018; Cuijpers und Cristea 2016; Cuijpers et al. 2016). Aus methodischer Sicht könnte also die praktische Empfehlung für Interventionen im Unternehmenskontext heißen, dass es günstiger sein könnte, unterschiedliche Maßnahmen parallel zu implementieren und hinsichtlich der Wirksamkeit zu vergleichen – eine Vorgehensweise, die in Organisationen tendenziell auch auf eine höhere Akzeptanz treffen könnte.

Leider ist jedoch eine konsequente Umsetzung von methodisch rigorosen und hochqualitativen Interventionsstudien im Kontext von Arbeit und Gesundheit schwierig und deshalb keine durchgängige Praxis. So fassen die AutorInnen eines kürzlich erschienenen qualitativen Reviews der Literatur zu achtsamkeitsbasierten Trainings für ArbeitnehmerInnen (Eby et al. 2019) beispielsweise zusammen, dass von den 67 publizierten Studien 61 % eine Kontrollgruppe implementiert hatten, davon aber nur 76 % als randomisierte Kontrollgruppe; wobei die Umsetzung als randomisiertes Wartegruppendesign nur auf 27 % aller Studien zutraf. In diesem Sinne resümierten Briner und Walshe (2015): „[. . .] body of work on interventions is therefore not only small but of quite low quality. Taken as a whole, it has significant weaknesses [. . .], which means that its academic worth and practical value are quite limited“ (S. 564) und sie plädierten für einen evidenzbasierten Ansatz sowohl bei der Planung und Umsetzung als auch beim Berichten der Ergebnisse von Interventionen. Diese Punkte wollen wir im nächsten Teil unseres Beitrags aufgreifen.

4 Was wollen wir wie verändern?

In ihrem Plädoyer für einen evidenzbasierten Ansatz in der Interventionsforschung (Briner und Walshe 2015) schlagen die Autoren vor, dass schon im Entscheidungsprozess für die Wahl einer bestimmten Maßnahme systematisch vorgegangen werden sollte und damit ähnlich der Vorgehensweise bei einer Risikobewertung. Bereits in diesem ersten Schritt sollte sowohl das Wissen aus der Organisation als auch Evidenz aus der wissenschaftlichen Literatur berücksichtigt werden. Auf organisationaler Ebene bedeutet dies, dass Informationen darüber vorliegen müssen, welches Problem (z. B. hohe Erschöpfung und Arbeitsunzufriedenheit) in welcher Zielgruppe (z. B. welche Abteilungen) vorliegt und dass dieses identifizierte Problem prinzipiell mit Hilfe der intendierten Intervention veränderbar ist. Dieser letzte Punkt impliziert auch, dass für das vorliegende Problem nicht andere Gründe vorliegen, welche von der Intervention völlig unberührt blieben. Beispielsweise wäre es weder evidenzbasiert noch zielführend, wenn zur Reduktion von Burnout und Arbeitsunzufriedenheit in einer Abteilung, die stark unterbesetzt ist, bloß eine achtsamkeitsbasierte Intervention geplant würde. Mit anderen Worten ist zum Erwerb des erforderlichen Wissens aus der Organisation eine systematische und analytische Vorgehensweise nötig, die im besten Fall durch entsprechende Daten (bspw. aus Mitarbeiterbefragungen und Arbeitsplatzbeobachtungen) gestützt werden sollte (siehe auch Kap. ▶ Gefährdungsbeurteilung psychischer Belastung im Arbeitskontext und nachfolgende Maßnahmen). Für die externe Evidenz aus der wissenschaftlichen Literatur und zur Identifikation von relevanten theoretischen Modellen und Mechanismen eignen sich sowohl

Primärstudien zu Interventionen im Arbeitskontext als auch metaanalytische Überblicksartikel. Die Nützlichkeit von Primärstudien bei der Replikation sowohl in der Forschung als auch in der Praxis, oder auch bei der Weiterentwicklung bestehender Interventionsansätze wird dadurch determiniert, wie vollständig relevante Aspekte der durchgeführten Studie berichtet wurden (Briner und Walshe 2015; siehe auch Abschn. 8 in diesem Beitrag). Metaanalysen können – zumindest in der Theorie – über die Größe der erwarteten Effekte und potenzielle Moderatoren informieren und als Grundlage dafür dienen, den Stichprobenumfang für eine inferenzstatistische Absicherung des Treatment-Effekts zu planen. In den letzten Jahren ist jedoch das Instrument der Metaanalysen wegen der fragwürdigen Qualität der eingehenden Primärliteratur in die Kritik geraten. Probleme wie kleine Stichproben, publication bias, HARKing, p-hacking und Ähnliche (für einen Überblick siehe z. B. Nelson et al. 2018; Simmons et al. 2011) schaden einem Fortschritt in der (Interventions-)Forschung auf mindestens zweierlei Arten. Zum Einen werden „nicht erfolgreiche" Interventionsstudien – d. h. Studien, die hinsichtlich der untersuchten Outcomes einen Nulleffekt gezeigt haben – nicht publiziert (und verschwinden damit in der Schublade – „file drawer-"-Problem). Die nicht publizierten Studien führen jedoch dazu, dass dieses Wissen in der Community nicht vorhanden ist und bei der Planung neuer Interventionen nicht berücksichtigt werden kann. Zum Anderen wird damit der zu erwartete Effekt stark überschätzt – und damit sowohl die statistische als auch die praktische Relevanz der Maßnahme. Beispielsweise zeigten White et al. (2019), dass die Effekte von Interventionen, die der positiven Psychologie zuzuordnen sind, de facto viel geringer ausfallen als aus der Primärliteratur und früheren Metaanalysen zu erwarten gewesen wäre. Nachdem die bisherigen metaanalytischen Ansätze für Probleme der Primärliteratur zu korrigieren nicht hinreichend sind (Carter et al. 2019), müssen wir sowohl in der Forschung als auch in der Praxis deren Ergebnisse mit Vorsicht und Zurückhaltung rezipieren und bei den Interventionen von einer geringeren Effektivität als gemeinhin gedacht ausgehen.

5 Operationalisierung und Messung

Sobald die Inhalte der gesundheitsfördernden Maßnahme nach evidenzbasierten Kriterien identifiziert wurden, stellt sich die Frage einer geeigneten Operationalisierung. Bei der Evaluation einer Intervention entscheidet die Sensitivität der Messung von Prädiktoren und Outcomes darüber ob überhaupt und wenn ja, wie valide und reliabel ein Effekt der Intervention gezeigt werden kann. Wir möchten im Folgenden auf einige besonders relevante Aspekte im Kontext von Interventionen eingehen: auf die Unterscheidung zwischen Trait- und Statemaßen, den Umgang mit heterogenen Skalen und die Frage nach Messinvarianz der untersuchten Konstrukte über die Zeit.

Trait versus State. Bei der Operationalisierung von Prädiktoren und Outcomes in einer Interventionsmaßnahme muss berücksichtigt werden, wie stabil versus fluktuierend die betreffenden Konstrukte sind. Von dieser sehr grundsätzlichen Überlegung hängen weitere Entscheidungen bezüglich einer adäquaten Messung des Konstrukts und bezüglich eines adäquaten Interventionsdesigns ab. Obwohl generell auch Veränderungen stabiler Eigenschaften (Traits) im Rahmen von Interventionen adressiert werden können (Bleidorn et al. 2019; Roberts et al. 2017), setzen diese sehr wahrscheinlich eine höhere Dosis und Dauer der Intervention voraus, als sie im Arbeitskontext und im Rahmen von gesundheitsfördernden Maßnahmen tendenziell realisierbar sind. Aus diesem pragmatischen Grund – aber auch aufgrund eines genuinen theoretischen Interesses an dynamischen Prozessen im Kontext von Arbeit, fokussieren zahlreiche Interventionsstudien in der Wahl ihrer Outcomes auf stärker fluktuierende und veränderbare State-Maße, wie z. B. wöchentliches oder tägliches Burnout (Nurmi et al. 2008), Arbeitsengagement (Knight et al. 2019), Stimmung (Meier et al. 2016), psychologisches Abschalten von der Arbeit (Althammer et al. under review), Schlafqualität (Hülsheger et al. 2015), oder Job Crafting (van den Heuvel et al. 2015). Das stärkere Interesse für fluktuierende Maße und Veränderungen, die innerhalb von Individuen stattfinden, korrespondiert mit einer Messung der

Konstrukte als State – entweder mit bezüglich des relevanten Zeitraums adaptieren Trait-Skalen, oder aber mit speziell konzipierten State-Skalen. Bei der Planung von Interventionen sollte damit auch die relevante Laufzeit einer Intervention berücksichtigt werden, einerseits um bei der Messung die Variabilität des Konstrukts abbilden zu können und andererseits damit die Dauer und damit auch die Dosis der Intervention nicht zu gering sind. Obwohl die Verwendung von stärker fluktuierenden Maßen zur Evaluation von Maßnahmen der Gesundheitsförderung sowohl inhaltlich als auch methodisch gerechtfertigt ist, müssen in der Forschung und Praxis auch die Kosten-Nutzen-Perspektive, die Nachhaltigkeit – und damit letztlich auch ein ethischer Aspekt kritisch diskutiert werden: Selbst wenn es gelingt, relativ kurzfristig eine Verbesserung der Befindlichkeit beispielsweise durch eine Reduktion des allgemeinen Stresslevels oder arbeitsbezogener Rumination (Bono et al. 2013; Querstret et al. 2017) zu erreichen, fehlen in zahlreichen Studien Daten aus Follow-up Befragungen – und damit die Information darüber, wie stabil die erzielten Treatment-Effekte sind. Entsprechendes Wissen wäre aber sowohl für die Theoriebildung als auch für die Praxis von zentraler Bedeutung.

Generelle Konstrukte versus Facetten/Symptome. Die stärkere Fokussierung auf State-Konstrukte hat hinsichtlich ihrer Erfassung weitere günstige Effekte für die Evaluation von Interventionen. Nachdem die wiederholten Befragungen im Alltag aus Gründen der Zumutbarkeit für die Teilnehmenden kurz gehalten werden müssen (Gabriel et al. 2019), werden häufig aus einem heterogeneren Pool an Items von Trait-Skalen bewusst entweder einzelne Items selektiert (siehe z. B. Reis et al. 2016), oder aber nur bestimmte Facetten (z. B. nur ausgewählte Strategien der Emotionsregulation) untersucht. Durch diese Vorauswahl werden die zu verändernden Konstrukte oftmals unidimensionaler gemessen. Dadurch wird auf inhaltlicher Ebene deutlicher, was genau (welche (un-)günstige Strategie, welches Verhalten u. Ä.) durch die Intervention beeinflusst werden kann, während aus methodischer Sicht die Bildung von Gesamtscores eher gerechtfertigt erscheint. Im Gegensatz dazu können mit Hilfe der bei heterogenen Konstrukten gebildeten Mittel- oder Summenwerte die Veränderungsprozesse oftmals nicht abgebildet werden. Ein prominentes Beispiel für diese Problematik stellt die Messung von Depression mit gängigen Messinstrumenten dar: durch die Zusammenlegung sehr unterschiedlicher Symptome wie Schlafprobleme, Verschlechterung von Stimmung und Essprobleme zu einem Gesamtscore kann bei einer Reduktion der Werte durch eine Intervention der eigentliche Prozess/Mechanismus kaum identifiziert werden (Fried et al. 2016). Die Fokussierung auf weniger breit erfasste Konstruktfacetten hat darüber hinaus den Vorteil, dass Mechanismen mit geringerer konzeptueller Überlappung untersucht werden können. Die Messung auf Symptom- oder Facettenebene korrespondiert gut mit einer Auswertung solcher Designs mit sogenannten Netzwerkmodellen (siehe Abschn. 6).

Messinvarianz im Längsschnitt. In klassischen Prä-Post-Follow-Up Designs wird die Evaluation einer Maßnahme durch einen Vergleich von manifesten Mittelwerten vorgenommen. Einer solchen Auswertung im Kontext von t-tests, ANOVAs – aber auch manifesten Mehrebenenmodellen oder Wachstumskurvenmodellen liegt die Annahme zugrunde, dass eine Messung zu allen Messzeitpunkten das gleiche Konstrukt in gleicher Weise repräsentiert. Analog zu Gruppenvergleichen (Meredith 1993) sollte diese Annahme aber nicht per se vorausgesetzt werden. Sobald mehrere Indikatoren eines Konstrukts in den Daten vorliegen, kann die Annahme gleich bleibender psychometrischer Eigenschaften der wiederholt erhobenen Messinstrumente formal überprüft werden. Die Frage der Messinvarianz bezieht sich hierbei auf die Verknüpfung der Indikatoren mit den latenten Variablen und damit auf die Konstanz der Faktorladungen und Intercepts über die Zeit (Widaman et al. 2010). Erst ab dem Vorliegen von sogenannter starker faktorieller Invarianz – die erfüllt ist, wenn sowohl die Ladungen als auch die Intercepts über die Zeit hinweg gleich bleiben – können Unterschiede in den mittleren Ausprägungen der Outcomes ohne Einschränkungen als Veränderungen der wahren Merkmalsausprägung (*true score*) interpretiert werden (Widaman und Reise 1997). Wenn starke faktorielle Invarianz nicht vorliegt (oder nicht getestet wurde),

müssen potenziell gefundene Effekte der Intervention vorsichtiger interpretiert werden – denn ein Teil dieser Veränderungen kann auch auf veränderte Messeigenschaften zurückgehen (z. B. niedrigere Itemschwierigkeiten durch Lern- oder Gewöhnungseffekte in der Messwiederholung).

6 Evaluationsdesigns

Neben den klassischen Prä-Post-Follow-up-Designs möchten wir im folgenden Abschnitt auf weitere – teilweise erst kürzlich vorgeschlagene – Alternativen eingehen, wie Evaluationsdesigns konzipiert (und anschließend ausgewertet) werden können. Dabei bieten neue Wege der Datenerhebung (Experience Sampling, Daten von Sensoren) die Möglichkeit, nicht nur Informationen über das Alltagsgeschehen der Teilnehmenden zu erhalten, sondern auch direkt im Alltag („in real life and real time", Heron und Smyth 2010) zu intervenieren. Damit können die Interventionshäufigkeit und -dosis an das Auftreten (un-)erwünschten Verhaltens angepasst werden (Lischetzke et al. 2015) und Empfehlungen an Organisationen und ArbeitnehmerInnen können die volle Dynamik der Prozesse berücksichtigen (Sosnowska und Griep 2019). Dies kann insbesondere dann nützlich sein, wenn wir vorab Informationen über den Alltag der Teilnehmenden erheben um die Intervention zu „personalisieren". In solchen Designs kann der Interventionseffekt sowohl in Relation zu einer Kontrollgruppe (*between-subjects*) als auch in innerhalb von Personen (*within-subjects*) untersucht werden (Lischetzke et al. 2015), erfordern in der Regel aber auch den Einsatz von (multivariaten) Multilevelmodellen als Auswertungsstrategie.

Einen weiteren vielversprechenden Ansatz in der Evaluation von Interventionen stellen Netzwerkanalysen dar (Borsboom und Cramer 2013; Fried und Cramer 2017; Schmittmann et al. 2013). In diesen stellt das Ziel der Evaluation nicht die Einschätzung der Veränderung in einem generellen (latenten) aber heterogenen (entweder per definitionem multidimensionalem oder in der Theorie unidimensionalem) Konstrukt dar (wie beispielsweise „Burnout"); es soll viel mehr die Veränderung in dem Zusammenwirken einzelner Symptome sichtbar gemacht werden (chronische Müdigkeit oder Erschöpfung, Desillusionierung, Resignation oder Rückzug). Netzwerkanalysen erscheinen gerade für Maßnahmen der Gesundheitsförderung nützlich: bei komplexen Problemlagen, die auf mehreren Ebenen interagieren, können so viel spezifischer und feingliedriger zentrale Aspekte (bei einer querschnittlichen Analyse) und Prozesse (bei einer längsschnittlichen Analyse, siehe z. B. Bringmann et al. 2013) in Augenschein genommen werden, aus deren Verbesserung weitere positive Veränderungen resultieren können. So konnten beispielsweise Stochl und Kollegen (Stochl et al. 2019) mit Netzwerkanalysen an Daten von vier Kohorten zeigen, dass die zentralsten Aspekte psychologischen Wohlbefindens eine positive Selbst-Wahrnehmung und positive Stimmung sind und sprechen deshalb die Empfehlung aus, dass diese bei der Planung von Interventionsprogrammen der öffentlichen Hand in den Vordergrund rücken sollten.

7 Generalisierbarkeit

Sowohl in der Praxis als auch in der Forschung stellt sich häufig die Frage nach der Generalisierbarkeit durchgeführter Interventionsmaßnahmen. Das Konzept der Generalisierbarkeit ist hier eng verknüpft mit differentieller und konditionaler/situativer Effektivität von Interventionen (Lischetzke et al. 2015). Dies gilt damit sowohl hinsichtlich erfolgreicher als auch nicht erfolgreicher Maßnahmen. In Interventionsstudien wird häufig eine Einschränkung der Generalisierbarkeit lediglich hinsichtlich der Zusammensetzung der Stichprobe als Limitation diskutiert – z. B. inwieweit sich die Effekte einer an Lehrkräften durchgeführten Maßnahme zur Stressreduktion auf andere Berufsgruppen übertragen lassen. Hierbei werden oftmals zwei unterschiedliche Probleme vermischt: zum einen die Frage, inwieweit die Verfahren frequentistischer Statistik bei nicht zufällig gezogenen Stichproben zulässig bzw. interpretierbar sind (für eine differenzierte Diskussion siehe Weakliem 2016) und zum anderen die Frage, inwieweit die fehlende Repräsentativität Schlüsse auf die Population wegen systematischer Fehler

gefährdet. Weitere Aspekte der Generalisierbarkeit betreffen ihre differentielle Wirksamkeit („Für wen ist die Maßnahme erfolgreich/für wen sind die Interventionsinhalte nützlich oder hilfreich?“) und konditionaler oder situativer Wirksamkeit („An welchen Tagen oder in welchen Situationen reduziert die Intervention Stress/verbessert die Intervention die Befindlichkeit der Teilnehmenden?“). Bei gesundheitsfördernden Maßnahmen in Unternehmen stellt sich die Frage der Generalisierbarkeit darüber hinaus auf unterschiedlichen Ebenen. Interventionen, die inhaltlich auf der Individualebene konzipiert sind, werden meist über aggregierte Gruppenunterschiede evaluiert – eine Vorgehensweise, die durchaus kritisch diskutiert wird (siehe z. B. Fisher et al. 2018; Hamaker 2012). Es ließe sich aber durchaus auch argumentieren, dass der Hauptfokus solcher Interventionen auch direkt auf übergeordneter Ebene zu sehen ist (z. B. die Reduktion der Fehlzeiten oder die Verbesserung des Teamklimas in einer Abteilung). Konsequenterweise werden im Rahmen der Ansätze von „realistischen Evaluationen“ (Pawson und Tilley 1997) so genannte Kontext-Mechanismus-Outcome Konfigurationen untersucht. Dies geschieht unter der Annahme, dass der Kontext eine zentrale Rolle darin spielt, wie Mechanismen bestimmte Outcomes triggern können und damit die Interventionsoutcomes de facto aufgrund einer Interaktion zwischen Kontext und Mechanismus generiert werden. Damit impliziert Generalisierbarkeit in solchen komplexen Evaluationsdesigns sowohl die Frage nach der Übertragbarkeit in einen anderen Kontext als auch eine genuin mehrschichtige Struktur der Daten.

8 Was sollten wir berichten?

Briner und Walshe (2015) argumentieren, dass in einem evidenzbasierten Ansatz in der Interventionsforschung nicht nur die Art und Weise, wie eine Maßnahme hinsichtlich ihres Designs geplant wird, relevant sind, sondern auch wie die Maßnahmen berichtet werden. Sie schlagen eine möglichst transparente Vorgehensweise vor, die es sowohl Praktikern als auch Forschenden ermöglichen soll, durchgeführte Studien hinsichtlich ihrer Qualität zu bewerten, um zu begründeten Entscheidungen bei der Auswahl von Studien für die Anwendung in der Praxis oder zur Replikation zu gelangen. Dieser Sicht möchten wir uns anschließen und argumentieren, dass ein möglichst vollständiges, umfassendes und offenes Berichten von Interventionen zentral ist.

Ein transparentes und standardisiertes Berichten kann sich an zahlreichen vorhandenen Checklisten orientieren – so zum Beispiel an der „TIDieR intervention reporting checklist“ von Hoffmann et al. 2014 oder der „TREND Checkliste“ von Des Jarlais et al. 2004 bei nicht-randomisierten Designs, oder aber der CONSORT-SPI (Montgomery et al. 2018), die eine Erweiterung der CONSORT-Empfehlungen für psychologische und behaviorale Interventionen darstellt. Die Transparenz hinsichtlich angenommener Wirkmechanismen, verwendeter Materialien, Stichprobenumfangsplanung, Rekrutierungsstrategien, Umgang mit Abbrechern, Festlegung von relevanten Evaluationskriterien etc. sind gleichermaßen für PraktikerInnen und Forschende bedeutsam. Die Offenheit hinsichtlich der Vorgehensweise bei der Datenanalyse bezüglich Datenmanagement, Umgang mit fehlenden Werten, alternativen analytischen Ansätzen etc. ist sowohl für Synthese von Interventionsstudien im Rahmen von Metaanalysen oder Reviews, für Replikationen als auch Adaptation von Maßnahmen an neue Kontexte oder Zielgruppen entscheidend. Einen weiteren wichtigen Aspekt der Transparenz stellt die Bereitstellung anonymisierter Originaldaten zur Nachnutzung dar. Dieser Aspekt wird insbesondere bei kleineren Stichproben oder bei potenziell identifizierbaren Organisationen und Auftraggebern gesundheitsfördernder Maßnahmen oftmals mit Bedenken gesehen. Eine mögliche Antwort auf solche Bedenken kann das Generieren von synthetischen Datensätzen sein (Quintana 2020), die zwar hinsichtlich der statistischen Eigenschaften den Originaldaten entsprechen, aber keinerlei persönliche oder identifizierbare Informationen mehr beinhalten.

Das Problem der nicht ausreichenden Berichterstattung von Interventionsstudien hat zahlreiche Facetten. Während manche Studien einfach nur wegen unvollständiger Beschreibung der Prozedur nicht rezipiert werden können, sind die Kon-

sequenzen von nicht-publizierten negativen Ergebnissen (*publication bias, file-drawer problem*), oder des selektiven Veröffentlichens von Interventionen ausschließlich bezogen auf die Outcomes, „die funktioniert" haben (*selective reporting* oder *„picking"*, manchmal einhergehend mit HARKing) in der psychologischen Forschung der letzten Jahre im Rahmen der *reproducibility crisis* sehr salient geworden (siehe z. B. Munafò et al. 2017). In der biomedizinischen (Interventions-)Forschung spricht man hierbei auch von *research waste* und beziffert diesen auf 85 % der gesamten Forschungsergebnisse (Chalmers und Glasziou 2009). Letztlich müssen wir als Forschende und PraktikerInnen die gravierenden ethischen Implikationen wahrhaben: die Prinzipien evidenzbasierter und transparenter Forschung tragen dazu bei, dass nicht förderliche Maßnahmen und damit sogar potenzieller Schaden für die TeilnehmerInnen vermieden – während tatsächlich erfolgreiche Interventionen schneller rezipiert werden können.

Literatur

Althammer, S., Reis, D., van der Beek, S., Beck, L., & Michel, A. (under review). A mindfulness intervention promoting work-life balance: How segmentation preference moderates change trajectories of work-life balance, detachment, and well-being.

Barlow, D. H. (2004). Psychological treatments. *American Psychologist, 59*(9), 869–878. https://doi.org/10.1037/0003-066X.59.9.869.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science, 4*(1), 19–21. https://doi.org/10.1111/j.1745-6924.2009.01088.x.

Bland, M. J., & Altman, D. (2015). Best (but oft forgotten) practices: Testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. *The American Journal of Clinical Nutrition, 102*(5), 991–994. https://doi.org/10.3945/ajcn.115.119768.

Bleidorn, W., Hill, P. L., Back, M. D., Denissen, J. J. A., Hennecke, M., Hopwood, C. J., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Wagner, J., Wrzus, C., Zimmermann, J., & Roberts, B. (2019). The policy relevance of personality traits. *American Psychologist, 74*(9), 1056–1067. https://doi.org/10.1037/amp0000503.

Bono, J. E., Glomb, T. M., Shen, W., Kim, E., & Koch, A. J. (2013). Building positive resources: Effects of positive events and positive reflection on work stress and health. *Academy of Management Journal, 56*, 1601–1627. https://doi.org/10.5465/amj.2011.0272.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608.

Briner, R. B., & Walshe, N. D. (2015). An evidence-based approach to improving the quality of resource-oriented well-being interventions at work. *Journal of Occupational and Organizational Psychology, 88*(3), 563–586. https://doi.org/10.1111/joop.12133.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., et al. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE, 8*(4), e60188. https://doi.org/10.1371/journal.pone.0060188.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*(2), 115–144. https://doi.org/10.1177/2515245919847196.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet, 374*(9683), 86–89. https://doi.org/10.1016/S0140-6736(09)60329-9.

Cristea, I. A. (2018). The waiting list is an inadequate benchmark for estimating the effectiveness of psychotherapy for depression. *Epidemiology and Psychiatric Sciences, 1–2*. https://doi.org/10.1017/S2045796018000665.

Cuijpers, P., & Cristea, I. (2016). How to prove that your therapy is effective, even when it is not: A guideline. *Epidemiology and Psychiatric Sciences, 25*(05), 428–435. https://doi.org/10.1017/S2045796015000864.

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., & Huibers, M. J. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry, 15*(3), 245–258. https://doi.org/10.1002/wps.20346.

Cuijpers, P., Reijnders, M., & Huibers, M. J. H. (2018). The role of common factors in psychotherapy outcomes. *Annual Review of Clinical Psychology*. https://doi.org/10.1146/annurev-clinpsy-050718-095424.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & Trend Group (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94*(3), 361–366.

Eby, L. T., Allen, T. D., Conley, K. M., Williamson, R. L., Henderson, T. G., & Mancini, V. S. (2019). Mindfulness-based training interventions for employees: A qualitative review of the literature. *Human Resource Management Review, 29*(2), 156–178. https://doi.org/10.1016/j.hrmr.2017.03.004.

Fava, M., Evins, A. E., Dorer, D. J., & Schoenfeld, D. A. (2003). The problem of the placebo response in clinical

trials for psychiatric disorders: Culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics, 72*, 115–127. https://doi.org/10.1159/000069738.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115.

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science, 12*(6), 999–1020. https://doi.org/10.1177/1745691617705892.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28*(11), 1354.

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods, 22*(4), 969–1006. https://doi.org/10.1177/1094428118802626.

Hamaker, E. L. (2012). Why researchers should think „within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Hrsg.), *Handbook of research methods for studying daily life* (S. 43–61). New York, NY: Guilford.

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology, 15*, 1–39. https://doi.org/10.1348/135910709X466063.

Heuvel, M. van den, Demerouti, E., & Peeters, M. C. W. (2015). The job crafting intervention: Effects on job resources, self-efficacy, and affective well-being. *Journal of Occupational and Organizational Psychology, 88*(3), 511–532. https://doi.org/10.1111/joop.12128.

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., . . . & Lamb, S. E. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348*, g1687. https://doi.org/10.1136/bmj.g1687.

Hülsheger, U. R., Feinholdt, A., & Nübold, A. (2015). A low-dose mindfulness intervention and recovery from work: Effects on psychological detachment, sleep quality, and sleep duration. *Journal of Occupational and Organizational Psychology, 88*(3), 464–489. https://doi.org/10.1111/joop.12115.

Jayaratne, S. (1977). Single-subject and group designs in treatment evaluation. *Social Work Research and Abstracts, 13*(3), 35–42.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3*(1), 1–27. https://doi.org/10.1146/annurev.clinpsy.3.022806.091432.

Kazdin, A. E. (2011). Evidence-based treatment research: Advances, limitations, and next steps. *American Psychologist, 66*(8), 685. https://doi.org/10.1037/a0024975.

Knight, C., Patterson, M., & Dawson, J. (2019). Work engagement interventions can be effective: A systematic review. *European Journal of Work and Organizational Psychology, 28*(3), 348–372. https://doi.org/10.1080/1359432X.2019.1588887.

Lischetzke, T., Reis, D., & Arndt, C. (2015). Data-analytic strategies for examining the effectiveness of daily interventions. *Journal of Occupational and Organizational Psychology, 88*(3), 587–622. https://doi.org/10.1111/joop.12104.

Magnusson, K. (2019). *Methodological issues in psychological treatment research. Applications to gambling research and therapist effects*. Stockholm: Karolinska Institutet.

Meier, L. L., Cho, E., & Dumani, S. (2016). The effect of positive work reflection during leisure time on affective well-being: Results from three diary studies. *Journal of Organizational Behavior, 37*(2), 255–278.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Miller, L. E., & Stewart, M. E. (2011). The blind leading the blind: Use and misuse of blinding in randomized controlled trials. *Contemporary Clinical Trials, 32*(2), 240–243. https://doi.org/10.1016/j.cct.2010.11.004.

Mohr, D. C., Ho, J., Hart, T. L., Baron, K. G., Berendsen, M., Beckner, V., . . . & Schroder, K. E. (2014). Control condition design and implementation features in controlled trials: A meta-analysis of trials evaluating psychotherapy for depression. *Translational Behavioral Medicine, 4*(4), 407–423. https://doi.org/10.1007/s13142-014-0262-3.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1.

Montgomery, P., Grant, S., Mayo-Wilson, E., et al. (2018). Reporting randomised trials of social and psychological interventions: The CONSORT-SPI 2018 Extension. *Trials, 19*, 407. https://doi.org/10.1186/s13063-018-2733-1.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research*. New York, NY: Cambridge University Press.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., . . . & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836.

Nielsen, K., & Miraglia, M. (2017). What works for whom in which circumstances? On the need to move beyond the ‚what works?' question in organizational intervention research. *Human Relations, 70*(1), 40–62.

Nurmi, J. E., Salmela-Aro, K., Keskivaara, P., & Näätänen, P. (2008). Confidence in work-related goals and feelings of exhaustion during a therapeutic intervention for burnout: A time-series approach. *Journal of Occupational and Organizational Psychology, 81*(2), 277–297.

Pawson, R. (2013). *The science of evaluation: A realist manifesto*. Thousand Oaks/London: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks/London: Sage.

Piccirillo, M. L., Beck, E. D., & Rodebaugh, T. L. (2019). A clinician's primer for idiographic research: Considerations and recommendations. *Behavior Therapy, 50*(5), 938–951. https://doi.org/10.1016/j.beth.2019.02.002.

Querstret, D., Cropley, M., & Fife-Schaw, C. (2017). Internet-based instructor-led mindfulness for work-related rumination, fatigue, and sleep: Assessing facets of mindfulness as mechanisms of change. A randomized waitlist control trial. *Journal of Occupational Health Psychology, 22*(2), 153–169. https://doi.org/10.1037/ocp0000028.

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *Elife, 9*, e53275.

Reis, D., Hoppe, A., Arndt, C., & Lischetzke, T. (2016). Time pressure with state vigour and state absorption: Are they non-linearly related. *European Journal of Work and Organizational Psychology*, 1–13. https://doi.org/10.1080/1359432X.2016.1224232.

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin, 143*-(2), 117–141. https://doi.org/10.1037/bul0000088.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350.

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology, 31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Sosnowska, J., & Griep, Y. (2019). Well-being at work: Applying dynamics of affect in positive psychological interventions. In L. E. Van Zyl & S. Rothmann Sr. (Hrsg.), *Theoretical approaches to multi-cultural positive psychological interventions*. Cham: Springer.

Stochl, J., Soneson, E., Wagner, A. P., Khandaker, G. M., Goodyer, I., & Jones, P. B. (2019). Identifying key targets for interventions to improve psychological well-being: Replicable results from four UK cohorts. *Psychological Medicine, 49*(14), 2389–2396. https://doi.org/10.1017/S0033291718003288.

Wadhwa, M., & Cook, T. D. (2019). The set of assumptions randomized control trials make and their implications for the role of such experiments in evidence-based child and adolescent development research. *New directions for child and adolescent development, 2019* (167), 17–37.

Weakliem, D. L. (2016). *Hypothesis testing and model selection in the social sciences*. New York: Guilford Publications.

White, C. A., Uttl, B., & Holder, M. D. (2019). Meta-analyses of positive psychology interventions: The effects are much smaller than previously reported. *PLOS ONE, 14*(5), e0216588. https://doi.org/10.1371/journal.pone.0216588.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Hrsg.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (S. 281–324). Washington, DC: American Psychological Association.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x.