Handbuch Friedenspsychologie

Christopher Cohrs, Nadine Knab & Gert Sommer (Hrsg.)

Christopher Cohrs • Nadine Knab • Gert Sommer (Hrsg.)

Handbuch Friedenspsychologie

ISBN 978-3-8185-0565-3

DOI: https://doi.org/10.17192/es2022.0021

Lektorat und Formatierung: Michaela Bölinger und Johanna Hoock

Titelbild und Kapitelgestaltung: Nadine Knab

Umschlagbild: Hoffnung (Esperanza). Frieden, Dankbarkeit, Kreativität und Wiederstandfähigkeit sind die Symbole und Elemente, die in diesem Kunstwerk in Einklang gebracht werden. Es ist als Großformat in der Gemeinde 13 in Medellín, Kolumbien, Teil der Graffiti-Tour. Das Kunstwerk vermittelt eine wichtige Botschaft der Hoffnung sowohl an die lokale Gemeinde als auch an ausländische Besucher/innen.
@medapolo.trece @fateone96 @radycalshoes @pemberproducciones

https://handbuch-friedenspsychologie.de
Website-Gestaltung: Tamino Konur, Iggy Pritzker, Nadine Knab

Forum Friedenspsychologie
https://www.friedenspsychologie.de

Methoden der friedenspsychologischen Forschung

Kea S. Brahms[1] & Manfred Schmitt[2]

Zusammenfassung

Das Kapitel bietet einen Überblick über für die Friedenspsychologie relevante Aspekte des Forschungsprozesses. Zunächst werden verschiedene Erkenntnisziele und Forschungsdesigns vorgestellt. Im Anschluss daran erfolgt eine ausführliche Diskussion von sowohl quantitativen als auch qualitativen Gütekriterien sowie die Darstellung verschiedener Samplingtechniken, Erhebungs- und Auswertungsmethoden. Es werden zahlreiche Beispiele von Studien mit friedenspsychologischem Bezug vorgestellt und – wenn nötig – auf dem Feld eigene Besonderheiten im Forschungsprozess, z.B. beim Sampling, hingewiesen.

Schlüsselwörter: Forschungsmethoden; qualitativ; quantitativ; Untersuchungsplanung; Gütekriterien; Sampling; Datenerhebung; Datenauswertung

Abstract

In this chapter, we provide an overview of the research process insofar as it is relevant to peace psychology. To begin, different research goals and designs are introduced. Subsequently, we discuss different quality criteria, sampling techniques, as well as methods of data collection and analysis. In doing so, we cover aspects of both qualitative and quantitative methodology. The chapter is complemented with various examples of studies from the field of peace psychology. Where applicable, we refer to methodological particularities specific to the area.

Keywords: research methods; qualitative; quantitative; research design; quality criteria; sampling; data collection; data analysis

Die Friedenspsychologie trägt zum wissenschaftlichen Verständnis der Entstehung und Dynamik von kollektiven Konflikten bei. Dieses Verständnis wird benötigt, um Konflikten vorzubeugen und sie gewaltfrei und konstruktiv zu lösen. Zur Generierung empirischen Wissens benötigt die Friedenspsychologie Forschungsmethoden und die Kompetenz, diese richtig anzuwenden. Methodenexpertise ist unerlässlich, um die Güte und Aussagekraft vorhandener Untersuchungen zu beurteilen und neue Studien so planen und durchführen zu können, dass wissenschaftliche Gütekriterien bestmöglich erfüllt sind.

1 keabrahms@gmail.com
2 Fachbereich Psychologie, Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, schmittm@uni-landau.de

Es ist die Auffassung der Autor*innen, dass eine strikte Trennung der beiden in unterschiedliche „Forschungswelten“ wenig zielführend ist und stattdessen eine an der Fragestellung orientierte Auswahl der Untersuchungsmethoden erfolgen sollte. Zudem lassen sich quantitative und qualitative Verfahren oft sinnvoll ergänzen. Multimethodale Forschung ist methodisch einseitiger Forschung meistens überlegen. Dies gilt für die Auswertung von Daten ebenso wie für ihre Erhebung (Eid & Diener, 2006). Insbesondere in der Friedenspsychologie kann die Anwendung einer breiten Methodenvielfalt wesentlich zum Erkenntnisgewinn beitragen. Deshalb wird in diesem Kapitel eine gleichwertige Darstellung qualitativer wie quantitativer Methoden angestrebt. Die Durchführung empirischer Untersuchungen lässt sich in der Regel in die Schritte der Untersuchungsplanung, Datenerhebung und Datenanalyse unterteilen. Entsprechend gliedern wir dieses Kapitel.

Untersuchungsplanung

Jede empirische Untersuchung bedarf eines geeigneten Untersuchungsplans. Maßgeblich für dessen Wahl sind die Fragestellungen und Ziele einer Studie sowie die materiellen, organisatorischen und personellen Voraussetzungen aufseiten der Untersuchten und der Forschenden. Untersuchungspläne von Studien, die kausale Schlussfolgerungen erlauben sollen, müssen andere Anforderungen erfüllen als Untersuchungspläne von Studien, die der Erkundung eines neuen Phänomens dienen. Untersuchungen von veränderlichen Phänomenen müssen anders angelegt sein als Untersuchungen von stabilen Phänomenen. Bei Untersuchungen, die im Labor durchgeführt werden, können Störeinflüsse weitgehend kontrolliert werden. Bei Studien, die im Feld stattfinden, ist die Kontrolle von Drittvariablen nur mit Einschränkungen möglich; dafür bieten sie andere Vorteile.

Ethische Vorüberlegungen

Bereits zu Beginn einer (friedens-)psychologischen Untersuchung sollten sich Forschende umfangreiche Gedanken über die ethischen Implikationen ihres Vorhabens machen und eine kritische Reflektion ihrer eigenen Standpunkte vornehmen. In der Friedenspsychologie haben ethische Fragen einen besonderen Stellenwert. Die Arbeit mit vulnerablen Zielgruppen, die sensiblen Themen und nicht zuletzt der normative Anspruch, den die meisten friedenspsychologisch Forschenden vertreten, werfen wichtige Fragen auf: Welchen Nutzen bringt ein geplantes Forschungsprojekt? Welche Nachteile können daraus entstehen – für die Forschungsteilnehmenden, für stigmatisierte Gruppen oder für die Gesellschaft? Welche unbeabsichtigten Konsequenzen kann eine Studie haben? Wie können Ergebnisse möglicherweise missbraucht werden? Wie wird mit „ungewollten“ Ergebnissen umgegangen? Forschung ist längst nicht mehr auf den wissenschaftlichen Elfenbeinturm begrenzt, sondern findet regelmäßig Eingang in öffentliche Diskurse. Die damit einhergehende Legitimierungsmacht ist nicht zu unterschätzen.

Die zunehmende Priorisierung ethischer Kriterien im Forschungsalltag drückt sich im Vorliegen zahlreicher gesetzlicher, institutioneller und publikationsbasierter Rahmenbedingungen aus. So fordern die meisten Fachzeitschriften mittlerweile die Erfüllung ethischer Mindeststandards als Voraussetzung zur Publikation einer Studie. Für Details zu den einschlägigen Forderungen, z.B. die informierte Einwilligung von Teilnehmenden oder den sensiblen Umgang mit personenbezogenen Daten, sei auf die berufsethischen Richtlinien des Berufsverbands Deutscher Psychologinnen und Psychologen e.V. der Deutschen Gesellschaft für Psychologie e.V. (2016) verwiesen.

Erkenntnisziele und Versuchsdesigns

Am Beginn einer empirischen Untersuchung stehen ein bestimmtes Erkenntnisinteresse und eine damit verbundene Forschungsfrage. Der Forschungsgegenstand und welche Arten von Antworten Forschende suchen, bestimmen ganz wesentlich die Wahl des Forschungsdesigns, der konkreten Methoden und letztlich auch, welchen Bewertungsmaßstäben ein Untersuchungsplan unterliegt (siehe Abschnitt „Was macht einen guten Untersuchungsplan aus?“). Im Folgenden werden die wichtigsten Erkenntnisziele und die damit einhergehenden Forschungsdesigns kurz erläutert.

Exploration und Theoriebildung. Häufig beginnt ein Forschungsprogramm damit, dass ein neues Phänomen, z.B. die Möglichkeiten des Konfliktmanagements in sozialen Medien oder die Bewegung der Querdenker*innen, die Neugierde von Forschenden weckt oder gesellschaftliches Interesse seine wissenschaftliche Ergründung anregt. In solchen Fällen ist es sinnvoll, sich dem Phänomen zunächst explorativ zu nähern, das heißt ohne Einschränkung durch Vorannahmen alle Informationen zusammenzutragen, die helfen könnten, das Phänomen zu verstehen. Ein möglichst natürlicher Zugang zu den Alltags- und Lebenswelten der involvierten Akteure ist bei solchen Untersuchungen häufig besonders wichtig, um die Generierung von neuem, den Forschenden unbekanntem Wissen zu unterstützen. Die so gewonnenen Informationen müssen anschließend geordnet, verdichtet und gezielt angereichert werden, um Hypothesen zu generieren oder sogar Theorien zu entwickeln und diese systematisch zu untersuchen. Leitfragen in einem solchen Erkundungsprozess könnten sein: Welche gegensätzlichen Interessen und welche konkurrierenden Akteure sind beteiligt? Welche Koalitionen gehen diese ein? Ähnelt das Konfliktgeschehen früheren Konflikten der Beteiligten oder Konflikten zwischen anderen Parteien? Welche aus der Konfliktforschung bekannten Faktoren lassen sich auf den aktuellen Fall übertragen? Insbesondere in der qualitativen Forschung gibt es diverse Methoden (z.B. die Grounded Theory-Methodologie, Glaser & Strauss, 1967), die sich besonders zur Exploration und Theorieentwicklung eignen. Dennoch gilt, dass explorative Forschung nicht auf bestimmte Methoden festgelegt werden kann. Vielmehr steigt ihr Wert mit der Vielfalt an Methoden, mit denen potentiell relevante Informationen zusammengetragen werden.

Deskriptive Untersuchungen. Erkundenden (explorativen) und beschreibenden (deskriptiven) Untersuchungen ist gemeinsam, dass Theorieprüfung keine Priorität hat. Der wesentliche Unterschied zwischen ihnen besteht in der Standardisierung. Während friedenspsychologische Erkundungsstudien zum Beispiel durch einen aktuellen Konflikt ausgelöst werden können und ein auf diesen Konflikt zugeschnittenes Methodenarsenal verwenden, kann eine beschreibende Untersuchung z.B. dem langfristigen Monitoring von Konflikten im Vergleich über Staaten und Regionen dienen. Letzteres verlangt eine Standardisierung der verwendeten Methoden. Als Beispiele lassen sich die Untersuchungen von Amnesty International zu Menschenrechtsverletzungen (z.B. Amnesty International, 2021), das regelmäßige Korruptionsranking von Transparency International (z.B. Transparency International, 2022) und die Datenbanken der Online-Plattform Our World in Data (z.B. Herre, Ortiz-Ospina & Roser, 2013) anführen. Bei deskriptiven Studien hat die Verallgemeinerbarkeit der Ergebnisse auf den angestrebten Gültigkeitsbereich (z.B. eine bestimmte Population) besonders hohe Priorität.

Korrelative Designs. Korrelative Designs bieten sich an, wenn Zusammenhänge zwischen Merkmalen von Interesse sind, um eine Theorie zu entwickeln oder zu prüfen. Um theoretisch einschlägige Zusammenhänge zu bestimmen, werden Untersuchungseinheiten anhand der interessierenden Merkmale beschrieben und miteinander verglichen. Man bezeichnet die Untersuchungseinheiten deshalb auch als Merkmalsträger. Merkmalsträger können individuelle Untersuchungsteilnehmende sein oder Kollektive aller Art (reale Gruppen wie Fanclubs, soziale Kategorien wie Religionsgemeinschaften oder Staaten). Wesentlich ist, dass die Merkmalsträger sich in mindestens zwei Merkmalen unterscheiden, z.B. einem Indikator für materiellen Wohlstand und einem Maß für Friedfertigkeit. In diesem Fall wäre eine bivariate (zwei Variablen umfassende) Zusammenhangsanalyse möglich. In der Regel sind korrelative Designs jedoch multivariat, d.h. mehr als zwei Variablen betreffend, angelegt. Zur Illustration eignet sich eine Studie von Bashiriyeh (2010). Anhand eines Vergleichs von 81 Staaten korrelierte Bashiriyeh die Häufigkeit von Tötungsdelikten mit dem Bruttonationaleinkommen, Werthaltungen aus dem World Values Survey (Schwartz, 2004; Schwartz & Bardi, 2001: Religiosität, Omnipotenz, Absolutismus, Nationalismus, Wettbewerbsorientierung, Autoritarismus) und den Kulturdimensionen Kollektivismus und Machtdistanz nach Hofstede (2001). Aus den neun Prädiktoren ließen sich zwei Faktoren extrahieren, die gemeinsam 20% der Varianz zwischen den Staaten in der Häufigkeit von Tötungsdelikten erklärten. Auf der Ebene der einzelnen Prädiktoren hatte Omnipotenz den stärksten (positiven) und das Bruttonationaleinkommen den zweitstärksten (negativen) Effekt.

Korrelative Designs haben zwei Schwächen. Erstens lassen Korrelationen zwischen querschnittlich erhobenen Variablen keine kausalen Schlüsse zu. Wissen um kausale Prozesse, die hinter einem Zusammenhang stehen, ist für die Einleitung von Präventions- und Interventionsmaßnahmen aber unerlässlich. Zur Illustration der Problematik ziehen wir einen Befund aus der Forschung zu Jugendkriminalität heran (Köster, 2009). Die Häufigkeit, mit der ein Jugendlicher Opfer einer kriminellen Handlung wird (Opferschaft), korreliert mit der Häufigkeit eigener krimineller Handlungen (Täterschaft). Diese Korrelation sagt nichts über das

ursächliche Verhältnis beider Variablen aus. Es könnte sein, dass eigene kriminelle Handlungen dazu führen, dass man eher Opfer von Rachehandlungen wird. Es könnte auch sein, dass Opfererfahrungen die Bereitschaft zu kriminellen Handlungen erhöhen. Beide Variablen könnten sich auch gegenseitig (bidirektional) beeinflussen. Schließlich könnte es sich um eine Scheinkorrelation handeln. Von einer Scheinkorrelation spricht man, wenn zwischen zwei Variablen kein direkter kausaler Zusammenhang besteht, sondern beide durch eine dritte Variable beeinflusst werden. Im unserem Beispiel könnte diese Drittvariable das kriminelle Milieu sein, in dem ein Jugendlicher aufwächst. Trotz ihres Nachteils, die hinter einer Korrelation stehenden kausalen Mechanismen nicht aufdecken zu können, kommen korrelative Designs sehr häufig zum Einsatz – unter anderem auch, weil sie in der Regel deutlich einfacher umzusetzen sind als kausale Designs. Ist man sich der mit korrelativen Designs verbundenen Limitationen bewusst, sind sie ein sehr nützliches Instrument der Erkenntnisgewinnung.

Kausale Designs. In der (friedens-)psychologischen Forschung spielen kausale Fragestellungen eine besonders wichtige Rolle. Plant man beispielsweise die Implementierung aufwändiger Interventionsmaßnahmen zur Konfliktbeilegung, so besteht seitens der Geldgeber ein berechtigtes Interesse an der Frage, ob die Maßnahme tatsächlich wirkt oder ein vergleichbarer Effekt auch mit anderen Mitteln erzielt werden könnte. Das Konzept der Kausalität wird in der Methodenliteratur kontrovers diskutiert. Die folgende Darstellung orientiert sich am in der Psychologie aktuell dominierenden Kausalitätsverständnis; es sei jedoch auch auf die Diskussion des Konzeptes, z.B. aus handlungstheoretischer Perspektive, in den breiteren Sozialwissenschaften verwiesen (Kelle, 2007).

Ein kausaler Zusammenhang erfüllt die Kriterien Kovariation (die Variablen korrelieren), zeitlicher Vorrang (die verursachende Variable ist der beeinflussten Variable zeitlich vorgeordnet) und interne Validität (alternative Erklärungen können ausgeschlossen werden). So definierte kausale Effekte lassen sich am besten mit echten Experimenten nachweisen (Shadish, Cook & Campbell, 2002). Echte Experimente zeichnen sich durch das Vorhandensein einer Vergleichsgruppe (häufig auch Kontrollgruppe genannt) sowie die gezielte Manipulation der unabhängigen Variablen durch die Forschenden aus. Außerdem werden bei echten Experimenten die Untersuchungsteilnehmenden den experimentellen Bedingungen zufällig zugewiesen (Randomisierung). So wird bei einer hinreichend großen Stichprobe (diejenige Teilmenge einer Grundgesamtheit, die an einer Untersuchung teilnimmt) gewährleistet, dass die Häufigkeitsverteilungen von Drittvariablen in allen experimentellen Bedingungen gleich sind und Drittvariablen nicht für den Einfluss der verursachenden (unabhängigen) Variable auf die beeinflusste (abhängige) Variable verantwortlich gemacht werden können.

Ein Beispiel für ein echtes Experiment mit friedenspsychologischem Hintergrund findet sich bei Halperin, Porat, Tamir und Gross (2013). Hier sollte überprüft werden, ob die – aus der Psychotherapie entlehnte – Emotionsregulationstechnik der kognitiven Umbewertung auch als Intervention in Intergruppenkonflikten genutzt werden kann. Zu diesem Zweck wurden jüdische Israelis zufällig in zwei Gruppen unterteilt, von denen die eine ein Training

zur Regulation von Wut erhielt und die andere ihren Emotionen freien Lauf lassen sollte (Kontrollgruppe). Anschließend wurden ihre Einstellungen in Bezug auf den Israel-Palästina-Konflikt abgefragt. Es zeigte sich, dass die Teilnehmer*innen in der Trainingsbedingung im Anschluss signifikant positivere Einstellungen gegenüber Palästinenser*innen äußerten als in der Kontrollgruppe (Studie 1) und dieser Effekt sogar noch 5 Monate später nachweisbar war (Studie 2).

Nicht alle Merkmalsträger eignen sich für Experimente. Mit größeren sozialen Einheiten sind Experimente nicht möglich. Staaten lassen sich nicht zufällig Versuchsbedingungen zuordnen. Dennoch sollte bei jeder Prüfung einer kausalen Hypothese überlegt werden, ob sie experimentell möglich ist. Denn keine andere Methode ist kausalanalytisch so mächtig wie das echte Experiment. Zur Untersuchung von Konflikten auf der Ebene bestehender Gruppen (z.B. Gangs) und sozialer Kategorien (z.B. politische Parteien, Religionsgemeinschaften) können quasi-experimentelle Versuchspläne verwendet werden. Wie in echten Experimenten werden die Gruppen unterschiedlichen Bedingungen (z.B. Informationen, Schulungen, Trainings) ausgesetzt, von denen unterschiedliche Effekte auf eine abhängige Variable, z.B. die Bereitschaft zum Gewaltverzicht bei der Austragung eines Konflikts, erwartet werden. Da die Gruppen nicht randomisiert gebildet wurden, sondern vorliegen, können sie sich bereits im Vorfeld in Drittvariablen unterscheiden, z.B. in Einstellungen, Werthaltungen und Persönlichkeitseigenschaften. Dadurch kommt es zu Korrelationen zwischen den Drittvariablen und der experimentellen Bedingungsvariable. Eine solche Konfundierung beeinträchtigt die interne Validität, wenn die Drittvariablen die abhängige Variable beeinflussen. Es handelt sich bei den Drittvariablen dann um systematische Störvariablen. Zur Sicherung der internen Validität von Quasi-Experimenten können zwei methodische Wege beschritten werden. Der erste besteht in der Erhebung systematischer Störvariablen und ihrer statistischen Kontrolle, z.B. mittels multipler Regressionsanalyse oder Kovarianzanalyse (siehe Abschnitt „Korrelations- und Regressionsanalysen“). Die zweite und zunehmend häufiger verwendete Methode heißt Propensity Score Matching (Austin, 2011). Je nach Zahl der Gruppen werden mit dieser Methode Paarlinge, Tripel, Quadrupel usw. aus Mitgliedern der Gruppen gebildet, die sich in potentiellen Störvariablen so ähnlich wie möglich sind. Beide Strategien verfolgen das Ziel, die Vergleichbarkeit der Gruppen und damit die interne Validität eines Quasi-Experiments zu erhöhen.

Neben der Unterscheidung in echte und Quasi-Experimente wird häufig auch zwischen Labor- und Feldexperimenten differenziert. Im Rahmen von Feldexperimenten wird auf die Kontrolle der Umgebungsvariablen im Labor zugunsten eines möglichst natürlichen Settings verzichtet, was in der Regel zulasten der internen Validität geht, aber dafür deutlich zur externen Validität (der Übertragbarkeit von Studienergebnissen auf natürliche Lebenskontexte; siehe Abschnitt „Was macht einen guten Untersuchungsplan aus?“) einer Untersuchung beiträgt. Feldexperimente sind häufig auch Quasi-Experimente, da meist natürliche Gruppen untersucht werden. Paluck (2009) untersuchte beispielsweise die Auswirkungen einer auf Versöhnung zwischen Hutu und Tutsi zielenden Radiosendung in Ruanda. Sie wählte dazu verschiedene Gemeinschaften aus, von denen ca. die Hälfte ein Jahr lang regelmäßig

das Radioprogramm „New Dawn“ hörte, die andere Hälfte hingegen eine Sendung über Gesundheitsthemen. Es handelte sich bei den Teilnehmenden um intakte Gemeinschaften, die regelmäßig zum Hören der Radiosendung an einem auch sonst üblichen Treffpunkt, z.B. im Dorfgemeinschaftshaus, zusammenkamen. Nach Abschluss der Intervention wurden Interviews, Fokusgruppen und Verhaltensbeobachtungen (siehe Abschnitt „Datenerhebung“) mit den Teilnehmenden durchgeführt und es zeigte sich ein gewisser Einfluss der Sendung auf wahrgenommene soziale Normen und Verhaltensweisen, nicht jedoch auf persönliche Einstellungen der Teilnehmenden in Bezug auf eine Reihe konfliktbezogener Themen.

Während Korrelationen zwischen gleichzeitig (querschnittlich) erhobenen Variablen grundsätzlich keine kausalen Schlüsse erlauben, sind Zusammenhänge zwischen zeitlich verzögert (längsschnittlich) erhobenen Variablen eher kausal interpretierbar. Längsschnittliche Designs können z.B. die Form eines sogenannten kreuzverzögerten Längsschnittplans (Cross-lagged-panel-design; Kearney, 2017) annehmen. Mindestens zwei Variablen X und Y werden mindestens zweimal (zu T1 und T2) in einem zeitlichen Abstand erhoben, in dem mit differentiellen (zwischen Merkmalsträgern unterschiedlichen) Veränderungen von Y gerechnet werden kann, die durch X kausal erklärt werden sollen. Vorhergesagt werden die Veränderungen von Y aus der zu T1 gemessenen Ursache X. Wenn sich beide Variablen über den betrachteten Zeitraum differentiell verändern, können auch reziproke kreuzverzögerte Effekte auftreten, von denen auf eine wechselseitige kausale Beziehung zwischen X und Y geschlossen wird.

Hierarchische Designs. Verhalten in Konfliktsituationen hängt von äußeren Bedingungen (Situation, Kontext) und Merkmalen der beteiligten Akteure ab. Die Friedenspsychologie interessiert sich vor allem für Konflikte zwischen Kollektiven (Gruppen, Interessensgemeinschaften, Staaten). Um diese zu verstehen, kann es unzureichend sein, nur Merkmale der individuellen Akteure zur Erklärung ihres Verhaltens heranzuziehen. Denn auch Merkmale der Gruppen, denen sie angehören, können sich auf ihr Verhalten auswirken. Wie eine Studienserie zum Einfluss von Intergruppenkontakt auf individuelle Vorurteile zeigt, kann der Erklärungswert von Gruppenmerkmalen denjenigen individueller Merkmale durchaus übertreffen (Christ et al., 2014). Designs, die individuelle und Gruppenmerkmale berücksichtigen, nennt man hierarchische Designs oder Mehrebenendesigns. Die Unterscheidung von Ebenen kann auch aufseiten der äußeren Umstände (Situation, Kontext) sinnvoll sein, um einen Konflikt umfassend zu verstehen. Beispielsweise können zu Konflikten zwischen den Fanclubs zweier Fußballvereine Merkmale einzelner Mitglieder (Ebene 1) wie ihre Verträglichkeit oder ihre Identifikation mit dem Verein beitragen, aber auch Merkmale der Clubs wie deren Leitbilder, Mottos und Selbstverständnisse (Ebene 2). Aufseiten der äußeren Umstände können Merkmale des Spielverlaufs (Fouls, Schiedsrichterentscheidungen) und das Spielergebnis (Ebene 1) einen Konflikt anheizen. Zusätzlich dürften die Wettbewerbsgeschichte der Vereine (Gewinn-Verlust-Bilanz, Status in einer Liga) und die Konfliktgeschichte ihrer Fanclubs (Ebene 2) eine Rolle spielen.

Hierarchische Designs eignen sich auch für kulturvergleichende Untersuchungen. Zur Veranschaulichung eines Mehrebenendesigns mit kulturvergleichenden Elementen kann

eine Studie von van Assche, Roets, de Keersmaecker und van Hiel (2017) dienen. Ziel der Studie war, mehr über den Zusammenhang zwischen rechtsgerichteten ideologischen Einstellungen und verschiedenen Vorurteilen gegenüber Fremdgruppen (bei Individuen, Ebene 1) herauszufinden. Dabei wurde unterschieden zwischen individuellen ideologischen Einstellungen (Ebene 1) und dem regional bzw. national vorherrschenden politischen Klima (Ebene 2). Als Datengrundlage dienten der European Social Survey (Studie 1) und der World Values Survey (Studie 2). Als Cross-level-Effekt zeigte sich, dass individuelle Vorurteile (Ebene 1) nicht nur – wie aus vorherigen Studien bekannt – mit individuellen ideologischen Einstellungen (Ebene 1) zusammenhängen, sondern auch durch das vorherrschende ideologische Klima im jeweiligen regionalen bzw. nationalen Kontext (Ebene 2) beeinflusst werden. Menschen, die in rechtsgerichteten sozialen Kontexten leben, haben in der Regel auch mehr Vorurteile gegenüber Fremdgruppen. Über solche Cross-level-Effekte hinaus sind in hierarchischen Designs auch sogenannte Cross-level-Interaktionen von großem Interesse. Diese besagen, ob der Einfluss eines Merkmals auf einer Ebene durch ein Merkmal auf einer anderen Ebene moderiert, also verstärkt oder abgeschwächt wird. In der eben erwähnten Studie lag eine solche Cross-level-Interaktion vor, und zwar dahingehend, dass individuelle ideologische Einstellungen in rechtsgerichteten Kontexten einen schwächeren Einfluss auf Vorurteile haben als in linksgerichteten Kontexten.

Was macht einen guten Untersuchungsplan aus?

Gute Forschung verwirklicht den unter den Randbedingungen bestmöglichen Untersuchungsplan. Seine Qualität und Umsetzung wird anhand von Gütekriterien beurteilt. Für quantitative und qualitative Untersuchungspläne gelten dabei in der Regel unterschiedliche Bewertungsmaßstäbe, was auf die zugrundeliegenden epistemologischen Grundannahmen zurückzuführen ist (Kapitel „Forschungsparadigmen“, Billmann-Mahecha, n.d.). Paradigmenübergreifend werden von guter Forschung neben den klassischen Gütekriterien auch die Erfüllung ethischer Standards (siehe Abschnitt „Ethische Vorüberlegungen“) sowie Offenheit und Transparenz (siehe nächster Abschnitt „Open Science“) gefordert.

Open Science. Die 2010er Jahre waren geprägt von einer Reihe von Skandalen in der Sozialpsychologie, in denen unter anderem Datenmanipulationen und bewusste Täuschungen von prominenten Sozialpsychologen aufgedeckt wurden (Steffens, 2012). Gleichzeitig nahm die sogenannte „Replikationskrise“ ihren Anfang, als einige Forschende unabhängig voneinander Experimente aus einem Artikel mit dem provokanten Titel „Feeling the future“ (Bem, 2011) nachstellten, in dem vermeintlich parapsychologische Phänomene nachgewiesen wurden. Versuche, die im Paper berichteten Ergebnisse zu replizieren – also unter gleichen Bedingungen erneut nachzuweisen – scheiterten (Ritchie, Wiseman & French, 2012). In den kommenden Jahren systematisch unternommene Replikationsversuche vieler klassischer Untersuchungen aus der Psychologie ließen erhebliche Zweifel an der Replizierbarkeit vieler Studien aufkommen (Open Science Collaboration, 2015). Vor diesem Hintergrund ist das Erstarken der Open Science-Bewegung zu verstehen, deren Ziel die Verbesserung der

Qualität wissenschaftlicher Forschung durch Erwirkung maximaler Transparenz und Nachvollziehbarkeit des Forschungsprozesses und seiner Ergebnisse ist.

Als Orientierung zur Erfüllung entsprechender Open Science-Kriterien können beispielsweise die TOP-Guidelines von Nosek und Kolleg*innen (2015) dienen. Viele der dort geforderten Maßnahmen, wie z.B. die Präregistrierung von Studien oder das zur Verfügung stellen von Originaldatensätzen und Auswertungsskripten in entsprechenden Repositorien (z.B. Open Science Framework), werden mittlerweile auch von den einschlägigen Fachzeitschriften und Einrichtungen der Forschungsförderung wie der Deutschen Forschungsgemeinschaft eingefordert.

Gütekriterien in der quantitativen Forschung. Für die Mehrheit der quantitativ arbeitenden Psycholog*innen bildet der kritische Rationalismus (Popper, 2005) die wissenschaftstheoretische Grundlage empirischer Forschung. Aus diesem Grund konnte sich ein allgemein anerkannter Kanon quantitativer Gütekriterien durchsetzen. Die wichtigsten davon sind die interne Validität, externe Validität, Generalisierbarkeit, Reliabilität / Replizierbarkeit und Teststärke.

Interne Validität. Interne Validität ist gegeben, wenn Effekte und Zusammenhänge „echt“ sind, das heißt, dass sie nicht, auch nicht teilweise, aufgrund von systematischen Störvariablen zustande kommen. Soll z.B. ein Konfliktlösetraining evaluiert werden und findet sich am Ende des Trainings eine hohe Kompetenz der Teilnehmenden, kann nicht automatisch auf die Wirkung des Trainings geschlossen werden. Erstens muss gezeigt werden, dass die Kompetenz der Teilnehmenden über den Zeitraum des Trainings gestiegen ist. Dieser Nachweis erfordert z.B. einen Kompetenztest vor dem Training. Zweitens muss belegt werden, dass die Kompetenz von Personen, die nicht am Training teilgenommen haben, im Zeitraum des Trainings auch nicht (oder zumindest weniger stark) zugenommen hat. Dieser Nachweis erfordert eine Kontrollgruppe. Drittens muss sichergestellt werden, dass sich die Mitglieder der Trainingsgruppe von denen der Kontrollgruppe nicht in Merkmalen unterscheiden, die auch eine Kompetenzzunahme bewirken könnten. Die Motivation, Konfliktlösekompetenz zu erwerben, könnte eine solche Störvariable sein. Es würde eine Konfundierung des experimentellen Faktors (Training vs. kein Training) mit einem individuellen Faktor (Motivation) vorliegen, die sich am besten durch Randomisierung vermeiden ließe (siehe Abschnitt „Kausale Designs“). Interne Validität ist besonders dann relevant, wenn eine Studie einen Anspruch auf kausale Schlussfolgerungen erhebt. Auch die Schlüsse von Korrelationsstudien können durch einen Mangel an interner Validität bedroht werden, wenn beispielsweise Scheinkorrelationen vorliegen (siehe oben). Aus diesem Grund wird in Korrelationsstudien häufig für einschlägige Drittvariablen kontrolliert.

Externe Validität. Externe Validität ist gegeben, wenn Effekte und Zusammenhänge, die unter kontrollierten Laborbedingungen ermittelt wurden, auch in natürlichen Lebenskontexten gelten. Ökonomische Spiele wie das Diktator-, Ultimatum- oder Drei-Personen-Spiel können als Beispiele dienen (Becker, 1976; Camerer, 2003). Im Drei-Personen-Spiel bekommt eine Person A Geld, das sie zwischen sich und einer Person B nach Belieben aufteilen darf.

Person C bekommt ebenfalls Geld, das sie entweder behalten oder verwenden darf, um Person A für eine unfaire Aufteilung zu bestrafen oder Person B für eine unfaire Behandlung durch A zu entschädigen (Baumert, Schlösser & Schmitt, 2014). Zahlreiche Studien haben in dieser Situation eine Präferenz für die altruistische Bestrafung einer unfairen Person A durch Person C gefunden. Personen in der Rolle von C verzichten auf Geld, um die von A gegenüber B verletzte Gerechtigkeit wiederherzustellen. Extern valide wäre dieser Befund, wenn Menschen oder Gruppen in der Rolle von C auch im realen Leben eigenes Geld einsetzen würden, um Menschen oder Gruppen in der Rolle von A für eine unfaire Behandlung von anderen Menschen oder Gruppen in der Rolle von B zu bestrafen.

Generalisierbarkeit. Unter der Generalisierbarkeit eines Befundes werden seine Reichweite und Verallgemeinerbarkeit verstanden. Erstrebenswert sind Untersuchungspläne, die hohe Generalisierbarkeit gewährleisten oder das Ausmaß der Generalisierbarkeit erkennen lassen. Externe Validität ist eine Variante von Generalisierbarkeit. Externe Validität ist gegeben, wenn Befunde, die unter kontrollierten Laborbedingungen erzielt wurden, auf die Welt außerhalb (extern) des Labors generalisiert werden können. Neben externer Validität gibt es weitere Formen der Generalisierbarkeit. Beispielsweise bedeutet historische Generalisierbarkeit, dass ein Befund unabhängig davon gilt, wann er erzielt wurde. Bei den Experimenten von Milgram (1974) zur Verabreichung schmerzhafter Stromschläge auf Anweisung einer Autorität scheint die historische Generalisierbarkeit über die letzten 50 Jahre gegeben zu sein (Doliński et al., 2017). Generalisierbarkeit kann weiterhin die Frage betreffen, ob die Befunde einer Untersuchung auch für Personen oder Gruppen gelten, die an der Untersuchung nicht teilgenommen haben. Angenommen, eine intern valide Untersuchung in Schule A habe ergeben, dass mit Rollenspielen Empathie und Perspektivenübernahme gesteigert werden können und dadurch die Konfliktbearbeitungskompetenz. Generalisierbar wäre dieser Befund, wenn er in anderen Schulen ebenfalls gelten würde. Dies ist nicht selbstverständlich, denn Schulen unterscheiden sich in vielfältiger Weise und einige Unterschiede könnten den Effekt des Rollenspiels moderieren, also stärken oder schwächen. Es könnte z.B. sein, dass das Training in Schulen, deren Curricula viel Projektarbeit vorsehen, weniger wirkt als in Schulen, die überwiegend Frontalunterricht praktizieren. Der Unterschied zwischen den Schulen könnte daher rühren, dass Projektunterricht Empathie und Perspektivenübernahme schult und mit Rollenspielen kein zusätzlicher Lerneffekt erzielt werden kann. Die Unterrichtsform wäre in diesem Beispiel ein systematischer Moderator der Generalisierbarkeit.

Reliabilität / Replizierbarkeit. Das Kriterium der Reliabilität ist in dem Maße erfüllt, in dem eine Versuchsanordnung gewährleistet, dass ein Ergebnis sich unter gleichen Bedingungen replizieren lässt, es sich also nicht um einen zufälligen Befund handelt, sondern einen systematischen. Typischerweise wird der Nachweis der Reliabilität über die schließende Statistik erbracht (siehe Abschnitt „Quantitative Auswertungsverfahren“). Diese beruht auf der Idee, dass ein wahrer Effekt oder Zusammenhang, der in der Population existiert, in einer Stichprobe aus dieser Population nur näherungsweise richtig geschätzt wird. Deshalb weichen die Befunde mehrerer Stichproben voneinander ab. Verantwortlich hierfür sind Störva-

riablen, die im Unterschied zu konfundierten Drittvariablen (siehe kausale Designs und interne Validität) und systematischen Moderatoren der Generalisierbarkeit unsystematisch wirken. Der Standardfehler eines Effekt- oder Zusammenhangskoeffizienten gibt an, wie stark dieser über Stichproben gleicher Größe aus derselben Population streut (Eid, Gollwitzer & Schmitt, 2015). Der Standardfehler ist somit ein Maß für die Reliabilität eines Effekts oder Zusammenhangs und fließt in Signifikanztests ein.

Teststärke. Damit ein wahrer Effekt oder Zusammenhang in einer Population anhand einer Stichprobe aus dieser Population entdeckt werden kann, muss die Teststärke (Power) ausreichend sein. Sie bezeichnet die Wahrscheinlichkeit, dass ein in der Population tatsächlich bestehender Zusammenhang oder Effekt in der Stichprobe entdeckt wird. Die Teststärke hängt von der Größe der Stichprobe sowie der Stärke des Effekts oder Zusammenhangs ab. Je stärker ein Effekt oder Zusammenhang in der Population ist, desto wahrscheinlicher lässt er sich in einer Stichprobe entdecken. Zu einem guten (quantitativen) Untersuchungsplan gehört deshalb eine Poweranalyse oder Stichprobenumfangsplanung auf der Basis der bekannten Zusammenhänge zwischen der Effektstärke, der Stichprobengröße und dem vorab festgelegten Risiko, fälschlich die Annahme zu verwerfen, der Effekt oder Zusammenhang sei in der Population nicht vorhanden (Alpha-Fehler). Die Grundlagen der Poweranalyse werden in jedem Statistikbuch erklärt (z.B. Eid et al., 2015). Zur Berechnung der Teststärke eignet sich das Programm G-Power (Faul, Erdfelder, Buchner & Lang, 2009).

Qualitative Gütekriterien. Im Gegensatz zur quantitativen Forschung, wo der Kanon der relevanten Gütekriterien seit Jahrzehnten gut etabliert ist, herrscht in der qualitativen Forschung deutlich weniger Einigkeit bezüglich der Beurteilungsstandards guter Forschung. Ein wichtiger Grund dafür ist die in der qualitativen Forschung vorherrschende Vielfalt erkenntnistheoretischer Paradigmen, welche die Ausgangsbasis für die Beurteilung von Qualität bilden. Letztlich bestimmen die Ziele einer Untersuchung und die Vorannahmen darüber, was Erkenntnis bedeutet bzw. in welcher Form Erkenntnis überhaupt möglich ist, das Urteil darüber, ob eine Untersuchung in adäquater Art und Weise durchgeführt wurde. Dementsprechend gibt es in der qualitativen Forschung unterschiedliche Herangehensweisen an das Thema Güte (Steinke, 1999).

Eine recht kontroverse Auffassung ist die (postmoderne) Position, dass Gütekriterien generell nicht mit qualitativer Forschung vereinbar seien und deshalb gänzlich auf sie verzichtet werden sollte (Steinke, 1999). Diese Auffassung liegt in einer radikal sozial-konstruktivistischen Sicht auf die Welt begründet, welche die Existenz einer objektiven Realität ausschließt. Demzufolge würde die Anwendung von Kriterien zur Bestimmung der Güte einer epistemologischen Aussage das zugrundeliegende erkenntnistheoretische Paradigma ad absurdum führen.

Am anderen Ende des Spektrums stehen Versuche der Anwendung von quantitativen Gütekriterien in qualitativen Forschungsdesigns. Relativ unstrittig ist, dass die Gütekriterien aus der quantitativen Forschung keine direkte Anwendung auf qualitative Forschungsprojekte finden können (Steinke, 1999). Ein Beispiel stellt die klassische Konzeption der internen Validität dar, deren Ziel darin liegt auszuschließen, dass andere als die a priori angenommen

Variablen einen beobachteten Zusammenhang erzeugen (s.o.). Die damit einhergehende Kontrolle möglichst vieler Randbedingungen würde die Essenz vieler qualitativer Verfahren infrage stellen, deren Stärke gerade der möglichst direkte, authentische Zugang zu den Beforschten und ihrem Umfeld darstellt.

An Stelle einer direkten Übertragung stehen stattdessen Bestrebungen einer Reformulierung und Anpassung der quantitativen Gütekriterien für die qualitative Forschung (z.B. Miles & Hubermann, 1994). So wird beispielsweise interne Validität breiter gefasst und mehr im Sinne von Glaubwürdigkeit bzw. Authentizität verstanden (im weitesten Sinne auch als „Wahrheitsgehalt" – auch wenn dies für viele qualitativ Forschende ein schwieriges Konzept ist). Hier wird vor allem die Ergebnisdarstellung dahingehend überprüft, ob sie kritischen Hinterfragungen standhält, ob alternativen Erklärungen hinreichend Raum gegeben wurde, ob die Darstellung lückenlos und Konzepte systematisch miteinander verbunden sind.

Mit externer Validität bzw. Transferierbarkeit oder auch Passung qualitativer Studiendesigns ist – ähnlich wie bei quantitativen Designs – die Übertragbarkeit der Befunde über den konkreten Kontext der Studie hinaus gemeint, aber auch die theoretische Verallgemeinerbarkeit der Ergebnisse bzw. ihre Verbindung zu übergeordneten Theorien.

Das Kriterium der Objektivität bzw. Bestätigbarkeit qualitativer Untersuchung steht für einen reflektierten, transparenten Umgang mit der Subjektivität der Forschenden und die Frage, ob andere Forschende zu vergleichbaren Ergebnissen kommen würden. Dazu gehören beispielsweise die detaillierte Offenlegung des Forschungsprozesses, die Dokumentation der angewandten Methoden und die Darlegung persönlicher Werte und Vorannahmen (siehe Abschnitt „Reflektierte Subjektivität").

Reliabilität bzw. Verlässlichkeit (auch: Auditierbarkeit) steht für die Konsistenz und Stabilität des Forschungsprozesses über die Zeit und über verschiedene Forschende und Methoden hinweg. Hier stellen sich z.B. Fragen nach der Passung von Fragestellung und Studiendesign, Explikation der Rolle der Forschenden sowie der konzeptuellen Klarheit des Designs. Auch wird hier die Forderung nach Protokollen für die Datenerhebung, Überprüfung der Schlussfolgerungen verschiedener Beobachter*innen/Auswerter*innen, Konvergenz unterschiedlicher Datenquellen sowie prozessinternen Peer-Reviews formuliert.

Der Vollständigkeit halber sei erwähnt, dass Miles und Hubermann (1994) auch noch das Kriterium der Nützlichkeit (bzw. Anwendbarkeit oder Handlungsorientierung) ergänzen, welches auf die Vorteile anspielt, die eine Studie für Forschende, Beforschte und Konsument*innen der Forschungsergebnisse bereithält (siehe Abschnitt „ethische Vorüberlegungen").

Die Anwendung „klassischer" Gütekriterien auf die qualitative Forschung trifft bei vielen qualitativ Forschenden auf Skepsis (Steinke, 1999; Flick, 2010). Zum einen bleiben Zweifel bestehen, ob diese – auch in revidierter Fassung – qualitativer Forschung wirklich gerecht werden können. Zum anderen entstehen durch die Verwendung der entsprechenden Begrifflichkeiten in qualitativen Kontexten konzeptuelle Unschärfen, da diese doch anders interpre-

tiert werden, als es in der quantitativen Forschung der Fall ist (und zudem auch sehr unterschiedliche Interpretationen innerhalb der qualitativen Forschungscommunity vorliegen, Miles & Hubermann, 1994; Maxwell, 1992). Stattdessen finden sich zahlreiche Bestrebungen, eigene Gütekriterien für die qualitative Forschung zu entwickeln. Aufgrund der Vielfalt an qualitativen Ansätzen (Mey & Mruck, 2010a) sind sowohl verfahrensspezifische (z.B. Strauss & Corbin, 1990) als auch methodenübergreifende (sog. Kernkriterien, z.B. Steinke, 1999) Vorschläge gemacht worden. Eine vollständige Erläuterung aller so generierten Kriterien würde den Rahmen dieses Kapitels überschreiten, weshalb im Folgenden nur eine Auswahl vorgestellt wird (eine detailliertere Übersicht findet sich z.B. bei Steinke, 1999):

Kommunikative Validierung. Die kommunikative Validierung, häufig auch als „member check" bezeichnet, beschreibt das Rückspielen der erhobenen Daten und ihrer Interpretationen an die Beforschten selbst, sodass diese eine Rückmeldung bezüglich ihrer Gültigkeit geben können.

Triangulation. Die Triangulation (Flick, 2011) ist Validierungsinstrument und Methode zugleich. Die Grundidee dahinter ist, dass ein Phänomen klarer wird und Verzerrungen minimiert werden können, wenn man es aus unterschiedlichen Perspektiven betrachtet. Zur Triangulation können dabei die Sichtweisen unterschiedlicher Forschender dienen, häufig werden aber auch diverse Methoden (z.B. Interviews, Beobachtungen; aber auch qualitative und quantitative Methoden) oder sogar theoretische Perspektiven trianguliert.

Intersubjektive Nachvollziehbarkeit. Da qualitative Forschung in der Regel nicht standardisiert – und damit nicht im klassischen Sinne replizierbar – ist, wird stattdessen das Kriterium der intersubjektiven Nachvollziehbarkeit vorgeschlagen. Dieses beinhaltet zunächst, dass der Forschungsprozess akribisch dokumentiert wird, sodass eine solide Beurteilungsgrundlage vorliegt. Des Weiteren wird die Interpretation in Gruppen empfohlen, da so einseitige Sichtweisen reduziert werden und ein Mindestmaß an Nachvollziehbarkeit gewährleistet wird. Schließlich empfiehlt sich der Einsatz sog. „kodifizierter" Verfahren. Die qualitative Forschung hält eine Vielfalt klar umrissener, regelgeleiteter Methoden bereit (z.B. die qualitative Inhaltsanalyse, siehe Abschnitt „Qualitative Inhaltsanalyse"), die zu befolgen ratsam ist. Nicht selten findet man bei unerfahrenen Anwender*innen qualitativer Forschung z.B. den Hinweis, dass die Daten „kodiert" worden seien, was eine sehr unpräzise Aussage ist, da „Kodieren" per se keine Auswertungsmethode darstellt. Der Hinweis auf ein etabliertes (Kodier-)Verfahren mit entsprechender Quellenangabe ist für Konsument*innen qualitativer Forschung eine sehr hilfreiche Angabe.

Empirische Verankerung. Die Entwicklung von Hypothesen oder Theorien sollte sich immer möglichst nah an den zur Verfügung stehenden empirischen Daten orientieren. In diesem Rahmen findet auch ein gewisses Maß an deduktiver Überprüfung der getroffenen Aussagen statt, indem Vorhersagen formuliert werden, die dann wiederum anhand des vorliegenden (oder im weiteren Forschungsprozess neu erhobenen) Materials getestet werden. Als besonders streng gelten dabei Versuche, die eigenen Befunde zu falsifizieren, explizit nach Widersprüchen und Gegenbeispielen in den Daten zu suchen und die daraus gewonnen

Erkenntnisse in die Analyse zu integrieren und somit eine stärkere Verankerung in den Daten zu erzielen.

Reflektierte Subjektivität. Das Kriterium der reflektierten Subjektivität geht mit dem Zugeständnis einher, dass Forschende nichts als Tabula Rasa forschen, sondern stets ihre subjektiven Sichtweisen, Interessen und Agenden mitbringen. Mit dieser Erkenntnis geht die Forderung nach entsprechender Transparenz einher, die eigene Subjektivität, den eigenen Standpunkt als Forschende offen zu legen und den Umgang damit im Forschungsprozess zu reflektieren. So werden Dritte befähigt, sich ein Bild von der Rolle der Forschenden und ihrem möglichen Einfluss auf die Schlussfolgerungen der Studie zu machen.

Sampling

Da Ressourcen und Feldzugang von Forschenden limitiert sind und Untersuchungen an gesamten Populationen im Normalfall weder möglich noch nötig sind, ergibt sich für empirische Untersuchungen die Notwendigkeit der Stichprobenziehung bzw. Fallauswahl. Diese Notwendigkeit stellt sich unabhängig davon, ob die Studie qualitativ oder quantitativ angelegt ist. In der Regel wird zwischen Wahrscheinlichkeitssampling, unsystematischem Sampling und absichtsvollem Sampling unterschieden (Boehnke, Lietz, Schreier & Wilhelm, 2011). Die Güte einer empirischen Untersuchung ist auch dadurch bedingt, ob die gewählte Samplingstrategie zum Ziel der Studie passt und wie diese umgesetzt wurde. Untersuchungseinheiten sind in friedenspsychologischen Untersuchungen in der Regel Individuen, auch wenn in gewissen Forschungsdesigns ggf. größere Einheiten (Gruppen, Organisationen, Staaten) relevant sein können.

Neben der Wahl der Samplingstrategie spielen in friedenspsychologischen Untersuchungen auch ethische und forschungspraktische Überlegungen bezüglich der Rekrutierung von Teilnehmenden eine wichtige Rolle. So kann es beispielsweise sein, dass im Zentrum des Interesses eine Personengruppe steht, bei der die Population von Vorneherein sehr klein ist, z.B. bei Kriegsverbrecher*innen oder Terrorist*innen. Auch kann man es mit Personen zu tun haben, die schwer zu identifizieren sind (z.B. Menschen mit Traumaerfahrungen), oder Menschen, die ein berechtigtes Interesse haben, möglichst anonym zu bleiben (z.B. Menschen ohne legalen Aufenthaltsstatus). Die Frage des Feldzugangs bzw. der Rekrutierbarkeit von Teilnehmenden sollte im Rahmen jedes Untersuchungsplans sorgfältig bedacht und kritisch hinterfragt werden. Forschende neigen dazu, die Bereitschaft zur Teilnahme bzw. die Auffindbarkeit von geeigneten Forschungssubjekten zu überschätzen. Auch in interkulturellen Kontexten, die in friedenspsychologischen Studien häufig anzutreffen sind, kann der Zugang zu Teilnehmenden eine Herausforderung sein. So herrschen in unterschiedlichen Ländern unterschiedliche Normen bezüglich der Teilnahme an wissenschaftlichen Studien und die Vertrautheit mit verschiedenen Erhebungsinstrumenten variiert. Die Tatsache, dass Forschende als Ausländer*innen identifizierbar sind, kann ebenfalls Auswirkungen auf die Rekrutierung haben (positiv, z.B. im Sinne von Neugier oder positiven Stereotypen, aber auch negativ, im Sinne von Vertrauensverlusten, z.B. vor dem Hintergrund einer vorhergehenden

Kolonialisierung). Letzteres ist auch ein ethisches Thema (siehe Abschnitt „Ethische Vorüberlegungen“). Forschende sind also gefordert, das Thema Rekrutierung hinreichend zu bedenken und ggf. kreative Strategien zu entwickeln, um eventuelle Herausforderungen vorausschauend zu navigieren.

Wahrscheinlichkeitssampling. Diese Art des Sampling gilt als Königsweg in der quantitativen Forschung, wo das Ziel der Stichprobenziehung die Verallgemeinerbarkeit der Aussagen, die an einer Stichprobe gewonnen wurden, auf eine Grundgesamtheit ist. Aus diesem Grund wird angestrebt, dass die Stichprobe repräsentativ für die Grundgesamtheit ist, aus der sie entstammt. Bei hinreichender Größe der Stichprobe kann das Ziel der Repräsentativität durch die zufällige Ziehung einer Stichprobe gewährleistet werden. Das Ziel der zufälligen Stichprobenziehung wird in der Forschungsrealität jedoch selten erreicht. Es scheitert daran, dass nicht alle Menschen einer Population für eine wissenschaftliche Untersuchung erreichbar sind und nur ein Teil der erreichten Menschen bereit ist, an der Studie teilzunehmen.

Zur Beurteilung der Repräsentativität wird häufig auf demographische Variablen wie Alter, Geschlecht und (formale) Bildung zurückgegriffen, weil deren Verteilungen in der Population bekannt sind. Übersehen wird dabei, dass es meistens nicht auf die Repräsentativität einer Stichprobe hinsichtlich demographischer Variablen ankommt. Entscheidend ist vielmehr die Repräsentativität der Stichprobe hinsichtlich der Untersuchungsvariablen und möglicher theoretisch bedeutsamer Drittvariablen. Wenn die Verteilungen der Untersuchungsvariablen in der Population nicht bekannt sind, kann ihre Übereinstimmung mit den Verteilungen in der Stichprobe nicht überprüft werden. Daraus folgt, dass die Repräsentativität einer Stichprobe nur gewährleistet ist, wenn sie zufällig aus der Grundgesamtheit gezogen wurde und es keine selektive Teilnahmebereitschaft gab. Selektiv ist die Teilnahmebereitschaft, wenn sie mit Untersuchungsvariablen oder anderen theoretisch relevanten Drittvariablen korreliert. Bei einer friedenspsychologischen Studie könnte dies z.B. bedeuten, dass Personen, die gewaltfreie Konfliktlösung bevorzugen, mit größerer Wahrscheinlichkeit an einer Studie teilnehmen als Personen, die auf Gewaltfreiheit keinen Wert legen. Das Beispiel zeigt, wie schwer es bei sensiblen Themen wie Konflikt und Gewalt fällt, repräsentative Stichproben zu gewinnen.

Mit seinem Modell des repräsentativen Designs hat Brunswik (1955) das Bewusstsein für die Bedeutung der Repräsentativität jenseits von Personenstichproben geschaffen. Er argumentiert, dass alle Elemente psychologischer Untersuchungen dem Kriterium der Repräsentativität genügen sollten, also z.B. Reize, Situationen, Testaufgaben und Verhaltensweisen. Für eine friedenspsychologische Untersuchung könnte diese Forderung bedeuten, dass bei der Konstruktion eines Instruments zur Messung von gewaltfreier Konfliktbewältigung zunächst die Grundgesamtheit gewaltfreien Konfliktverhaltens definiert werden muss. Diese Aufgabe kann mühsam sein, wenn die beteiligten Forschenden unterschiedliche Vorverständnisse von Gewalt und Gewaltfreiheit mitbringen. Einige würden vielleicht die enga-

gierte Betätigung einer Vuvuzela während eines Fußballspiels als gewaltfreies Wettbewerbsmittel anerkennen, während andere mit Hinweis auf die ohrenbetäubende Lautstärke des Instruments von Körperverletzung sprechen würden.

Unsystematisches Sampling. Unsystematisches Sampling ist kein methodologisch fundierter Samplingansatz, sondern ein den Limitationen der Forschungspraxis geschuldetes Vorgehen. Tatsächlich fällt aber ein großer Teil (friedens-)psychologischer Studien in diese Kategorie. Durch die oben bereits diskutierten Herausforderungen fällt die Gewinnung von Zufallsstichproben häufig schwer und Forschende greifen stattdessen auf Teilnehmende zu, die leicht erreichbar sind, z.B. Psychologie-Erstsemesterstudierende. Insbesondere bei herausfordernden Zielgruppen kann der Einsatz dieser Art von Sampling durchaus legitim sein, die damit verbundenen Limitationen des Untersuchungsplans sollten aber transparent dargelegt werden. Der Einsatz von Quoten (auf Basis von – meist theoretischen – Vorüberlegungen a piori festgelegte Kategorien, die zu einem bestimmten Prozentsatz im Sample vorhanden sein sollten) kann bei diesem Vorgehen helfen, die Generalisierbarkeit des Samples zu verbessern. Einige Anwender*innen von unsystematischen Samplingtechniken argumentieren, dass in vielen psychologischen Studien die theoretische Verallgemeinerbarkeit – und damit die interne Validität – Vorrang hat vor der Repräsentativität der Stichprobe bzw. dem Transfer von Stichprobe auf Population. In diesem Sinne stellt das unsystematische Sampling dann keine Bedrohung der Güte des Designs dar, da die Zielsetzung der Studie nicht auf Generalisierbarkeit gerichtet ist.

Absichtsvolles Sampling. Als absichtsvolles Sampling wird eine Gruppe von Samplingtechniken bezeichnet, die vor allem in der qualitativen Forschung zum Einsatz kommen (aber nicht darauf beschränkt sind). Hier werden Fälle nicht nach dem Kriterium der Repräsentativität ausgewählt, da Verallgemeinerbarkeit auf die Grundgesamtheit nicht das primäre Ziel ist, sondern nach ihrer Relevanz für den Forschungsgegenstand. Als Beispiel sei das theoretische Sampling genannt, welches seinen Ursprung in der Grounded Theory-Methodologie (Glaser & Strauss, 1967) hat. Hier werden Fälle parallel zu einer sich entfaltenden Datenanalyse ausgewählt, und zwar anhand von theoretischen Überlegungen, die ganz spezifisch auf den aktuellen Stand der Analyse zugeschnitten sind. Auf diese Weise ist das theoretische Sampling ein sehr ökonomisches Verfahren, mit dessen Hilfe auch bei kleinen Fallzahlen ein hoch relevantes, informatives Sample generiert werden kann. Statt auf eine exakte Repräsentation der Population zielt das theoretische Sampling auf eine möglichst facettenreiche Repräsentation des zu untersuchenden Phänomens (Morse, 2007).

Datenerhebung

Neben der Wahl eines passenden Untersuchungsplans ist die Wahl geeigneter Instrumente zur Erhebung der Untersuchungsvariablen die zweite methodische Herausforderung in der psychologischen Forschung. Im Laufe ihrer über 100jährigen Geschichte hat die empirische Psychologie eine Fülle von Erhebungsinstrumenten für ihre Konstrukte entwickelt. Überblicke geben einschlägige Lehr- und Handbücher der psychologischen Diagnostik (z.B. Schmidt-

Atzert & Amelang, 2021; Schmitt & Gerstenberg, 2014) und Forschungsmethoden (z.B. Mey & Mruck, 2010a). Die schiere Fülle vorhandener Erhebungsmethoden erschwert die Erstellung systematischer Taxonomien. Im Folgenden begrenzen wir uns auf einige der für die Friedenspsychologie besonders relevanten Verfahren. Wir beginnen mit einem Überblick über das Messen, Skalieren und Testen, welche – in unterschiedlichem Ausmaß – die Grundlage für sämtliche quantitativen Formen der Datenerhebung darstellen. Anschließend folgt eine Betrachtung von direkten Befragungen, indirekten Erhebungsverfahren, Interviews, Beobachtungsmethoden und Archivdaten. Daran anschließend erfolgt eine Reflektion dieser Erhebungsmethoden im Wandel der Zeit, speziell im Hinblick auf die fortschreitende Digitalisierung der Gesellschaft und Trends wie „Big Data“.

Ein bei der Datenerhebung immer mit zu beachtender Umstand sind den Kontext betreffende forschungspraktische Überlegungen. Gerade in friedenspsychologischen Untersuchungen können Studien in forschungsfeindlichen Kontexten angesiedelt sein, beispielsweise unter repressiven Regimen, in denen jegliche Art von Forschung einer staatlichen Genehmigung bedarf (und damit ggf. einer Zensur unterworfen ist), oder in aktiven Konfliktzonen, wo reguläre Forschung nur unter leiblicher Gefährdung der Forschenden (und Beforschten) möglich ist. Solche besonderen Umstände, die in der Friedenspsychologie durchaus häufiger anzutreffen sind, können erheblichen Einfluss auf die Wahl der Erhebungsinstrumente haben. So können persönliche Interviews ggf. auch „unter dem Radar“ durchgeführt werden, wo eine offizielle, im großen Stil verbreitete Onlinebefragung evtl. die Aufmerksamkeit der Behörden auf sich ziehen würde. Oder aber Forschende entscheiden sich für die Erhebung von Daten über das Internet, weil die Umsetzung vor Ort zu riskant wäre. Wichtig ist, dass bei der Entscheidung, ob und in welcher Form Daten erhoben werden, immer auch die Risiken für die Beforschten im Blick behalten werden und der Nutzen der Studie nach strengen ethischen Maßstäben beleuchtet wird. Nicht zuletzt sollte die Wahl der Erhebungsinstrumente auch von den Bedürfnissen und Fähigkeiten der Beforschten abhängen. Forschung mit Kindern bedarf anderer Instrumente als Forschung mit Erwachsenen, einer Analphabetin einen schriftlichen Fragebogen vorzulegen wäre nicht nur aus methodischer Sicht problematisch, sondern auch ethisch sehr fragwürdig.

Messen, Skalieren, Testen

Wer quantitativ arbeiten und erhobene Daten statistisch auswerten möchte, muss sich vor der Datenerhebung mit den Grundlagen des Messens, Skalierens und Testens auseinandersetzen. Unter Messen versteht man die Zuordnung von Objekten zu Symbolen derart, dass die Relationen zwischen den Symbolen (numerisches Relativ) den Relationen zwischen den Objekten (empirisches Relativ) entsprechen. In friedenspsychologischen Anwendungen sind Objekte Personen oder Gruppen. Als Symbole werden Zahlen verwendet, weil sie mathematisch-statistische Operationen ermöglichen. Die Art der Relation zwischen den Objekten und Zahlen definiert das Skalenniveau. Dieses entscheidet, welche Aussagen über die Objekte

eindeutig und welche mathematischen Transformationen der ihnen zugewiesenen Messwerte zulässig sind (Steyer & Eid, 2001). Von Niveau spricht man, weil Skalen unterschiedlich informativ sind. Je höher ihr Niveau, desto informativer ist eine Skala.

Nominalskala. Nominalskalen sind am wenigsten informativ. Sie unterteilen Objekte in gleiche und ungleiche. Gleichen Objekten werden gleiche Zahlen zugeordnet, ungleichen Objekte ungleiche. Beispielsweise könnte man im Nordirlandkonflikt Unionisten den Messwert 1 und Irisch-Republikanern den Messwert 2 zuordnen. Zulässig sind alle Transformationen der Zahlen, die ihre Gleichheit oder Ungleichheit unverändert lassen. Bedeutsam sind Aussagen über die Gleichheit oder Ungleichheit der Objekte.

Ordinalskala. Rang- oder Ordinalskalen informieren zusätzlich zur (Un-)Gleichheit von Objekten über deren Größenordnung, wobei „Größe“ im übertragenen Sinne zu verstehen ist. Ordinalskalen ordnen Objekten Zahlen so zu, dass größeren Objekten eine größere Zahl zugeordnet wird als kleineren Objekten. Beispielsweise könnte man die Fanclubs der 18 Fußballvereine der ersten Bundesliga anhand der Häufigkeit, mit der sie bei Spielen in Schlägereien verwickelt sind, in eine Rangreihe der Friedfertigkeit bringen und ihnen ganzzahlige Messwerte von 1 bis 18 zuweisen. Zulässig sind bei Ordinalskalen alle monotonen Transformationen der Messwerte, die ihre Gleichheit (bei geteilten Rangplätzen) und ihre Größenordnung (bei ungeteilten Rängen) unverändert lassen. Bedeutsam sind Aussagen über die Gleichheit der Objekte und ihre Größenordnung.

Intervallskala. Intervallskalen informieren außer über die (Un-)Gleichheit von Objekten und deren Größenordnung auch darüber, wie unterschiedlich Objekte in einem quantitativen Sinn sind. In unserem Beispiel von Fanclubs bedeuten die Rangzahlen 1 bis 18 nicht, dass die Friedfertigkeitsunterschiede zwischen benachbarten Rangplätzen gleich sind. Erst eine Intervallskala der Friedfertigkeit enthält diese Information. Intervallskalen ordnen Objekten Zahlen so zu, dass die Verhältnisse der Differenzen zwischen den Messwerten zweier Objekte mit den Verhältnissen der Merkmalsunterschiede zwischen ihnen gleich sind. Solche Aussagen sind nur möglich, wenn den Objekten reelle Zahlen zugeordnet werden. Wenn der Friedfertigkeitsunterschied zwischen den Fanclubs A und B doppelt so groß ist wie derjenige zwischen C und D, handelt es sich bei einer Friedfertigkeitsskala dann um eine Intervallskala, wenn der Messwerteunterschied zwischen A und B genau doppelt so groß ist wie der zwischen C und D. Zulässig sind bei Intervallskalen positive lineare Transformationen. Bedeutsam sind Aussagen über die (Un-)Gleichheit und Rangordnung von Objekten und außerdem über die Verhältnisse ihrer Unterschiede, nicht aber über Ausprägungsverhältnisse. Die Aussage, Club A sei doppelt so gewaltbereit wie Club B, ist mit einer Intervallskala der Gewaltbereitschaft nicht möglich. Hierzu bedürfte es einer Verhältnisskala.

Verhältnisskala. Eine Verhältnisskala ordnet Objekten Zahlen so zu, dass Verhältnisse zwischen zwei Zahlen den Verhältnissen der Merkmalsausprägungen entsprechen. Verhältnisskalen haben einen natürlichen Nullpunkt, aber keine natürliche Maßeinheit. Zulässige Transformationen der Messwerte sind auf ihre Multiplikation mit einer positiven Konstante beschränkt. Bedeutsam sind Aussagen, um welchen Faktor ein Objekt A „größer“ oder „klei-

ner“ ist als ein Objekt B. In der Psychologie sind Verhältnisskalen selten. Eine Ausnahme bildet die Reaktionszeit. Sie könnte auf die Messung der Gewaltbereitschaft von Fanclubs angewendet werden, wenn man diese als Zeitpunkt der ersten Handgreiflichkeit eines Fanclubmitglieds nach Anpfiff eines Spiels messen wollte. Damit ist jedoch nichts darüber gesagt, wie valide eine solche Messung von Gewaltbereitschaft wäre.

Absolutskala. Absolutskalen haben einen natürlichen Nullpunkt und eine natürliche Maßeinheit. Die Häufigkeitsskala ist eine in der Psychologie oft verwendete Absolutskala. Sie kann auf alle psychologischen Merkmale angewendet werden, die sich sinnvoll über die Anzahl von Ereignissen messen lassen. So ließe sich die Gewaltbereitschaft eines Fanclubs an der Häufigkeit initiierter Schlägereien messen. Bedeutsam sind Aussagen über die absolute Ausprägung von Merkmalen. Transformationen von Absolutskalen sind unzulässig.

Testtheorien. Die Konstruktion diagnostischer Instrumente, die psychologische Konstrukte auf Intervallskalen messen, geschieht auf der Basis von Testtheorien. Die Bezeichnung dieser Theorien als Testtheorien rührt daher, dass sie bevorzugt an Aufgaben demonstriert und erprobt werden, die objektiv richtig gelöst werden können. Das Paradebeispiel hierfür ist der Intelligenztest. Testtheorien lassen sich aber auf alle diagnostischen Instrumente anwenden. Die bekanntesten und einflussreichsten Testtheorien sind die Klassische Testtheorie (KTT; Lord & Novick, 1968) und die probabilistischen Testtheorien, die auch Item-Response-Theorien genannt werden (IRT; Rost, 2004).

Klassische Testtheorie. Die KTT zerlegt einen Messwert additiv in einen wahren Wert und einen Messfehler:

$$y_{mi} = \tau_{mi} + \varepsilon_{mi}$$

Die Symbole der Modellgleichung bedeuten:

y_{mi}	Messwert *y* der Person *m* gemessen mit Instrument (Item, Test) *i*
τ_{mi}	wahrer Wert τ (tau) der Person *m* bezogen auf Instrument (Item, Test) *i*
ε_{mi}	Messfehlerwert ε (epsilon) der Person *m* bezogen auf Instrument (Item, Test) *i*

Der wahre Wert ist ein theoretischer Wert, den man nicht beobachten kann. Mathematisch definiert ist er als Erwartungswert des Messwerts, gegeben die Person, von der die Messwerte stammen. Könnte man eine Person unendlich oft die gleiche Frage beantworten oder die gleiche Aufgabe lösen lassen, wäre ihr wahrer Wert identisch mit der durchschnittlichen Antwort oder Lösung. Jeder einzelne Messwert ist durch einen Messfehler verfälscht. Dessen Richtung und Größe werden als zufällig angenommen. Deswegen mittelt sich der Messfehler bei unendlich vielen Messungen des gleichen Merkmals aus. Die Modellgleichung der KTT impliziert, dass wahre Werte und Messfehler unkorreliert sind (Steyer & Eid, 2001). Folglich setzt sich die Varianz der Messwerte additiv aus der Varianz der wahren Werte und der Varianz der Messfehler zusammen. Der Anteil der Varianz der Messfehler an der Varianz der

Messwerte ist ein objektives Maß für die Unzuverlässigkeit eines Messinstruments, der Anteil der Varianz der wahren Werte an der Varianz der Messwerte ein objektives Maß für seine Zuverlässigkeit oder Reliabilität.

Das Gros der psychologischen Messinstrumente basiert, zumindest implizit, auf der KTT. Leser*innen mit Interesse an einem vertieften Verständnis der KTT und ihrer Messmodelle müssen wir aus Platzgründen auf entsprechende Lehrbücher verweisen (z.B. Schmitt & Gerstenberg, 2014).

Item-Response-Theorien. Die KTT ist für intervallskalierte Items geeignet. Viele Items, die in der Psychologie verwendet werden, sind jedoch zweiwertig (Frage bejaht oder verneint; Lösung einer Aufgabe richtig oder falsch) und somit nicht intervallskaliert. Für solche Items wurde das Rasch-Modell entwickelt, das einfachste IRT-Modell. Es erklärt die Wahrscheinlichkeit, dass eine Aufgabe gelöst wird bzw. eine Frage bejaht wird. Da sich der Wertebereich der Wahrscheinlichkeit von 0 (Aufgabe wird nie gelöst) bis 1 (Aufgabe wird immer gelöst) erstreckt, handelt es sich trotz der unendlich vielen Abstufungen der Wahrscheinlichkeit nicht um eine Intervallskala, denn eine Intervallskala hat keinen begrenzten Wertebereich.

Das Rasch-Modell spezifiziert die Lösungswahrscheinlichkeit als Funktion der Schwierigkeit einer Aufgabe und der Fähigkeit der Person, der sie gestellt wird. Je schwieriger eine Aufgabe ist, desto unwahrscheinlicher wird sie gelöst. Je fähiger eine Person ist, desto wahrscheinlicher löst sie Aufgaben. Es wird angenommen, dass beide Faktoren die Lösungswahrscheinlichkeit additiv bedingen. Je mehr die Fähigkeit einer Person die Schwierigkeit einer Aufgabe übersteigt, desto wahrscheinlicher wird die Aufgabe gelöst. Je mehr die Schwierigkeit der Aufgabe die Fähigkeit der Person übersteigt, desto unwahrscheinlicher ist die Lösung. Die Lösungswahrscheinlichkeit ist somit eine Funktion der Differenz zwischen der Fähigkeit der Person und der Schwierigkeit der Aufgabe. Damit diese Differenz bedeutsam ist, müssen Fähigkeit und Schwierigkeit auf einer gemeinsamen Intervallskala abgetragen werden. Das Rasch-Modell übersetzt diese Ideen in die folgende logistische Funktion:

$$p(y_{mi} = 1) = \frac{e^{\tau_m - \alpha_i}}{1 + e^{\tau_m - \alpha_i}}$$

Die Symbole der Modellgleichung bedeuten:

$p(y_{mi} = 1)$	Wahrscheinlichkeit, dass Person *m* Item *i* löst oder bejaht und somit der Messwert y_{mi} = 1 beträgt („1“ ist die Kodierung für „gelöst“ oder „bejaht“)
e	Eulersche Zahl (2.718), Basis des natürlichen Logarithmus
τ_m	Fähigkeit τ (tau) der Person *m*

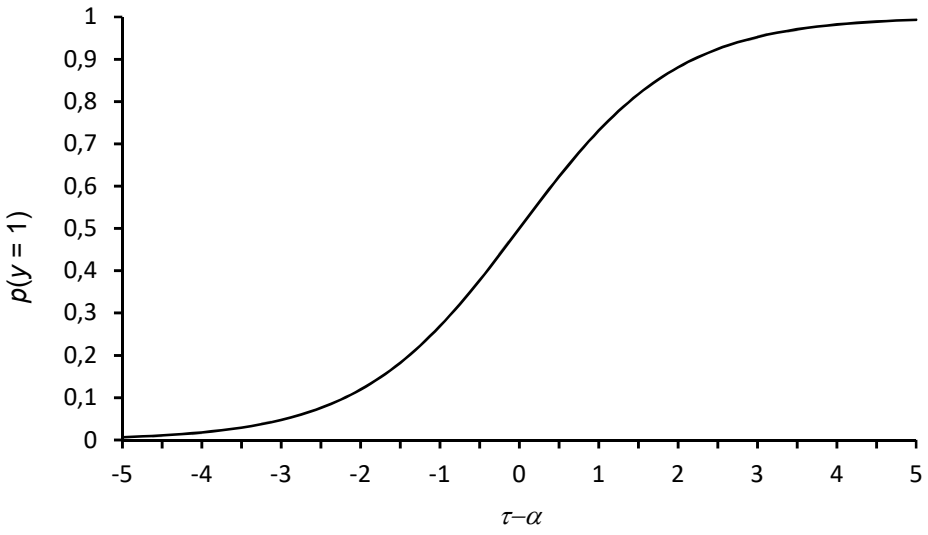


Abb. 1: Ogive des Rasch-Modells

Den geometrischen Verlauf dieser Funktion nennt man Ogive (Abb. 1). Die Lösungswahrscheinlichkeit ist auf der Y-Achse abgetragen, die Differenz zwischen der Fähigkeit und der Aufgabenschwierigkeit auf der X-Achse. Bemerkenswert sind zwei Eigenschaften der Funktion. Erstens nähert sich die Lösungswahrscheinlichkeit dem oberen bzw. unteren Ende der Wahrscheinlichkeitsskala umso stärker an, je mehr die Fähigkeit der Person die Schwierigkeit der Aufgabe übersteigt bzw. unterschreitet. Zweitens wird eine Aufgabe mit 50%iger Wahrscheinlichkeit (p = .50) gelöst, wenn Schwierigkeit und Fähigkeit identisch sind und ihre Differenz Null beträgt.

Da das Rasch-Modell nur einen Parameter enthält, die Differenz zwischen der Fähigkeit und der Aufgabenschwierigkeit, handelt es sich um das einfachste Modell der IRT. Zum tieferen Verständnis des Rasch-Modells und seiner mehrparametrischen Erweiterungen müssen wir aus Platzgründen auf spezielle Literatur verweisen (Rost, 2004; Strobl, 2012).

Gütekriterien von Tests. Wie gut die Konstruktion eines Messinstruments gelungen ist, wird anhand von Gütekriterien beurteilt. Diese werden in die Hauptgütekriterien der Objektivität, Reliabilität, Validität und die Nebengütekriterien der Normierung, Fairness, Ökonomie, Nützlichkeit und Akzeptanz unterteilt. Sind Vergleiche zwischen Gruppen oder Kulturen von Interesse, ist zusätzlich das Gütekriterium der Messäquivalenz von Bedeutung. Auch wenn die Begrifflichkeiten und Inhalte teilweise ähnlich sind, sollten die Testgütekriterien nicht mit den allgemeinen Gütekriterien von Untersuchungsplänen verwechselt werden. Letztere dienen der Gesamtbeurteilung einer Untersuchung, wohingegen sich die Testgütekriterien zur Feststellung der Konstruktvalidität, also lediglich eines Teilbereichs der Untersuchung, eignen. Die folgenden Definitionen sind eng an das Lehrbuch von Schmitt und Gerstenberg (2014) angelehnt. Es bietet, wie andere Lehrbücher auch (z.B. Schmidt-Atzert & Amelang, 2021), vertiefende Erläuterungen der Gütekriterien und ihrer Bestimmung.

Objektivität. Objektiv ist ein Messinstrument in dem Maße, in dem das erzielte Ergebnis unabhängig davon ist, wer das Messinstrument anwendet. Zur Objektivität trägt bei, wenn die Durchführung durch eine Anleitung reglementiert (Durchführungsobjektivität), die Auswertung nach festen Regeln vorgenommen (Auswertungsobjektivität) und das Messergebnis nach festen Regeln interpretiert wird (Interpretationsobjektivität). Die Objektivität ist bei verschiedenen Arten von Instrumenten unterschiedlich ausgeprägt. Beispielsweise lässt sie sich mit standardisierten Fragebögen besser gewährleisten als mit Interviews.

Reliabilität. Methoden der Datenerhebung sind zuverlässig (reliabel) in dem Maße, in dem sie genau messen, was sie messen. Es kommt nicht darauf an, was sie messen, sondern nur darauf, dass das Messergebnis reproduziert werden kann, wenn die Messung mit dem gleichen oder einem gleichwertigen Instrumenten wiederholt wird. Die Reliabilität ist umso stärker eingeschränkt, je stärker die Resultate mehrerer Messungen unter gleichen Bedingungen voneinander abweichen. Das Gütekriterium der Reliabilität hat bei Messinstrumenten somit eine ähnliche Bedeutung wie bei Untersuchungsplänen (siehe oben).

Validität. Valide ist ein Messinstrument in dem Maße, in dem es misst, was es messen soll. Während die Reliabilität nur durch unsystematische Messfehler eingeschränkt wird, sind für die Minderung der Validität neben unsystematischen auch systematische Störvariablen verantwortlich. Unsystematische Störvariablen beinträchtigen also die Reliabilität und die Validität, während systematische Störvariablen die Validität reduzieren, nicht aber die Reliabilität. Angenommen, man fragt die Mitglieder von Fanclubs, wie häufig sie von Mitgliedern anderer Fanclubs im letzten Jahr provoziert wurden. Antworten auf diese Frage können mehrere Ursachen haben. Sie spiegeln teilweise die objektive Zahl von Provokationen wider, dürften darüber hinaus aber auch durch Persönlichkeitseigenschaften wie Ängstlichkeit, die Gedächtnisleistung sowie die Einstellung gegenüber den anderen Fanclubs bedingt sein. Keines dieser Merkmale wird durch das Item vollkommen valide gemessen, da die anderen Merkmale die Antwort ebenfalls beeinflussen. Das Gütekriterium der Validität hat bei Messinstrumenten somit eine ähnliche Bedeutung wie die interne Validität eines Untersuchungsplans (siehe oben).

Normierung. Da es für psychologische Merkmale keine natürlichen Maßeinheiten gibt, müssen die verwendeten Messinstrumente normiert werden, um die Vergleichbarkeit von Messwerten zu gewährleisten. Dazu werden Normierungsstudien an möglichst großen und möglichst repräsentativen Stichproben durchgeführt. Anhand von Normtabellen ist es dann möglich, die Rohwerte in Standardwerte wie T-Werte zu transformieren. Dadurch werden auch die Messwerte unterschiedlicher Instrumente und Konstrukte vergleichbar.

Fairness. Fairness ist ein Aspekt von Validität, der besonders beim Vergleich von Gruppen thematisiert wird. Angenommen, man würde die physische Gewaltbereitschaft von Personen an der Stärke ihrer Faustschläge messen, würden Männer durchschnittlich höhere Werte erzielen als Frauen, auch wenn sie sich nicht in der Gewaltbereitschaft unterscheiden.

Ökonomie. Wenn zwei Instrumente ein Merkmal gleich gut messen, ist jenes ökonomischer, das geringeren Aufwand erfordert und weniger Kosten verursacht.

Nützlichkeit. Ein Messinstrument ist umso nützlicher, je weniger alternative Instrumente es für den gleichen diagnostischen Zweck gibt. Vor der Entwicklung eines neuen Instruments sollte deshalb gründlich nach bereits existierenden Instrumenten recherchiert werden.

Akzeptanz. Wenn zwei Instrumente ein Merkmal gleich gut messen, genießt jenes mehr Akzeptanz, mit dem eine höhere Teilnahmebereitschaft erreicht wird. Bei sensiblen Themen wie Diskriminierung, Feindseligkeit und Gewalt ist Akzeptanz besonders wichtig. Sie lässt sich steigern, indem Items zur Erfassung solcher Merkmale subtil und indirekt formuliert werden.

Messäquivalenz. Messäquivalenz (oder Messinvarianz) ist ein Aspekt von Fairness und damit auch von Validität, der in kulturvergleichenden Untersuchungen eine wichtige Rolle spielt. Messäquivalenz bedeutet, dass die in verschiedenen Kulturen verwendeten (sprachlichen) Varianten eines Messinstruments die gleichen Messeigenschaften aufweisen. Dies ist nicht selbstverständlich, da sich die Semantik von Items durch Übersetzungen ändern und die Bedeutung der Inhalte von Items über Kulturen variieren kann. Man denke an Begriffe wie Korruption, Pünktlichkeit, Gottesfurcht, Patriotismus, Ehre, Partnerschaft oder Gerechtigkeit. Geprüft wird Messäquivalenz anhand von Messmodellen, deren Parameter über die Gruppen verglichen werden. Dabei können unterschiedlich strenge Maßstäbe an deren Invarianz angelegt werden. An vertieftem Verständnis interessierte Leser*innen müssen wir aus Platzgründen auf spezielle Literatur verweisen (van de Vijver & Poortinga, 1997).

Direkte Befragungen

Schriftliche Befragungen in Form von Fragebögen gehören neben Tests zu den am häufigsten verwendeten Messerverfahren der Psychologie. Tests spielen in der Friedenspsychologie kaum eine Rolle, da sie Merkmale messen, für die es einen objektiven Gütemaßstab gibt (z.B. kognitive Fähigkeiten). Fragebögen hingegen sind für viele friedenspsychologische Forschungsfragen geeignet. Die häufige Verwendung dieser Methode hat mehrere Gründe. Erstens sind Fragebögen ökonomisch einsetzbar, da man sie vielen Personen gleichzeitig (auch online) zur Beantwortung vorlegen kann. Deshalb lassen sich Fragebögen zweitens leicht standardisieren. Drittens können sie mit wenig Aufwand konstruiert werden. Viertens verfügt die Psychologie bereits über eine große Menge an Fragebögen, auf die man bei vielen Fragestellungen zurückgreifen kann. Fünftens kennt kaum jemand einen Menschen so gut wie dieser sich selbst, weshalb die allermeisten Fragebögen auf Selbstbeschreibungen zurückgreifen.

Diesen Vorteilen – insbesondere von selbstbeschreibenden – Fragebögen stehen drei Nachteile gegenüber. Erstens ist die Selbstkenntnis von Menschen begrenzt. Zweitens erfordern Selbstbeschreibungen sprachliche Kompetenzen, die bei Kindern erst ab einem bestimmten Alter vorliegen und auch bei Jugendlichen und Erwachsenen nicht immer so gut ausgeprägt sind, dass Verständnisfehler ausgeschlossen werden können. Drittens setzen valide Selbstbeschreibungen Ehrlichkeit voraus. Bei sensiblen Themen neigen Menschen mehr

oder weniger zu beschönigenden Selbstbeschreibungen und sozial erwünschten Antworten, um Missbilligung durch sich selbst und andere zu vermeiden.

Selbstbeschreibungsverfahren gibt es in vielen unterschiedlichen Formaten (Mummendey, 2014), z.B. zur Messung der sozialen Identität, des individuellen und kollektiven Selbstkonzepts sowie von Persönlichkeitseigenschaften, Werthaltungen, Interessen und Motiven. Einstellungsfragebögen können als ihr Prototyp gelten, werden besonders häufig verwendet und eigenen sich auch für die Friedenspsychologie. Beispiele für ihren Einsatz finden sich in vielen der bereits vorgestellten Studien (z.B. Halperin et al., 2013). Geschuldet ist diese Vorrangstellung der Prominenz, Flexibilität und Erklärungsmacht des Einstellungskonstrukts, die Allport bereits 1935 konstatiert hat und die nach wie vor gelten (Bohner & Wänke, 2002). Unter einer Einstellung versteht man den Grad der Zuneigung versus Abneigung gegenüber einem Objekt, die sich in Gedanken, Gefühlen und Verhaltenstendenzen äußern. Einstellungen kann man gegenüber Dingen, Menschen, Ideen, Institutionen, Verhaltensweisen u.a.m. haben. Für die Friedenspsychologie können verschieden Einstellungen von Interesse sein, z.B. gegenüber den Parteien eines Konflikts, ihren Ideologien, Formen der Austragung des Konflikts, dem Verzicht auf Gewalt und Schlichtungsverfahren.

Eine Möglichkeit, einige der Nachteile von Selbstberichten auszugleichen, besteht im Einsatz von Fremdberichten. Bei dieser Methode schätzen Personen, die informierte Aussagen über eine Zielperson treffen können (z.B. Eltern, Lehrer*innen, Expert*innen), beispielsweise Persönlichkeitseigenschaften oder Verhaltensweisen ein. Fremdberichte können als schriftliche Fragebögen vorgelegt werden, aber auch andere Erhebungsformate (z.B. Interviews) sind in diesem Kontext denkbar. Fremdberichte können ebenso wie Selbstberichte Urteilsfehler enthalten, dennoch kann eine Triangulation (siehe Abschnitt „Qualitative Gütekriterien“) beider Perspektiven wertvolle Informationen zu Tage fördern.

Indirekte Erhebungsverfahren

Indirekte oder objektive Erhebungsverfahren sind Tests, die so konzipiert werden, dass sie nicht durch die Teilnehmenden, z.B. einen Mangel an Ehrlichkeit, verfälscht werden können (Häcker, 2017), weil der Zusammenhang zwischen den Inhalten des Tests und dem zu messenden Konstrukt nicht unmittelbar erkennbar oder beeinflussbar ist. Statt auf Selbsteinschätzungen basieren diese Erhebungsverfahren auf Verhaltensmaßen. Dies können physiologische Maße sein, wie z.B. die Hautleitfähigkeit oder die Herzrate bei der indirekten Messung von Emotionen (Mauss & Robinson, 2009, für eine Übersicht), Reaktionszeiten oder verschiedene Leistungsindikatoren, z.B. bei objektiven Persönlichkeitstests (Schuerger, 2008). Für die Friedenspsychologie von besonderem Interesse sind die sogenannten impliziten Einstellungstests, wie der Implizite Assoziationstest (IAT), mit dem unter anderem Vorurteile gegenüber unterschiedlichen Zielgruppen gemessen werden können (z.B. Greenwald, Smith, Sriram, Bar-Anan & Nosek, 2009). Vorurteile sind besonders anfällig für Verfälschung durch soziale Erwünschtheit und Selbsttäuschung, die durch Versuche, direkte Erhebungs-

verfahren subtiler zu gestalten (z.B. die Modern Racism Scale, McConahay, 1986), nur bedingt ausgeräumt werden können. Dagegen haben implizite Erhebungsverfahren, wie z.B. der IAT, den Vorteil, dass Teilnehmende sich kaum erschließen können, wie der Test zu manipulieren wäre. Dieser Vorteil ist gleichzeitig auch ein Nachteil von impliziten Messverfahren, da es Zweifel an deren Validität gibt, die dem komplexen Weg von Konstrukt zu Operationalisierung geschuldet sind (Blanton et al., 2009; Clayton, Horillo & Sniderman, 2020).

Interviews und Gruppendiskussionen

Interviews stellen eine sehr heterogene Gruppe von Verfahren dar, bei denen die mündliche Kommunikation mit den Beforschten im Vordergrund steht. Die Grenze zu Fragebögen verläuft insofern fließend, als dass auch mündlich verlesene Fragebögen gelegentlich als (vollstandardisierte) Interviews bezeichnet werden. Dies kann z.B. bei Erhebungen per Telefon oder mit nicht alphabetisierten Personengruppen der Fall sein. Ihre große Stärke entfalten Interviews jedoch erst in teil- bzw. unstandardisierter Form, da sie den Beforschten eine Stimme geben und auch Äußerungen außerhalb der von den Forschenden antizipierten Pfade ermöglichen. Das Interview hat als Erhebungsform eine lange Tradition, und es existieren zahlreiche verschiedene Formen von Interviews mit teilweise sehr unterschiedlichen methodologischen Unterbauten und Zielsetzungen (für eine umfassendere Übersicht, siehe Mey & Mruck, 2010b). Hierzu zählen beispielsweise das narrativ-biographische Interview (Schütze, 1983), das problemzentrierte Interview (Witzel, 2000), das ethnographische Interview (Spradley, 1979) oder das Expert*inneninterview (Meuser & Nagel, 1991). In vielen, aber längst nicht allen Interviews kommen Leitfäden zum Einsatz, die den Interviewverlauf strukturieren und systematisieren.

Obwohl die Standardsituation im Interview eine aus Interviewer*in und Interviewtem bestehende Dyade ist, sind auch Interviewsettings mit mehreren Teilnehmer*innen üblich (sogenannte Gruppendiskussionen, Przyborski & Riegler, 2010). Auch hier gibt es – ähnlich wie beim dyadischen Interview – eine Vielzahl unterschiedlicher Varianten, von denen die wohl bekannteste die Fokusgruppe ist (Merton, Fiske & Kendall, 1956). Gruppendiskussionen ermöglichen Forschenden die Erhebung multipler Perspektiven in recht kurzer Zeit und können Einblicke in ein komplexes kommunikatives Interaktionsgeschehen geben. Es gilt mittlerweile als gute Praxis, dass sowohl Interviews als auch Gruppendiskussionen – die Zustimmung der Teilnehmenden vorausgesetzt – aufgezeichnet und transkribiert werden. Die Transkripte (auch hier gibt es verschiedene Regelsysteme, Dresing & Pehl, 2010) dienen dann als Grundlage für die weitere Datenanalyse.

Ein Beispiel für den Einsatz von sowohl (standardisierten) Einzelinterviews als auch Fokusgruppen ist die oben bereits vorgestellte Feldstudie zum Radioprogramm in Ruanda (Paluck, 2009). Besonders interessant ist hier die Triangulation der Ergebnisse der Einzelinterviews und der Fokusgruppen: Letztere wurden genutzt, um zu überprüfen, inwieweit die Präsenz anderer Gruppenmitglieder zu einer Veränderung zuvor getätigter Aussagen führte.

Dies war tatsächlich teilweise der Fall. Die Teilnehmenden waren im Rahmen der Gruppendiskussion beispielsweise weniger bereit zuzugeben, dass sie ihrer eigenen Gemeinschaft misstrauten. Die Studie ist zudem ein gutes Beispiel für die Kombination von Erhebungsformen, die eher in der qualitativen Forschung anzutreffen sind, mit statistischen Auswertungsmethoden.

Beobachtungsmethoden

Zum Verständnis eines Konflikts hilfreiche Informationen lassen sich nicht immer aus der Perspektive der Betroffenen mit Fragebögen oder Interviews erheben. Außer den bereits genannten Gründen können auch die Art oder das Stadium eines Konflikts gegen ihre Verwendung sprechen. Wenn Konflikte eskalieren, sind die Konfliktparteien oft so stark involviert, dass sie im Austausch mit Forschenden keinen Sinn sehen. Bei manchen Konflikten sind die Beteiligten nicht erreichbar, unterliegen der Schweigepflicht oder beschreiben den Konflikt bewusst falsch, um Wissenschaft und Medien in die Irre zu führen oder neutrale Instanzen von der eigenen Sicht zu überzeugen, z.B. der, dass die gegnerische Partei die alleinige Verantwortung für die Eskalation des Konflikts trägt.

Wenn solche Umstände die Verwendung von Fragebögen und Interviews erschweren oder vereiteln, sind Beobachtungsmethoden eine wertvolle Alternative. Bei kriegerischen Konflikten gehören sie zum *Standardrepertoire der Überwachung von Waffenstillstandsvereinbarungen*. Beispiele hierfür sind die Beobachtungen der United Nations (UN, Vereinte Nationen) in der militärischen Sperrzone auf dem Golan zwischen Israel und Syrien oder die Beobachtungen der Organisation für Sicherheit und Zusammenarbeit in Europa (OSZE) im Donbass zwischen der Ukraine und Russland. Auch die Kriegsberichtserstattung mit ihrer langen Tradition fällt in diese Kategorie.

Beobachtungsverfahren lassen sich auch im Labor unter standardisierten Bedingungen sinnvoll verwenden. Die Klinische Psychologie erhebt z.B. bei der Behandlung von Partnerschafts- und Familienkonflikten Beobachtungsdaten, um die Problematik besser zu verstehen, den Beteiligten ihr Kommunikationsverhalten zurückzumelden und Fortschritte bei der konstruktiven Konfliktlösung zu erkennen und zu dokumentieren.

Archivdaten, Dokumente und Verhaltensspuren

Unabhängig von den systematischen Datenerhebungsbemühungen von Forschenden produzieren Menschen in ihren natürlichen Lebenswelten unablässig „Daten“, von denen einige aufgezeichnet und archiviert werden und somit auch für die Forschung zugänglich werden. Dazu gehören schriftlich vorliegende Materialien, wie z.B. Zeitungsarchive, Tagebücher oder Briefe, aber auch audiovisuelle Daten, z.B. Fotos, Zeichnungen, Filme oder Tonaufzeichnungen. Durch den digitalen Wandel (siehe auch nächster Abschnitt) hat die Zahl der dokumentierten Verhaltensspuren drastisch zugenommen. Posts, Tweets, und Likes in sozialen Medien zählen ebenso dazu, wie Bewegungsprofile, Surfverhalten, Chatforen und nicht zuletzt

die zahlreichen Fotos und Videos, die tagtäglich durch Smartphones und Digitalkameras erstellt und hochgeladen werden. Für die Friedenspsychologie hat die Digitalisierung Möglichkeiten zur Dokumentation von Konflikten geschaffen, die vor wenigen Jahren noch undenkbar waren. Zum Beispiel entstehen Ton- und Bilddokumente von Konflikten, die neben der politischen und juristischen Aufarbeitung auch für (friedens-) wissenschaftliche Analysen von Wert sein können. Aktuelle Beispiele hierfür sind die Erstürmung des Kapitols der Vereinigten Staaten im Januar 2021, die Rückkehr der Taliban an die Macht in Afghanistan nach dem Abzug der Truppen der USA und anderer Staaten im August 2021 oder der russische Angriff auf die Ukraine im Februar 2022.

Erhebungsmethoden im Wandel der Zeit

Die Weiterentwicklung der Technik, die Digitalisierung und die rasant wachsende Verbreitung sozialer Medien haben unsere Welt nachhaltig verändert und nehmen tiefgreifenden Einfluss auf menschliches Handeln und gesellschaftlichen Wandel. Der Arabische Frühling im Dezember 2010 ist ein frühes Beispiel für die enge Verknüpfung digitaler Lebenswelten mit realpolitischen Konsequenzen. Auch für die Forschung entstehen aus dieser Entwicklung neue Möglichkeiten und Notwendigkeiten. Letztlich sind alle Elemente des Forschungsprozesses von den technischen und sozialen Konsequenzen der Digitalisierung betroffen (z.B. eröffnet diese ganz neue Möglichkeiten bei der Gewinnung von Studienteilnehmer*innen), aber die gravierendsten Veränderungen sind bei den Datenerhebungsmethoden zu beobachten. Bei fast allen Formen der Datenerhebung ist eine Verlagerung ins Digitale zu beobachten: Befragungen und Tests werden online verbreitet und ausgefüllt und auch indirekte Erhebungsverfahren werden computergestützt, teilweise sogar über das Internet (siehe z.B. Project Implicit, n.d.; https://implicit.harvard.edu/implicit/takeatest.html), bereitgestellt. Interviews und Fokusgruppen werden (spätestens seit der Coronapandemie) als Videochats durchgeführt. Auch das Universum der im vorherigen Abschnitt erwähnten, online generierten Verhaltensspuren scheint fast unerschöpflich. Diese Neuerungen bieten zahlreiche Vorteile: Forschung wird effizienter und ökonomischer, neue Zielgruppen, die wissenschaftlichen Untersuchungen sonst eher skeptisch gegenüberstehen, werden erschlossen, die Beobachtung von Verhaltensweisen außerhalb des Labors wird einfacher und durch die schiere Menge an verfügbarem Material entsteht der Eindruck, dass es deutlich leichter ist, überhaupt an Daten zu gelangen, als dies vor der Digitalisierung der Fall war. Gleichzeitig entstehen aber auch neue Herausforderungen bezüglich methodologischer Integration in bestehende Paradigmen, Datenqualität und technischer Umsetzung. Auch ethische (z.B. können Daten ohne Einwilligung der Teilnehmenden zu Forschungszwecken genutzt werden?) und datenschutzrechtliche Aspekte müssen mitbedacht werden.

Ein besonders hervorzuhebender Trend der letzten Jahre sind sogenannte Big-Data-Analysen. Hier wird explizit von den großen Datenmengen Gebrauch gemacht, die Nutzer*innen bei ihren Streifzügen durch das Internet interlassen. Diese Verhaltensspuren werden erfasst, aufbereitet und ausgewertet. Häufig umfassen solche Datensätze tausende oder mehr beforschte Nutzer*innen. Häufig handelt es sich bei Big-Data-Analysen eher um explorative

Studien, bei denen verschiedene statistische Techniken eingesetzt werden, um Muster und Strukturen in den vorliegenden Daten aufzudecken (Grimm, Jacobucci & McArdle, 2017). Insofern könnten Big-Data-Analysen ebenso gut bei den Auswertungsmethoden aufgeführt werden. Interessierten Leser*innen sei der Reader „Big data in psychological research“ (Woo, Tay & Proctor, 2020) empfohlen.

Zur Illustration des Potentials des Big-Data-Ansatzes soll eine Studie von Barberá, Jost, Nagler, Tucker und Bonneau (2015) dienen, die anhand von fast 150 Millionen Tweets von 3,8 Millionen Twitter-Nutzer*innen der Frage nachgingen, wie stark sich die politische Polarisierung von Liberalen und Konservativen in den USA in der Kommunikation in den sozialen Medien widerspiegelt. Sie fanden unter anderem, dass in Bezug auf politische Themen nur sehr wenig Informationsaustausch zwischen den Lagern stattfand, die Grenzen bei anderen Themen aber durchlässig waren. Diese Befunde widersprechen Thesen, die von einer strikten ideologischen Trennung in den sozialen Medien ausgehen.

Datenanalyse

Die mit den oben beschriebenen und weiteren Methoden erhobenen Daten müssen ausgewertet werden. Der folgende Überblick über Auswertungsverfahren ist aus Platzgründen selektiv und bedarf je nach Fragestellung und Datenart der Konsultation weiterführender Literatur (z.B. Eid et al., 2015; Mayring, 2015). Die folgende Darstellung trennt klar zwischen quantitativen und qualitativen Verfahren (Bauer & Blasius, 2019), da diese in der Auswertung – anders als bei den Erhebungsverfahren – kaum gemeinsame Schnittmengen aufweisen.

Quantitative Auswertungsverfahren

Quantitative Auswertungsverfahren lassen sich in deskriptivstatistische (beschreibende) und inferenzstatistische (schließende) unterteilen. Beschreibende Verfahren verdichten die erhobenen Rohdaten in einer Weise, die wesentliche Ergebnisse hervorhebt und unwesentliche ausblendet. Zur Beschreibung der Gewaltbereitschaft von Fanclubs ist die durchschnittliche Gewaltbereitschaft wesentlich und die Gewaltbereitschaft einzelner Mitglieder unwesentlich. Entsprechend vergleicht man in einer Studie die Mittelwerte der Gewaltbereitschaft von Fanclubs und sieht von der Darstellung der individuellen Gewaltbereitschaften ab.

Die schließende Statistik dient zur Prüfung der Generalisierbarkeit von Befunden aus einer Stichprobe auf die Grundgesamtheit, aus der sie gezogen wurde, oder auf andere Stichproben aus dieser Population. In der klassischen frequentistischen Statistik werden hierzu statistische Tests verwendet, die zwei Hypothesen prüfen, die als Nullhypothese und Alternativhypothese bezeichnet werden. Die Nullhypothese besagt, dass es einen bestimmten Effekt, Zusammenhang oder Unterschied in der Population nicht gibt und er in einer Stichprobe nur zufällig gefunden wurde. Die Alternativhypothese besagt, dass der fragliche Effekt, Zusammenhang oder Unterschied in der Population besteht. Das Testergebnis gibt an, mit welcher Wahrscheinlichkeit diese Hypothesen richtig oder falsch sind. Die schließende Statistik kennt eine große Zahl von Tests, die sich in ihren Zielsetzungen sowie ihren Voraussetzungen

an das Skalenniveau und die Verteilungen der Variablen in der Population unterscheiden. Statistiklehrbücher enthalten Übersichten über diese Tests (z.B. Eid et al., 2015). Hypothesen können auch mit der Bayesianischen Statistik beurteilt werden. Zum Vergleich dieser beiden Paradigmen der schließenden Statistik müssen wir aus Platzgründen auf spezielle Literatur verweisen (Tschirk, 2014).

Neben der Unterscheidung von Auswertungsverfahren in beschreibende und schließende werden sie auch anhand der Fragestellung unterteilt, die mit den Daten beantwortet werden soll. Dabei wird grob unterschieden zwischen Verfahren zur Analyse von Unterschieden zwischen Gruppen und Verfahren zur Analyse von Zusammenhängen zwischen Variablen. Diese Klassifikation entspricht den vorrangigen Erkenntnisinteressen der Allgemeinen Psychologie und der Differentiellen Psychologie. Die Unterscheidung ist allerdings künstlich, da die Prototypen beider Analyseverfahren, die Varianzanalyse und die Regressionsanalyse, auf dem Allgemeinen Linearen Modell beruhen (Rencher & Schaalje, 2008). Dennoch werden beide Verfahren in Lehrbüchern meistens getrennt dargestellt.

Varianzanalytische Verfahren. Varianzanalysen dienen dem Vergleich von Mittelwerten aus experimentellen Versuchsplänen. Varianzanalysen zerlegen die Gesamtvarianz einer abhängigen Variablen (AV) in jenen Anteil, der von den experimentellen Bedingungen erzeugt wird, und jenen, der auf individuelle Unterschiede innerhalb der Bedingungen zurückzuführen ist. Die Varianz zwischen den Bedingungen ist systematisch und kann mit der unabhängigen Variablen (UV), dem experimentellen Faktor, erklärt werden. Die Varianz innerhalb der Bedingungen ist unsystematisch, da sie nicht mit der UV erklärt werden kann. Sie wird deshalb als Restvarianz oder Fehlervarianz bezeichnet. Je größer der Anteil der systematischen Varianz, desto wahrscheinlicher kann die Nullhypothese verworfen werden, dass die experimentelle Manipulation wirkungslos war und sich die Bedingungsmittelwerte der AV nur zufällig unterscheiden.

Varianzanalysen werden nach der Zahl der Faktoren (einfaktoriell versus mehrfaktoriell), der Manipulation der UV zwischen versus innerhalb von Personen (Varianzanalyse ohne versus mit Messwiederholung) und der Zahl der abhängigen Variablen (univariate versus multivariate Varianzanalyse) unterschieden. Diese Varianten können kombiniert werden. Bei der klassischen Varianzanalyse ist die AV intervallskaliert. Es gibt aber auch Varianzanalysen für ordinalskalierte AV (Rangvarianzanalysen).

Korrelations- und Regressionsanalysen. Korrelations- und Regressionsanalysen dienen der Ermittlung von Zusammenhängen zwischen Variablen. Wird keine Zusammenhangsrichtung angenommen, also nicht zwischen UV und AV unterschieden, verwendet man Korrelationskoeffizienten. Diese müssen zum Skalenniveau passen. Für die Korrelation binärer Nominalskalen verwendet man den Koeffizienten Phi. Für Rangskalen eignen sich Rangkorrelationskoeffizienten, für Intervallskalen die Produktmomentkorrelation. Die Formeln zur Berechnung dieser Koeffizienten findet man in jedem Statistikbuch. Wurden mehr als zwei Variablen erhoben, überträgt man alle bivariaten Korrelationen in eine Korrelationsmatrix.

Für gerichtete Zusammenhänge zwischen einer UV und einer AV wird statt der Korrelation die Regressionsanalyse benötigt. Zusammenhänge mehrerer UV mit einer AV bestimmt man mit der multiplen Regressionsanalyse. Dieses Verfahren wird in der Psychologie sehr häufig genutzt. Gerichtete Zusammenhänge zwischen mehreren UV und mehreren AV werden mit der multivariaten multiplen Regression ermittelt. Dieser Variante der Regressionsanalyse begegnet man in der Psychologie eher selten. Das Grundprinzip der Regressionsanalyse ist mit dem der Varianzanalyse eng verwandt. Die Varianz einer AV wird zerlegt in jenen Teil, der mit der UV erklärt werden kann und systematisch ist, und jenen Teil, der mit der UV nicht erklärt werden kann, unsystematisch ist und als Fehlervarianz oder Residualvarianz bezeichnet wird. Bei der multiplen Regressionsanalyse ist die Lösung dieser Aufgabe mathematisch aufwändiger als bei der mehrfaktoriellen Varianzanalyse. Denn die UV einer multiplen Regressionsanalyse sind in der Regel untereinander korreliert. Deshalb überschneiden sich ihre Beiträge zur Erklärung der AV. Hingegen gewährleisten vollständig gekreuzte experimentelle Versuchspläne mit gleichen Zellhäufigkeiten die Unabhängigkeit der Faktoren, die deshalb additiv und überschneidungsfrei zur Erklärung der AV beitragen.

Strukturentdeckende und datenreduzierende Verfahren. Wenn die Zahl von Variablen groß ist und man alle bivariaten Korrelationen berechnet, weist die Korrelationsmatrix häufig ein charakteristisches Muster auf. Es lassen sich Gruppen von Variablen identifizieren, die hoch miteinander korrelieren und gleichzeitig niedrig mit den Variablen anderer Gruppen. Dies bedeutet, dass sich die Variablen, die der gleichen Gruppe angehören, überschneiden und ihre Information redundant ist. Da Sparsamkeit in der wissenschaftlichen Beschreibung von Phänomenen als erstrebenswert gilt, greift man bei solchen Mustern zu datenreduzierenden Analysen. Die Faktorenanalyse ist ihre prominenteste Vertreterin. Sie basiert auf der Idee, dass hohe Korrelationen zwischen Variablen zustande kommen, weil sie einen gemeinsamen Faktor teilen, der ihre Korrelationen stiftet. Die Reduktion hoch korrelierter Variablen auf einen gemeinsamen Faktor vereinfacht den Datensatz ohne gravierenden Verlust an Information.

Faktorenanalysen tragen außer zur Reduktion von Variablen auch zum theoretischen Verständnis und zur diagnostischen Nutzung ihrer Zusammenhangsstruktur bei. Angenommen, man trägt zur umfassenden Beschreibung und Diagnose von Feindseligkeit alle erdenklichen Indikatoren zusammen, formuliert diese zu Items eines Fragebogens und legt diesen einer großen Stichprobe von Personen vor. Eine Faktorenanalyse des Datensatzes würde zur Entdeckung der Struktur von Feindseligkeit führen und die theoretische Interpretation dieser Struktur erleichtern. Vermutlich würde sich zeigen, dass Feindseligkeitsindikatoren anhand ihrer Modalität in feindselige Gedanken, Gefühle und Handlungen unterteilt werden können. Diese Erkenntnis würde die Entwicklung eines Fragebogens zur multimodalen Diagnose von Feindseligkeit empfehlen. Tatsächlich ist das beschriebene Vorgehen gang und gäbe in allen Bereichen der differentiellen Psychologie (Persönlichkeit, Einstellungen, Werthaltungen, Motive, Emotionen, Interessen, kognitive Leistungen).

Pfadanalysen und Strukturgleichungsanalysen. Pfadmodelle sind Regressionsmodelle, die außer unabhängigen und abhängigen Variablen auch Mediatorvariablen enthalten.

Mediatoren vermitteln Effekte unabhängiger Variablen (Prädiktoren) auf abhängige Variablen (Kriterien) und sind somit selbst beides. Pfadmodelle können auch mehrere hintereinander geschaltete Mediatoren als mittlere Glieder einer Kausalkette spezifizieren. Außerdem können sie um Moderatoren erweitert werden, die Effekte von UV auf AV stärken oder schwächen. Pfadanalysen stellen mit diesen vier Typen von Variablen ein flexibles Werkzeug zur Untersuchung komplexer Zusammenhangsmodelle bereit. Das Pfadmodell von Duckitt und Sibley (2009) kann als ein friedenspsychogisch einschlägiges Beispiel dienen. Es führt antisoziale Dispositionen auf Merkmale des sozialen (Intergruppen-)Kontextes und auf Persönlichkeitseigenschaften zurück und nimmt an, dass deren Einfluss durch Weltbilder sowie rechtsgerichteten Autoritarismus (RWA: Right-Wing Authoritarianism) und Soziale Dominanzorientierung (SDO) vermittelt wird (Abb. 2).

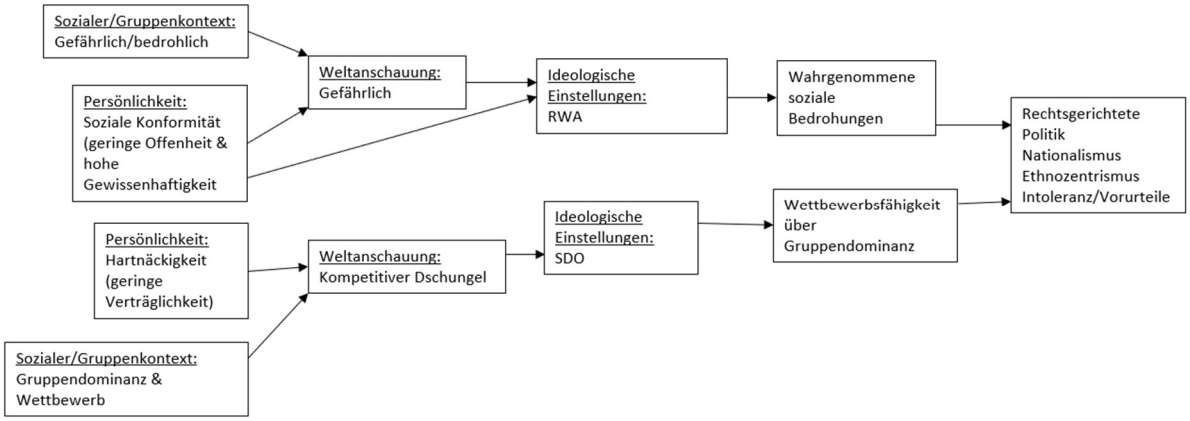


Abb. 2: Pfadmodell in Anlehnung an Duckitt & Sibley (2009)

Pfadmodelle gehören zur Familie der Strukturgleichungsmodelle. Pfadmodelle sind durch eine Menge von Regressionsgleichungen mathematisch eindeutig formuliert. Auch das Modell gemeinsamer Faktoren der Faktorenanalyse ist ein Strukturgleichungsmodell. In diesem Modell verknüpfen die Regressionsgleichungen die Faktoren (UV) mit ihren manifesten Indikatoren (AV). Die Regressionsgewichte heißen in diesem Modell Faktorladungen.

Obwohl Pfadmodelle und Faktorenmodelle Strukturgleichungsmodelle sind, wird in der Literatur der Begriff des Strukturgleichungsmodells üblicherweise für eine Kombination aus beiden verwendet. Während die klassische Pfadanalyse Effekte und Zusammenhänge zwischen manifesten Variablen betrachtet, spezifizieren Strukturgleichungsmodelle Effekte und Zusammenhänge auf der Ebene latenter Variablen mit dem Ziel, sie messfehlerbereinigt zu schätzen. Strukturgleichungsmodelle setzen sich somit aus Messmodellen für ihre Konstrukte und einem Strukturmodell für die Zusammenhänge und Effekte zwischen den Konstrukten zusammen.

Mehrebenenanalysen. Wenn Daten eine hierarchische Struktur aufweisen, sind Mehrebenen- oder Multilevel-Analysen die Methode der Wahl (Eid et al., 2015). Eine hierarchische Struktur liegt vor, wenn Untersuchungsteilnehmer*innen (Ebene 1) verschiedenen Gruppen entstammen (Ebene 2, z.B. Kulturen, Schulklassen, Fanclubs, Parteien) und das zu

erklärende Verhalten nicht nur durch individuelle Merkmale, sondern auch durch Gruppenmerkmale beeinflusst wird. Beispiele hierfür haben wir im Abschnitt über hierarchischen Designs gegeben (siehe oben). Multilevel-Analysen werden auch verwendet, wenn ein Merkmal oder Verhalten häufig wiederholt gemessen wird, wie dies bei Experience Sampling-Designs der Fall ist (Bolger & Laurenceau, 2013). In diesem Fall bilden die Messzeitpunkte die unterste Ebene der Datenhierarchie.

Typenbildende Verfahren. Bei vielen friedenspsychologischen Fragestellungen sind die Gruppen, zwischen denen es zu Konflikten kommt, bekannt, z.B. die Mitglieder einer rechtsextremistischen Organisation und Migrant*innen. Solche offensichtlich unterscheidbaren Gruppen sind jedoch nicht zwangsläufig homogen. Vielmehr kann es innerhalb der Gruppen Untergruppen geben, die sich durch spezifische Merkmalsprofile voneinander unterscheiden. Solche Unterschiede können zur Erklärung des interessierenden Verhaltens relevant sein. Beispielsweise sind sich Mitglieder rechtsextremistischer Gruppierungen einig in ihrem Ziel, Immigration einzuschränken. Sie können sich aber erheblich in den Mitteln unterscheiden, die sie zur Erreichung dieses Ziels für effektiv und legitim halten. Zur Entdeckung von Gruppen, die sich in Merkmalsprofilen unterscheiden, werden Klassifikationsverfahren verwendet. Die drei am häufigsten verwendeten Verfahren sind die Clusteranalyse (Bacher, Pöge & Wenzig, 2010), die latente Klassenanalyse (Rost & Eid, 2009) und die latente Profilanalyse.

Qualitative Auswertungsverfahren

In der qualitativen Sozialforschung kommt eine Vielzahl verschiedener Auswertungsmethoden zum Einsatz, von denen fast alle auch für die friedenspsychologische Forschung geeignet sind. Die folgende Auswahl beschränkt sich auf einige der bekannteren Verfahren. Für eine umfangreichere Darstellung sei auf das Handbuch Qualitative Forschung in der Psychologie (Mey & Mruck, 2010a) verwiesen. Unabhängig vom gewählten Verfahren ist der Prozess des Kodierens in der qualitativen Forschung von besonderer Bedeutung. Eine sehr praktisch orientierte Übersicht über verschiedenste Kodierstrategien findet sich bei Saldaña (2016). Dem Abschnitt sei darüber hinaus der Hinweis vorangestellt, dass Forschungsprozesse in der qualitativen Forschung nicht immer linear, sondern teilweise auch zirkulär bzw. iterativ verlaufen. Ein Beispiel dafür ist die Grounded Theory-Methodologie (Glaser & Strauss, 1967), die zwar häufig als Auswertungsmethode aufgeführt wird, tatsächlich aber ein umfassendes Forschungsprogramm darstellt, welches Sampling, Erhebung und Auswertung miteinschließt.

(Qualitative) Inhaltsanalysen. Die qualitative Inhaltsanalyse (z.B. Schreier, 2012; Mayring, 2015) ist ein Verfahren, welches sowohl im qualitativen als auch im quantitativen Forschungsparadigma Anschluss findet. Historisch gesehen ist sie aus der quantitativen Inhaltsanalyse erwachsen, die in den Kommunikationswissenschaften zur Auswertung großer Textmengen verwendet wurde und immer noch wird. Allerdings steht bei der quantitativen Inhaltsanalyse das Auszählen von Häufigkeiten bestimmter Kategorien im Vordergrund, wo-

hingegen das Ziel der qualitativen Inhaltsanalyse die systematische Erfassung von Textbedeutungen ist. Auch sie ist ein datenreduzierendes Verfahren und auch hier können letztlich Häufigkeiten ausgezählt werden – die Erfassung von Bedeutung steht dabei aber weiterhin im Vordergrund. Kern der Inhaltsanalyse ist ein mit umfassenden Erläuterungen versehenes Kategoriensystem, welches im Kodierprozess systematisch zum Einsatz kommt. Solche Kategoriensysteme können theoriegeleitet erstellt und auf das Material angewandt werden (deduktives Vorgehen), gänzlich aus den vorliegenden Daten entwickelt werden (induktives Vorgehen) oder in einer Kombination aus Theorie und Empirie entwickelt werden (deduktiv-induktives Vorgehen). Die qualitative Inhaltsanalyse zeichnet sich drüber hinaus als ein Verfahren aus, in dem häufig mehrere Kodierer*innen zum Einsatz kommen, so dass die Zuordnung von Textstellen zu bestimmten Kategorien transparent und nachvollziehbar ist. In der Kombination aus Kategoriensystem und Interraterreliabilitäten bietet die Inhaltsanalyse einen Weg aus dem Dilemma der Subjektivität von Textinterpretationen an. Dabei wird durchaus anerkannt, dass Texte mehrere Bedeutungen haben und individuell interpretiert werden können, sie zeigt aber gleichzeitig den Weg hin zu einem überindividuell-intersubjektiven Verstehen auf.

Da die qualitative Inhaltsanalyse ein sehr populäres Verfahren ist, finden sich zahlreiche Beispiele in der friedenpsychologischen Forschung. Im Rahmen der Gezi-Park-Proteste in der Türkei analysierten Acar und Uluğ (2016) das Zusammenkommen ganz unterschiedlicher, teilweise auch im Konflikt miteinander stehendender Gruppen im Rahmen eines Protestes mit einem gemeinsamen Ziel. So entwickelten beispielsweise türkische und kurdische Protestierende neue Sichtweisen und mehr Verständnis füreinander. Die qualitative Inhaltsanalyse trug hier zu einem differenzierten Verständnis und einer umfassenden Beschreibung komplexer Intergruppenprozesse in einer zunächst recht unüberschaubaren Situation bei.

Diskursanalysen. Während bei der Inhaltsanalyse die Erfassung von Bedeutungen im Vordergrund steht, richtet sich der Fokus bei Diskursanalysen auch auf den *Prozess* der Bedeutungskonstruktion mittels Sprache. In Diskursanalysen wird also untersucht, wie sich Realität sprachlich konstruiert und konstituiert. Der Diskursbegriff wird sehr umstritten diskutiert, kann aber als „an interrelated set of texts, and the practices of their production, dissemination, and reception, that brings an object into being“ (Phillips & Hardy, 2002, S. 3) verstanden werden. Als Beispiel kann der Patriotismusdiskurs in Deutschland herangezogen werden. Dieser setzt sich zusammen aus (textuell manifestierten) Positionen unterschiedlichster Akteure, deren Produktionen sich als Netzwerk darstellen und – auch wenn sie sich teilweise widersprechen – aufeinander bezogen sind. Dabei ist dieser Diskurs speziell auf den hiesigen Kontext zugeschnitten und eng mit der deutschen Geschichte verwoben, was zu Bedeutungskonstruktionen führt (z.B. der Kritik am Zurschaustellen der Deutschlandfahne), die ohne die Berücksichtigung dieses Kontextes nicht zu verstehen wären.

Innerhalb der Verfahrensklasse der Diskursanalysen gibt es sehr unterschiedliche Traditionen, für die im Detail auf weitere Literatur verwiesen wird (z.B. Keller, Hierseland, Schneider & Viehöver, 2011). Beispielhaft seien die diskursive Psychologie, die eher in einer

linguistischen Tradition steht, sowie die kritische, an Foucault angelehnte Diskursanalyse genannt. In der Friedenspsychologie sind in den vergangenen Jahren Rufe nach einer stärkeren Beachtung diskursiver Perspektiven laut geworden (Gibson, 2011a, b). Gibson (2011b) illustriert ihren Mehrwert anhand einer Analyse von Konstruktionen von Krieg in britischen TV-Debatten im Vorfeld des Irak-Krieges von 2003. Hier wird herausgearbeitet, dass kriegsbefürwortende Positionen häufig mit einer explizit ablehnenden Haltung gegenüber kriegerischen Interventionen im Allgemeinen einhergehen, was als strategisches Manöver zur Selbstlegitimierung sowie Unterstreichung der Notwendigkeit einer Intervention in diesem speziellen Fall gewertet werden kann. Die diskursive Nähe von generell pazifistischen Positionen und Befürwortung eines konkreten militärischen Eingreifens ist zwar einleuchtend, findet sich in der traditionellen Einstellungsforschung zum Thema Krieg aber bisher nicht abgebildet.

Narrative Analysen. In der narrativ-orientierten Forschung stehen biographische Erzählungen und sprachlich umgesetzte Darstellungen von Zeitlichkeit im Vordergrund. Narrative Analysen – häufig auch als Erzählanalysen bezeichnet – sind zudem im Licht der sogenannten „narrativen Wende“ (Polkinghorne, 1988) in den Sozialwissenschaften zu betrachten, die dem Narrativen den Status eines eigenen Forschungsparadigmas zuweisen. Die unterschiedlichen narrativen Analyseverfahren eint der Versuch, herauszufinden, wie die Beforschten in der Darstellung biographischer Erfahrung Sinn und Identität konstruieren, Bezüge herstellen, ihnen Widerfahrenes interpretieren und in zeitlichen Zusammenhang setzen. Die Forschungspraxis narrativer Analysen reicht dabei von deduktiv geprägten, regelgeleiteten Ansätzen hin zu stark induktiv orientierten, der Ethnographie entlehnten Vorgehensweisen.

Ein Beispiel für eine in der narrativen Tradition stehende Analyse findet sich z.B. bei Hammack (2006) in der Betrachtung von Identitätskonstruktionen junger Palästinenser*innen und Israelis. Von besonderem Interesse in dieser Analyse war der Stellenwert der Teilnahme an sogenannten „Koexistenz“-Programmen für die Identitätsentwicklung. Hier zeigte sich bei den meisten Jugendlichen im Rahmen des Programms zwar eine intensive Auseinandersetzung mit der Gegenseite und auch ein gewisses Verständnis für die andere Position, mittelfristig erfolgte aber meist doch eine Orientierung an den dominanten Narrativen der Eigengruppe. Herauszuheben ist an dieser Arbeit, dass der narrative Ansatz um Elemente aus der Grounded Theory-Methodologie ergänzt wurde – eine Integration von methodischen Ansätzen, wie sie in der qualitativen Forschung häufiger anzutreffen ist.

Fazit

(Friedens-)Psychologisch Forschenden steht ein sehr umfangreiches methodisches Angebot zur Verfügung. In diesem Kapitel haben wir einen breiten (wenngleich notwendigerweise unvollständigen) Überblick geliefert, der als Ausgangspunkt für weitere, vertiefende Auseinandersetzungen mit unterschiedlichen Forschungsmethoden dienen kann. Das Forschen in konflikthaften Kontexten, mit vulnerablen Zielgruppen und zu sensiblen Themen sowie der Anspruch an eine hohe Praxistauglichkeit der gewonnenen Erkenntnisse macht die Friedenspsychologie zu einem besonders lohnenswerten, aber in methodischer Hinsicht auch herausfordernden Feld. Umso wichtiger ist, dass friedenspsychologisch Forschenden ein möglichst breites Repertoire an unterschiedlichen Methoden zur Verfügung steht, die sie in kreativer und kompetenter Weise umzusetzen wissen.

Literatur

Acar, Y. G. & Uluğ, Ö. M. (2016). Examining prejudice reduction through solidarity and togetherness experiences among Gezi Park activists in Turkey. *Journal of Social and Political Psychology*, *4*(1), 166-179. https://doi.org/10.5964/jspp.v4i1.547

Allport, G. W. (1935). Attitudes. In C. Murchison (Hrsg.), *A Handbook of Social Psychology* (S. 792–844). Worcester, MA: Clark University Press.

Amnesty International (2021). *Amnesty International Report 2020/21 zur weltweiten Lage der Menschenrechte.* Berlin: Amnesty International Deutschland e.V. Verfügbar unter: https://www.amnesty.de/informieren/amnesty-report/amnesty-report-2020

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*, 399–424. https://doi.org/10.1080%2F00273171.2011.568786

Bacher, J., Pöge, A. & Wenzig, K. (2010). *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren.* München: Oldenbourg. https://doi.org/10.1524/9783486710236

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531-1542. https://doi.org/10.1177/0956797615594620

Bashiriyeh, H. (2010). *Culture and violence: Psycho-cultural variables involved in homicide across nations.* Landau: Universität Koblenz-Landau (Dissertation).

Bauer, N. & Blasius, J. (Hrsg.) (2019). *Handbuch der empirischen Sozialforschung*. Berlin: Springer.

Baumert, A., Schlösser, T. & Schmitt, M. (2014). Economic games: Performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment, 30*, 178-192. https://doi.org/10.1027/1015-5759/a000183

Becker, G. S. (1976). *The economic approach to human behavior*. Chicago, IL: University of Chicago Press.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407-425. http://dx.doi.org/10.1037/a0021524

Billmann-Mahecha, E. (n.d.). Forschungsparadigmen. In C. Cohrs, N. Knab & G. Sommer (Hrsg.), *Handbuch Friedenspsychologie*. Verfügbar unter: https://handbuch-friedenspsychologie.de/

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G. & Tetlock, P. E. (2009). Strong claims and weak evidence: reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, *94*(3), 567-582. https://doi.org/10.1037/a0014665

Boehnke, K., Lietz, P., Schreier, M. & Wilhem, A. (2011). Sampling. The selection of cases for culturally comparative psychological research. In D. Matsumoto & F. J. R. van de Vijver (Hrsg.), *Cross-cultural methods in psychology* (S. 101-129). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511779381.007

Bohner, G. & Wänke, M. (2002). *Attitudes and attitude change*. Hove, UK: Psychology Press.

Bolger, N. & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research.* New York, NY: Guilford Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193-217. https://doi.org/10.1037/h0047470

Camerer, C. F. (2003). *Behavioral game theory*. New York, NY: Russell Sage Foundation.

Christ, O., Schmid, K., Lolliot, S., Swart, H., Stolle, D., Tausch, N., ... & Hewstone, M. (2014). Contextual effect of positive intergroup contact on outgroup prejudice. *Proceedings of the National Academy of Sciences, 111*(11), 3996-4000. http://dx.doi.org/10.1073/pnas.1320901111

Clayton, K., Horrillo, J. & Sniderman, P. M. (2021). The BIAT and the AMP as measures of racial prejudice in political science: A methodological assessment. *SSRN Electronic Journal.* http://dx.doi.org/10.2139/ssrn.3744338.

Berufsverband Deutscher Psychologinnen und Psychologen e.V. und Deutsche Gesellschaft für Psychologie e.V. (2016). Berufsethische Richtlinien. Verfügbar unter: https://www.dgps.de/fileadmin/user_upload/PDF/berufsethik-foederation-2016.pdf

Doliński, D., Grzyb, T., Folwarczny, M., Grzybała, P., Krzyszycha, K., Martynowska, K. & Trojanowski, J. (2017). Would You Deliver an Electric Shock in 2015? Obedience in the Experimental Paradigm Developed by Stanley Milgram in the 50 Years Following the Original Studies. *Social Psychological and Personality Science*, *8*, 927–933. https://doi.org/10.1177/1948550617693060

Dresing, T. & Pehl, T. (2010). Transkription. In G. Mey & G. Mruck (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (S. 723-733). Wiesbaden: VS Verlag.

Duckitt, J. & Sibley, C. G. (2009). A dual-process motivational model of ideology, politics, and prejudice. Psychological Inquiry, 20, 98–109. http://dx.doi.org/10.1080/10478400903028540

Eid, M. & Diener, E. (Hrsg.) (2006). *Handbook of multimethod measurement in psychology*. New York, NY: American Psychological Association. https://psycnet.apa.org/doi/10.1037/11383-000

Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden*. Weinheim: Beltz.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Flick, U. (2010). Gütekriterien qualitativer Forschung. In G. Mey & K. Mruck (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (S. 395-407). Wiesbaden: VS Verlag.

Flick, U. (2011). *Triangulation. Eine Einführung* (3. Aufl.). Wiesbaden: VS Verlag.

Gibson, S. (2011a). Social psychology, war and peace: Towards a critical discursive peace psychology. *Social and Personality Psychology Compass, 5*, 239–250. https://doi.org/10.1111/j.1751-9004.2011.00348.x

Gibson, S. (2011b). 'I'm not a war monger but…': Discourse analysis and social psychological peace research. *Journal of Community and Applied Social Psychology*, *22*(2), 159-173. https://doi.org/10.1002/casp.1099

Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory. Strategies for qualitative research*. Chicago, IL: Aldine. https://doi.org/10.4324/9780203793206

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y. & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 U.S. presidential election. *Analyses of Social Issues and Public Policy, 9*(1), 241–253. https://doi.org/10.1111/j.1530-2415.2009.01195.x

Grimm, K., Jacobucci, R. & McArdle, J. J. (2017). Big data methods and psychological science. *Psychological Science Agenda*. Verfügbar unter : https://www.apa.org/science/about/psa/2017/01/big-data-methods

Häcker, H.O. (2017). Objektiver Test. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (S. 1197-1198). Göttingen: Hogrefe.

Halperin, E., Porat, R., Tamir, M. & Gross, J. J. (2013). Can emotion regulation change political attitudes in intractable conflicts? From the laboratory to the field. *Psychological Science*, *24*(1), 106-111. https://doi.org/10.1177/0956797612452572

Hammack, P. L. (2006). Identity, conflict, and coexistence: Life stories of Israeli and Palestinian adolescents. *Journal of Adolescent Research*, *21*(4), 323-369. https://doi.org/10.1177/0743558406289745

Herre, B., Ortiz-Ospina, E. & Roser, M. (2013). *Democracy*. Published online at OurWorldInData.org. Verfügbar unter: https://ourworldindata.org/democracy

Hofstede, G. (2001), *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks, CA: Sage.

Kearney, M. W. (2017). Cross-lagged panel analysis. In M. R. Allen (Hrsg.), *The SAGE Encyclopedia of Communication Research Methods* (S. 312–314). Thousand Oaks, CA: Sage.

Kelle, U. (2007). *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte.* Wiesbaden: VS Verlag.

Keller, R., Hirseland, A., Schneider, W. & Viehöver, W. (Hrsg.) (2011). *Handbuch sozialwissenschaftliche Diskursanalyse: Theorien und Methoden* (Bd.1, 3. Aufl.). Wiesbaden: VS Verlag.

Köster, S. (2009). *Männer als Opfer und Täter*. Frankfurt: Verlag für Polizeiwissenschaft.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mauss, I. B. & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

Maxwell, J. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, *62*(3), 279-301.

Mayring, P. (2015). *Qualitative Inhaltsanalyse*. Weinheim: Beltz.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Hrsg.), *Prejudice, discrimination, and racism* (S. 91–125). Cambridge, MA: Academic Press.

Merton, R., Fiske, M. & Kendall, P. (1956). *The focused interview. A manual of problems and procedures*. Glencoe, IL: Free Press.

Meuser, M. & Nagel, U. (1991). Experteninterviews – vielfach erprobt, wenig bedacht. Ein Beitrag zur qualitativen Methodendiskussion. In D. Garz & K. Kraimer (Hrsg.), *Qualitativ-empirische Sozialforschung. Konzepte, Methoden, Analysen* (S. 441-471). Opladen: Westdeutscher Verlag.

Mey, G. & Mruck, K. (Hrsg.). (2010a). *Handbuch qualitative Forschung in der Psychologie*. Wiesbaden: VS Verlag.

Mey, G. & Mruck, K. (2010b). Interviews. In G. Mey & Mruck, K. (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (S. 423-433). Wiesbaden: VS Verlag.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2. Aufl.). Thousand Oaks, CA: Sage.

Milgram, S. (1974). *Obedience to Authority. An Experimental View.* New York: Harper.

Morse, J. M. (2007). Sampling in grounded theory. In A. Bryant & K. Charmaz (Hrsg.), *The Sage handbook of grounded theory* (S. 229-244). London: Sage.

Mummendey, H. D. (2014). *Die Fragebogen-Methode. Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung*. Göttingen: Hogrefe.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S. et al. (2015). Scientific standards. Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716.

Our world in data (n.d.). Verfügbar unter: https://ourworldindata.org/

Paluck, E. L. (2009). Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of Personality and Social Psychology*, *96*(3), 574-587. https://doi.org/10.1037/a0011989

Phillips, N. & Hardy, C. (2002). *Discourse analysis: Investigating processes of social construction*. Thousand Oaks, CA: Sage.

Polkinghorne, D. E. (1988). *Narrative knowing and the human sciences*. Albany, NY: Suny Press.

Popper, K. R. (2005). *Logik der Forschung*. Tübingen: Mohr Siebeck.

Project Implicit (n.d.). Verfügbar unter: https://implicit.harvard.edu/implicit/takeatest.html

Przyborski, A. & Riegler, J. (2010). Gruppendiskussion und Fokusgruppe. In G. Mey & K. Mruck (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (S. 436-448). Wiesbaden: VS Verlag.

Rencher, A. C. & Schaalje, G. B. (2008). *Linear models in statistics*. New York, NY: John Wiley.

Ritchie, S. J., Wiseman, R. & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PloS one*, *7*(3), e33423. https://doi.org/10.1371/journal.pone.0033423

Rost, J. & Eid, M. (2009). Mischverteilungsmodelle. In H. Holling (Hrsg.), *Grundlagen und statistische Methoden der Evaluationsforschung (Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie IV: Evaluation, Bd. 1,* S. 483–524). Göttingen: Hogrefe.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.

Saldaña, J. (2016). *The coding manual for qualitative researchers* (3. Aufl.). London: Sage.

Schmidt-Atzert, L. & Amelang, M. (2021). *Psychologische Diagnostik*. Berlin: Springer.

Schmitt, M. & Gerstenberg, F. X. R. (2014). *Psychologische Diagnostik kompakt*. Weinheim: Beltz.

Schreier, M. (2012). *Qualitative content analysis in practice*. London: Sage.

Schuerger, J. M. (2008). The objective-analytic test battery. In G. J. Boyle, G. Matthews & D. H. Saklofske (Hrsg.), *The SAGE handbook of personality theory and assessment, Vol. 2. Personality measurement and testing* (S. 529–546). SAGE Publications. https://doi.org/10.4135/9781849200479.n25.

Schütze, F. (1983). Biographieforschung und narratives Interview. *Neue Praxis, 13*(3), 283-293.

Schwartz, S. H. & Bardi, A. (2001). Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross Cultural Psychology, 32*, 268-290. https://doi.org/10.1177/0022022101032003002

Schwartz, S. H. (2004). Mapping and interpreting cultural differences around the world. In H. Vinken, J. Soeters & P. Ester (Hrsg.), *Comparing cultures, dimensions of culture in a comparative perspective* (S. 43-73). Leiden: Brill.

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

Spradley, J. P. (1979). *The ethnographic interview*. New York, NY: Holt, Rinehart & Winston.

Steffens, M. (2012). *Die Affaire Stapel*. Verfügbar unter : https://www.spektrum.de/magazin/die-affaere-stapel/1140956

Steinke, I. (1999). *Kriterien qualitativer Forschung. Ansätze zur Bewertung qualitativ-empirischer Sozialforschung.* Weinheim: Juventa.

Steyer, R. & Eid, M. (2001). *Messen und Testen*. Berlin: Springer.

Strauss, A. & Corbin, J. (1990). *Basics of qualitative research. Grounded theory procedures and techniques.* Newbury Park, CA: Sage.

Strobl, C. (2012). *Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis*. München: Rainer Hampp Verlag.

Transparency International (2022). *Korruptionswahrnehmungsindex 2022*. Berlin: Transparency International Deutschland e.V. Verfügbar unter: https://www.transparency.de/publikationen/detail/article/korruptionswahrnehmungsindex-2022

Tschirk, W. (2014). *Statistik: Klassisch oder Bayes*. Berlin: Springer.

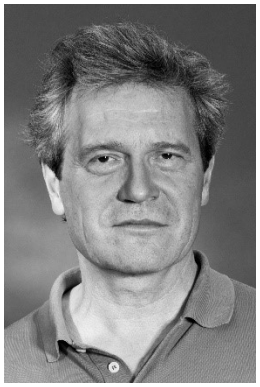
Van Assche, J., Roets, A., De Keersmaecker, J. & Van Hiel, A. (2017). The mobilizing effect of right-wing ideological climates: Cross-level interaction effects on different types of outgroup attitudes. *Political Psychology, 38*, 757-776. https://doi.org/10.1111/pops.12359

van de Vijver, F. J. R. & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29-37. http://dx.doi.org/10.1027/1015-5759.13.1.29

Witzel, A. (2000). Das problemzentrierte Interview. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(1), Art. 22. https://doi.org/10.17169/fqs-1.1.1132

Woo, S. R., Tay, L. & Proctor, R. W. (Hrsg.) (2020). *Big data in psychological research.* Washington, DC: American Psychologial Association.

Kea Sarah Brahms arbeitet als Beraterin in einem Programm zur Radikalisierungsprävention mit Schwerpunkt Islamismus. Sie ist außerdem als Lehrbeauftragte am Zentrum für Konfliktforschung der Philipps-Universität Marburg tätig. Nach dem Psychologiestudium in Osnabrück arbeitete sie in Forschungsprojekten in Bremen und Tel Aviv sowie als Lektorin im VS Verlag für Sozialwissenschaften. Sie interessiert sich besonders für (kulturübergreifende) Gerechtigkeitsüberzeugungen, Emotionen in Intergruppenprozessen, sowie die Frage, warum sich Menschen radikalen Strömungen zuwenden.



Manfred Schmitt ist Professor i.R. für Diagnostik und Persönlichkeitspsychologie an der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau. Zuvor hatte er Professuren für Entwicklungspsychologie (Saarbrücken), Methodenlehre (Magdeburg) und Sozialpsychologie (Trier) inne. Seine Forschungsinteressen umfassen Emotionen (Eifersucht, Schuldgefühle, Ärger, Angst, Ekel, Depressivität), Gerechtigkeitsüberzeugungen und Ungerechtigkeitssensibilität, das Zusammenspiel expliziter und impliziter Persönlichkeitseigenschaften bei der Informationsverarbeitung und Verhaltenssteuerung, Wechselwirkungen zwischen Persönlichkeitseigenschaften und Situationsmerkmalen, objektive Persönlichkeitstests, psychologische Beiträge zum Umweltschutz und Einflüsse der Persönlichkeit von Lehrkräften auf die Unterrichtsqualität.