3. **Auswahl der untersuchungsrelevanten Aspekte**: Da die dimensionale Analyse als Hilfsmittel der Konzeptspezifikation bzw. der Gegenstandsstrukturierung eine empirische Studie vorbereiten soll, müssen schließlich auch theoretische und forschungspraktische Erwägungen herangezogen werden: Welche Aspekte lassen sich mit den vorhandenen zeitlichen, finanziellen und personellen Mitteln überhaupt untersuchen? Welche Aspekte sind besonders interessant und vielversprechend hinsichtlich ihres theoretischen oder praktischen Erkenntniswertes? So mag man sich z. B. dafür entscheiden, den Aspekt der Videotelefonie per Handy nicht in die Studie einzubeziehen, weil zum Untersuchungszeitpunkt wenige Kinder diese Option nutzen. Oder man wird insbesondere jene Aspekte der Handykompetenz umfassend untersuchen, deren praktische Förderung – z. B. im Rahmen des Schulunterrichts – als besonders gut möglich erscheint.
4. **Entwicklung eines deskriptiven Begriffsschemas**: Nachdem die in die empirische Untersuchung einzubeziehenden Subdimensionen des Konzepts identifiziert wurden, müssen für diese jeweils schlüssige Nominaldefinitionen formuliert werden. Soll z. B. im Bereich Handywissen der Aspekt „Kenntnis über die Kosten der Mobilkommunikation" erfasst werden, so muss genau definiert werden, ob damit z. B. gemeint ist, dass Kinder wissen, wie viel ihr Handy als Endgerät kostet, wie teuer Inlands- und Auslandsgespräche mit dem Handy sind oder welche verschiedenen Bezahlmodelle existieren und welchen Tarif sie selbst nutzen.

Neben explorativen empirischen Studien sind es vor allem **Theorie- und Methodenstudien** (► Abschn. 7.3), die komplexe Konzepte einer systematischen dimensionalen Analyse unterziehen, um deren Messung vorzubereiten (► Studienbeispiel „Konzeptspezifikation von ‚Globalisierung' mittels dimensionaler Analyse").

8.3 Grundlagen zur Durchführung einer Operationalisierung

Auf die Konzeptspezifikation mittels Nominaldefinition, die mithilfe von Bedeutungsanalyse sowie dimensionaler Analyse zu erarbeiten ist, folgt die Operationalisierung, die in eine operationale Definition mündet. Im Folgenden wird die operationale Definition eingeführt. Dabei werden Besonderheiten bei abhängigen und unabhängigen Variablen hervorgehoben und verbreitete Fehlinterpretationen problematisiert.

8.3.1 Operationale Definition von theoretischen Konzepten

Für ein und dasselbe theoretische Konzept existieren meistens verschiedene Möglichkeiten der Operationalisierung. So kann das **theoretische Konstrukt „Intelligenz"** mit unterschiedlichen **Messinstrumenten** (Intelligenztests) operationalisiert werden, die sich jeweils aus verschiedenen **Indikatoren** (Intelligenztestaufgaben) zusammensetzen, welche einzeln und gemeinsam dazu dienen, die Intelligenzleistung einer Person zu erfassen. Eine **operationale Definition** wie „Intelligenz ist, was der Intelligenztest misst" bzw. genauer: „Intelligenz ist, was der Hamburg-Wechsler-Intelligenztest misst" oder „Intelligenz ist, was der Raven-Test misst" erscheint für sich genommen meistens unbefriedigend. Deswegen sollten operationale Definitionen nicht die Konzeptspezifikation ersetzen. Hat man zuerst im Rahmen der Konzeptspezifikation theoretisch festgelegt, welches Intelligenzkonzept man zugrunde legen möchte und die entsprechende Nominaldefinition angegeben, dann wird man auf dieser Basis gezielt die passende Operationalisierung bzw. den passenden Intelligenztest auswählen. Bei dieser Vorgehensweise ist die verwendete operationale Definition nicht zirkulär, sondern **in einem theoretischen Verständnis von Intelligenz verankert**.

Operationalisierung – Die Operationalisierung („operationalization") eines theoretisches Konzepts bzw. einer latenten Variable legt fest, anhand welcher beobachtbaren Variablen **(Indikatoren)** die Ausprägung des theoretischen Konzepts bei den Untersuchungsobjekten festgestellt werden soll. Neben der Auswahl der Indikatoren gehört zur Operationalisierung auch die **Festlegung der Messinstrumente**, mittels derer den Ausprägungen der einzelnen Indikatoren jeweils entsprechende numerische Werte zugeordnet und zu einem Gesamtmesswert für das Konstrukt verrechnet werden. Komplexe theoretische Konstrukte werden selten mit einem einzigen Indikator (**Einzelindikator** als Messinstrument) operationalisiert, sondern meist über einen Satz von Indikatoren (d. h. über eine **psychometrische Skala** oder einen **Index**). Mit der Festlegung der Operationalisierung wird für ein theoretisches Konzept (dargelegt über seine Nominaldefinition) eine konkretisierende **operationale Definition („operational definition")** vorgenommen.

Angenommen, die **Intensität des Mobbings von Schulkindern** wurde im Zuge der Konzeptspezifikation über die Dimensionen a) Anzahl der Aggressoren, b) Dauer des Mobbings und c) Schwere der aggressiven Handlungen definiert. Indikatoren für diese theoretischen Aspekte können nun entsprechend detaillierte Fragen in einem Elternfragebogen oder in einem Schülerinterview sein. Welche Operationalisierungsvariante und damit auch **Datenerhebungsmethode** (► Kap. 10) gewählt wird, hängt u. a. von forschungspraktischen sowie von theoretischen Erwägungen ab. So mag es zwar weniger aufwändig sein,

Eltern kollektiv beim Elternabend einen standardisierten Fragebogen zur Mobbingbetroffenheit ihres Kindes ausfüllen zu lassen als alle Kinder einzeln in kindgerechter Weise zu interviewen. Da jedoch nicht davon auszugehen ist, dass Eltern über alle Mobbingvorfälle ihrer Kinder genau informiert sind, wären vermutlich Indikatorvariablen, die direkt an den Kindern selbst erhoben werden, zu bevorzugen.

Die zur Konzeptspezifikation der **Globalisierung unterschiedlicher Länder** angegebenen theoretischen Dimensionen (z. B. „Touristenströme"; ► Studienbeispiel „Konzeptspezifikation von ‚Globalisierung' mittels dimensionaler Analyse") sind im Zuge der Operationalisierung in konkret messbare Indikatoren zu übersetzen (z. B. Anzahl der Ankünfte internationaler Touristen pro 1 000 Einwohner pro Jahr), die z. B. den amtlichen Statistiken zu entnehmen wären. Zuweilen muss aus **forschungspraktischen bzw. forschungsökonomischen Gründen** auf Indikatoren zurückgegriffen werden, die das theoretische Konzept nur teilweise oder nur ungenau abbilden (etwa weil relevante Kennwerte nicht schnell oder kostengünstig genug beschaffbar sind). Abweichungen zwischen theoretischem Konstrukt und den zur Operationalisierung genutzten Indikatoren sind zu begründen und bei der Ergebnisinterpretation zu berücksichtigen (z. B. wenn die interessierenden internationalen „Touristenströme" über die Zahl der Hotelübernachtungen operationalisiert würden, wobei dann auch nationale Touristen sowie Geschäftsreisende enthalten wären).

Multiple Indikatoren tragen dazu bei, dass die verschiedenen Aspekte eines komplexen theoretischen Konstruktes möglichst vollständig abgebildet werden und somit wirklich das erfasst wird, was gemessen werden soll (Kriterium der Gültigkeit bzw. **Validität** des Messinstrumentes). Zudem können durch den Einsatz multipler Indikatoren auch Messfehler reduziert werden, so dass die Messgenauigkeit bzw. **Reliabilität** des Instruments steigt (zu psychometrischen Gütekriterien im Überblick ► Abschn. 10.4.1). Dementsprechend wird eine groß angelegte bevölkerungsrepräsentative Studie zur Verbreitung von Depressionen mit einem etablierten Depressionsmessinstrument arbeiten, welches das Konstrukt möglichst differenziert mithilfe einer Reihe sorgfältig entwickelter und aufeinander abgestimmter Selbstauskunftsfragen bzw. Indikatoren erfasst. Demgegenüber wird eine Studie, die sich dem Essverhalten widmet, das Konstrukt Depression – wenn es denn am Rande auch erhoben werden soll – möglicherweise nur mit einem **Einzelindikator** (z. B. mit einer einzigen Frage im Fragebogen) erheben, weil alles andere das ohnehin umfangreiche Erhebungsinstrument sonst zu sehr aufblähen, die Untersuchungspersonen demotivieren und somit letztlich den Erfolg der gesamten Studie einschränken könnte (zur Messung von latenten Merkmalen wie z. B. Depression mittels Einzelindikatoren ► Abschn. 8.5.2).

Wenn anstelle eines Einzelindikators ein aus mehreren Indikatoren bestehendes Messinstrument zur Operationalisierung eines theoretischen Konstruktes genutzt wird, so stellt sich die Frage, welche Indikatoren auf welche Weise zu einem Messinstrument zusammengefasst werden sollen. Die Auswahl, Modifikation oder Neuentwicklung von Messinstrumenten und Indikatoren ist die Kernaufgabe bei der Operationalisierung. Wissenschaftlich unbrauchbar sind **Ad-hoc-Instrumente**, die aus einer mehr oder minder willkürlichen und ungeprüften Menge an Indikatoren bestehen. Gefordert sind stattdessen wissenschaftliche Messinstrumente, deren Aufbau theoretisch begründet und deren Gütekriterien empirisch geprüft sind, so dass man sichergehen kann, dass sie tatsächlich genau das Zielkonstrukt erfassen (Validität) und wenig durch Messfehler verzerrt sind (Reliabilität). Die Forschung rund um die Voraussetzungen und Gütekriterien der Messung psychologischer Merkmale wird auch als **Psychometrie** („psychometrics") bezeichnet. Eine wichtige Grundlage für die Konstruktion und Bewertung von Messinstrumenten ist die **Testtheorie** (► Abschn. 10.4.4). In Abgrenzung von Ad-hoc-Instrumenten werden überprüfte Messinstrumente auch als **psychometrische Skalen** und **psychometrische Tests** bezeichnet. Typische Gegenstände psychometrischer Messung sind Einstellungen, aber auch Wissen und kognitive Leistungen, Persönlichkeitsmerkmale oder psychologische Störungen, zu denen eine Fülle von Messinstrumenten vorliegen (für eine Auswahl gebräuchlicher Testverfahren ► Abschn. 10.4.3). Auf die Konstruktionsprinzipien von unterschiedlichen psychometrischen Skalen (► Abschn. 8.6) sowie von Indizes (► Abschn. 8.7) wird im Verlauf dieses Kapitels noch genauer eingegangen.

Wichtig ist es im Hinblick auf das **Verhältnis zwischen theoretischem Konstrukt und Indikator**, zwischen zwei Typen von Indikatoren zu unterscheiden: Den reflektiven und den formativen Indikatoren (auch Bühner 2011, S. 37).

- Beim **reflektiven Messmodell**, das psychometrischen Skalen (► Abschn. 8.6) zugrunde liegt, wird das **theoretische Konstrukt als Ursache** und die **Indikatoren** werden **als Wirkungen** betrachtet. Beispiel: Dadurch, dass Menschen sich im Grad ihrer Schüchternheit unterscheiden (theoretisches Konstrukt als Ursache), beantworten sie Indikatorvariablen bzw. Skalen-Items wie „Mir fällt es schwer, mit Fremden ins Gespräch zu kommen" oder „Wenn ich mit Fremden spreche, fühle ich mich gehemmt" in systematischer Weise unterschiedlich (reflektive Indikatoren als Wirkungen).

In den Ausprägungen dieser Indikatoren „reflektiert" sich also die Ausprägung des latenten Merkmals: Weil eine Person schüchtern ist, stimmt sie entsprechenden Aussagen über schüchternes Verhalten und Erleben tendenziell zu. Dasselbe Ursache-Wirkungs-Prinzip zwischen Konstrukt und Indikatoren gilt für Skalen, mit denen z. B. durch mehrere Skalen-Items (Fragen oder Aufgaben) Konstrukte wie Depression, Intelligenz, Aggressivität oder Religiosität erfasst werden. Die in einer psychometrischen Skala enthaltenen reflektiven Indikatoren sind einander **formal und inhaltlich ähnlich** (z. B. bedeutungsähnliche Selbsteinschätzungen zur Schüchternheit) und relativ hoch miteinander korreliert.

- Demgegenüber betrachtet das **formative Messmodell**, welches den Indizes (► Abschn. 8.7) zugrunde liegt, die **Indikatoren als Ursachen** bzw. Determinanten des im Zuge der Konzeptspezifikation definierten theoretischen Konstruktes. Die Ausprägung des **Konstruktes ist eine Wirkung** der Indikatoren. Beispiel: Erst wenn nachweisbar ist, dass ein bestimmtes Land über zunehmende grenzüberschreitende Touristen-, Daten-, Handels-, Finanzströme etc. verfügt, dann wird diesem Land gemäß der Nominaldefinition und Konzeptspezifikation von Globalisierung ein hoher Globalisierungsgrad zugeschrieben. Also nicht weil ein Land globalisiert ist, entstehen dadurch Touristen- und Finanzströme, Wirtschafts- und Wissenschaftsaustausch etc., sondern das Auftreten der inhaltlich ganz verschiedenen grenzüberschreitenden sozialen Interaktionen erzeugt bzw. „formt" den Globalisierungsgrad eines Landes. Die in einem Index enthaltenen formativen Indikatoren können einander **formal und inhaltlich sehr unähnlich** sein und müssen auch nicht miteinander korrelieren. Ein weiteres Beispiel: Wenn eine Person einen geringen Bildungsstand hat, einen Beruf mit geringem Prestige ausübt und über ein unterdurchschnittliches Einkommen verfügt (drei formative Indikatoren als Determinanten), dann wird ihr definitionsgemäß ein niedriger sozioökonomischer Status zugeschrieben (theoretisches Konstrukt als Wirkung).

Neben Einzelindikatoren werden zur Operationalisierung theoretischer Konstrukte oft **Messinstrumente** verwendet, die aus mehreren Indikatoren bestehen und auf zwei grundlegend verschiedenen Messmodellen basieren.

Reflektives Messmodell – Bei einem reflektiven Messmodell geht man davon aus, dass das zu messende Konstrukt die Ursache für die Merkmalsausprägungen auf den gewählten Indikatoren ist. Als Messinstrument wird eine **psychometrische Skala** genutzt, die aus homogenen bzw. inhaltsähnlichen Fragen, Aussagen oder Aufgaben besteht (sog. **reflektive Indikatoren**, in denen sich das Konstrukt widerspiegelt).

Formatives Messmodell – Bei einem formativen Messmodell geht man davon aus, dass das zu messende Konstrukt die Wirkung oder Folge der Merkmalsausprägungen der Indikatoren ist. Als Messinstrument wird ein **Index** gebildet, in den heterogene Kennwerte eingehen (sog. **formative Indikatoren**, durch die das Konstrukt ursächlich gebildet wird).

Bereits bei der Konzeptspezifikation (► Abschn. 8.2), spätestens im Zuge der operationalen Definition, also bei der Auswahl oder Konstruktion von Indikatoren und Messinstrumenten, sollte man sich darüber klar geworden sein, ob man ein reflektives oder ein formatives Messmodell anzulegen hat, also ob eine psychometrische Skala oder ein Index zu verwenden ist.

8.3.2 Operationalisierung von abhängigen Variablen

In der quantitativen Sozialforschung messen wir Variablen vor allem, um im Kontext explanativer Studien Hypothesen über Variablenzusammenhänge, Gruppenunterschiede oder Veränderungen über die Zeit zu prüfen. Besonders aussagekräftig ist dabei die Prüfung von Kausalhypothesen, wie sie in experimentellen und quasi-experimentellen Designs angestrebt wird (► Abschn. 7.6). In diesen Designs wird mindestens eine unabhängige Variable systematisch variiert (z. B. unterschiedliche Behandlungen, Interventionen, Stimuli), um die Auswirkung auf die abhängige(n) Variable(n) zu prüfen. Dabei versucht man meistens, die Ausprägungen der abhängigen Variablen in **möglichst feinen Abstufungen** zu erfassen. Aber auch in nicht-experimentellen Studien ist eine differenzierte Messung der Variablenausprägungen nützlich. Zur Messung feiner Merkmalsabstufungen sind folgende sechs **Operationalisierungsvarianten** (modifiziert nach Conrad & Maul, 1981, S. 151) besonders geeignet:

1. **Häufigkeit:** Wie oft tritt ein bestimmtes Verhalten auf? (Beispiele: Anzahl der Fehler in einem Diktat, Häufigkeit der Blickkontakte beim Flirt, Häufigkeit von Sprechpausen in einer Vernehmungssituation, Häufigkeit von Ehestreits vor und nach einer Paartherapie)
2. **Reaktionszeit:** Wie viel Zeit vergeht, bis eine Person nach Auftreten eines Stimulus reagiert? (Beispiele: Reaktionslatenz nach Auftreten eines unerwarteten Verkehrshindernisses, Reaktionszeit bis zur Identifikation eines Wortes). Die Reaktionszeitmessung ist in der Regel experimentellen Laborstudien vorbehalten.
3. **Reaktionsdauer:** Wie lange reagiert eine Person auf einen Stimulus bzw. auf eine Intervention? (Beispiele: Lösungszeit für eine Mathematikaufgabe, Verweildauer des Auges auf einem bestimmten Bildausschnitt, Dauer des Nichtrauchens nach einem Anti-Rauch-Training)

4. **Reaktionsstärke:** Wie intensiv reagiert eine Person auf einen Stimulus bzw. auf eine Intervention? (Beispiele: Stärke der Muskelanspannung als Indikator für Aggressivität, geäußerte Stärke von Meinungen auf Ratingskalen, Höhe des Blutdruckanstiegs, Intensität der Zustimmung zu einer Partei vor und nach der Rezeption von Pressebeiträgen über einen Parteiskandal). Die Reaktionsstärke kann im Rahmen von Messwiederholungsdesigns bzw. Längsschnittstudien auch mehrfach erhoben und in ihrem Verlauf betrachtet werden (► Abschn. 7.8).
5. **Reaktionsqualität**: Welche Wertigkeit (Valenz) hat eine Reaktion auf einen Stimulus: Ist sie eher positiv oder negativ bzw. beinhaltet sie Zuwendung oder Abwendung? (Beispiele: Bewertung von Lebensmitteln hinsichtlich Geruch und Geschmack auf Schulnotenskalen; Einstufung der Sympathie oder Antipathie gegenüber einer Person auf einer Ratingskala; Einstufung der eigenen Stimmungslagen – angespannt, aufmerksam, fröhlich etc. – auf einer psychometrischen Skala.)
6. **Wahlreaktion:** Welche Wahl trifft eine Person angesichts mehrerer Wahlmöglichkeiten? (Beispiele: Bevorzugung eines von zwei Kunstwerken als ästhetischer im Paarvergleichsurteil; Bevorzugung eines von mehreren Reisezielen bei einer Mehrfachwahlaufgabe; Nennung der Lieblingsmarke aus einem Spektrum an Marken bei einem Präferenzurteil.)

Bei den Operationalisierungsvarianten für abhängige Variablen ist der **Zeitpunkt der Messung** zu beachten: Die Messung kann **nach Abschluss der Intervention** (z. B. Stimuluspräsentation) erfolgen, etwa wenn nach dem Betrachten eines Films, nach einem Bewerbungsinterview oder nach einer Unterrichtsstunde jeweils die Ausprägungen der interessierenden abhängigen Variablen erfasst werden (z. B. mit einer psychometrischen Skala, mit einem psychologischen Testverfahren oder per Expertenurteil). Es besteht aber auch die Möglichkeit, die Messung **prozessbegleitend** durchzuführen (d. h. **während** die Untersuchungsperson auf den Stimulus bzw. die unabhängige Variable reagiert wird z. B. die Reaktionsdauer per Stoppuhr gemessen und die Reaktionsqualität durch Beobachtung von Mimik und Gestik erfasst). Besonders gut geeignet für prozessbegleitende Messungen sind **physiologische Messverfahren** (► Abschn. 10.5), mit denen kontinuierlich während des gesamten Versuchsdurchlaufs z. B. der Blutdruck oder die Blickbewegungen aufgezeichnet werden. Es besteht auch die Möglichkeit, subjektive Bewertungen prozessbegleitend zu erheben. Hierfür werden als Operationalisierungsvarianten das sog. **Real Time Response (RTR) Measurement** bzw. das **Continous Response Measurement (CRM)** genutzt (zum Überblick Maier, Maier, Maurer, Reinemann, & Meyer, 2009). Bei diesen Verfahren erhält die Untersuchungsperson als Messinstrument einen **Dreh- oder Schieberegler**, über den sie fortlaufend stufenlos angeben kann, ob und wie stark sie einen präsentierten Reiz gerade positiv oder negativ bewertet (auch andere Reaktionsqualitäten wie interessant vs. langweilig etc. können erfasst werden). Die RTR-Measurement- bzw. CRM-Methode wurde bereits in den 1930er-Jahren entwickelt und wird bis heute u. a. zur prozessbegleitenden Messung von Publikumsreaktionen auf Medienangebote wie Radio- oder TV-Sendungen, Werbespots und Kinofilme genutzt (Maier, Maurer, Reinemann, & Faas, 2006; Reinemann, Maier, Faas, & Maurer, 2005; Schmeisser, Bente, & Isenbart, 2004; Schneider et al., 2011).

Je komplexer ein theoretisches Konstrukt und je wichtiger seine genaue Messung für die Studie ist, umso eher wird man zu seiner Operationalisierung statt auf einen **Einzelindikator** (► Abschn. 8.5) auf ein **Messinstrument mit mehreren Indikatoren** zurückgreifen. Sehr verbreitet zur Operationalisierung abhängiger Variablen in der empirischen Sozialforschung sind die **psychometrische Skala** (► Abschn. 8.6) sowie der **Index** (► Abschn. 8.7). Sie messen Qualitäten sowie mehr oder minder fein abgestuft die Intensitäten von Merkmalsausprägungen.

8.3.3 Operationalisierung von unabhängigen Variablen

Während abhängige Variablen oft stetige Merkmale sind, die mit möglichst vielen Abstufungen gemessen werden, handelt es sich bei unabhängigen Variablen meistens um diskrete Variablen mit wenigen Ausprägungen. Die unabhängige Variable bzw. der Ursachenfaktor fungiert in experimentellen und quasi-experimentellen Studien (► Abschn. 7.6) als **Gruppierungsvariable**, wobei der Vergleich von zwei bis ca. zehn Gruppen typisch ist. Bei nicht-experimentellen Studien werden diese Gruppen durch eine vorgefundene Variable gebildet, die z. B. mittels Beobachtung oder Befragung operationalisiert wird (z. B. Vergleiche zwischen Altersgruppen, Geschlechtern, Nationalitäten auf der Basis entsprechender soziodemografischer Angaben im Fragebogen).

Die Operationalisierung von experimentellen und quasi-experimentellen unabhängigen Variablen läuft oft auf die Produktion von unterschiedlichem **Stimulusmaterial** oder die Konzeption von unterschiedlichen **Behandlungsformen** bzw. Interventionen oder Treatments hinaus. Soll etwa der Grad der Gewalthaltigkeit von Computerspielen als experimentelle unabhängige Variable variiert werden, um den Effekt auf die abhängige Variable

Aggressivität zu erfassen, so müssen mehrere Varianten eines Computerspiels mit unterschiedlicher Gewalthaltigkeit a) gezielt ausgewählt oder b) selbst produziert werden. Dabei kommt es darauf an, dass sich die Spiele möglichst nur in ihrer Gewalthaltigkeit und nicht in anderen Merkmalen unterscheiden. Würde man im Experiment z. B. die Aggressivität nach einem Denkspiel mit der Aggressivität nach dem Spielen eines Ego-Shooters vergleichen und in der Shooter-Gruppe tatsächlich erhöhte Aggressionswerte messen können, so wäre damit nicht belegt, dass es sich um einen Effekt der Gewalthaltigkeit des Spiels handelt. Denn Denkspiel und Shooter unterscheiden sich auch in vielen anderen Aspekten – von der Hintergrundmusik über die Farbgebung bis zur Spieleraktivität – deutlich voneinander. Die verschiedenen Untersuchungsbedingungen einschließlich der Stimulusmaterialien müssen idealerweise so gestaltet sein, dass sie sich exakt nur hinsichtlich der unabhängigen Variable voneinander unterscheiden und **alle anderen Aspekte der Untersuchungsbedingungen gleich** sind.

Zu beachten ist zudem, dass die **Dosierung der unabhängigen Variable** maßgeblich darüber entscheidet, welche **Effektstärke** sich in einer Studie zeigt (z. B. Messung der Aggressivität nach 10 Minuten, nach 3 Stunden oder nach 6 Monaten Nutzungszeit eines gewalthaltigen Medienangebots; ► Abschn. 14.2).

Eine theoretisch fundierte Operationalisierung der unabhängigen Variablen ist in der Experimentalforschung (► Abschn. 7.6) eine besondere Herausforderung und gleichzeitig entscheidende Voraussetzung für die Aussagekraft eines Experiments oder Quasi-Experiments.

8.3.4 Fehlinterpretation von Operationalisierungen

Die Auswahl der Indikatoren und die Wahl des Messinstrumentes sind bei **latenten Merkmalen** erklärungsbedürftig, weil eine Brücke zwischen den beobachtbaren Sachverhalten einerseits und der theoretischen Konzeptualisierung andererseits geschlagen werden muss. Bei **manifesten Variablen**, die praktisch unmittelbar als beobachtbare Indikatoren vorliegen, besteht kein besonderes konzeptuelles Überbrückungsproblem. Dennoch müssen strenggenommen auch manifeste Variablen operationalisiert werden. So etwa, wenn zur Erfassung von Alter, Geschlecht oder Wohnort entsprechende Fragen in einem Fragebogen gestellt werden (zur Messung soziodemografischer Variablen ► Abschn. 8.5.1).

Indem die Operationalisierung angibt, über welche Indikatoren und mit welchem standardisierten Messinstrument (z. B. einem standardisierten Fragebogen oder psychologischen Test) ein theoretisches Konstrukt empirisch zu erfassen ist, wird dieses greifbarer. Dabei ist im Auge zu behalten, dass die beobachteten Merkmale keine voraussetzungslosen Tatsachen, sondern immer Ergebnis eines **theoretischen Konstruktionsprozesses** sind. Der Umstand, dass eine Person auf einem Messinstrument für „Internetsucht" eine hohe Punktzahl erreicht, bedeutet nicht, dass die Person tatsächlich internetsüchtig „ist". Es bedeutet, dass ihr Verhalten und Erleben auf der Basis bestimmter theoretischer Vorannahmen mit dem Konzept der Sucht beschrieben und erklärt wird. Internetsucht als reale Tatsache – anstatt als theoretische Konstruktion – aufzufassen, käme einer unzulässigen Verdinglichung bzw. **Reifizierung** („reification") gleich. Eine andere Theorie könnte dasselbe Verhalten nicht als „Sucht", sondern als „Zwang" oder auch als „Gewohnheit" auffassen, woraus sich dann andere Schlussfolgerungen hinsichtlich Entstehung oder Behandlung ergeben würden. Empirische Forschung, die gemessene Variablen als Tatsachen auffasst, mündet in einen **naiven Empirismus bzw. Positivismus**.

Deswegen ist die **theoretische Konstruiertheit aller wissenschaftlichen Messungen** bei der Diskussion von empirischen Forschungsprozessen und ihren Ergebnissen stets zu berücksichtigen. Dies wird im quantitativen Paradigma der empirischen Sozialforschung im Rahmen der Wissenschaftstheorie des Kritischen Rationalismus ausdrücklich betont (► Abschn. 2.2.3). Theoretische Konzepte zu operationalisieren läuft somit keineswegs auf ein datengläubiges „empiristisches" oder „positivistisches" Vorgehen hinaus, vielmehr verlangt eine seriöse wissenschaftliche Operationalisierung transparente und fundierte theoretische Argumente sowohl bei der Auswahl und Konstruktion von Indikatoren und Messinstrumenten als auch bei der Interpretation der so gewonnenen quantitativen Daten.

8.4 Messung und die vier Skalenniveaus

Wurden im Zuge der Operationalisierung für ein latentes Merkmal die manifesten Indikatorvariablen ausgewählt und die Art des Messinstruments und damit auch der Datenerhebungsmethode festgelegt (z. B. Beobachtungsschema, standardisierter Fragebogen, psychologischer Test), so steht noch die Messung im engeren Sinne an, d. h., die aussagekräftige Zuordnung von numerischen Messwerten zu den beobachteten Ausprägungen der Untersuchungseinheiten auf den einzelnen Indikatorvariablen. In Abhängigkeit von der Art der Messung unterscheiden wir vier **verschiedene Messniveaus bzw. Skalenniveaus**. Je höher das Skalenniveau der Messung, umso informationshaltiger sind die erzeugten Messwerte und umso vielfältiger die Möglichkeiten der statistischen

Tabelle 8.4 Die drei respektive vier wichtigsten Skalenarten bzw. Skalenniveaus

Drei Skalenarten bzw. Skalenniveaus	Vier Skalenarten bzw. Skalenniveaus	Zulässige Transformationen	Mögliche Aussagen	Beispiele
1. Nominalskala	1. Nominalskala	Eindeutigkeitstransformation	Gleichheit, Verschiedenheit	Automarken, Krankheitsklassifikationen, Familienstand
2. Ordinalskala	2. Ordinalskala	Monotone Transformation	Größer-Kleiner-Relationen	Militärische Ränge, Windstärken
3. Kardinalskala = metrische Skala	3. Intervallskala	Lineare Transformation	Gleichheit von Differenzen	Temperatur (z. B. Celsius), Kalenderzeit, Intensität von Einstellungen
	4. Verhältnisskala	Ähnlichkeitstransformation	Gleichheit von Verhältnissen	Längenmessung, Gewichtsmessung, Häufigkeiten pro Person

Datenanalyse. Generell ist es empfehlenswert, sich im Vorfeld der Datenerhebung über den Informationsgehalt der Messwerte und die damit verbundenen statistischen Auswertungsmöglichkeiten Gedanken zu machen.

In der Sozialforschung werden die vier Skalenarten bzw. Skalenniveaus zuweilen auch zu drei Varianten zusammengefasst (Tab. 8.4). Die Daten eines bestimmten Skalenniveaus ermöglichen unterschiedliche inhaltliche Aussagen über die Variablenausprägungen. Sie sind gegenüber den jeweils zulässigen skalenspezifischen Transformationen invariant, wodurch die Möglichkeiten und Grenzen einer sinnvollen statistischen Auswertung abgesteckt werden: Im Rahmen der zulässigen Transformationen können Messwerte verrechnet werden, ohne dass sich ihre inhaltliche Aussage verändert. Jede Skalenart ist durch spezifische messtheoretische Voraussetzungen definiert, auf die wir in den folgenden Abschnitten genauer eingehen.

Ein Vergleich der vier Skalenniveaus zeigt, dass die Messungen mit zunehmender Ordnungsziffer des Skalenniveaus genauer werden. Während eine Nominalskala lediglich Äquivalenzklassen von Objekten numerisch beziffert, informieren die numerischen Werte einer Ordinalskala zusätzlich darüber, bei welchen Objekten das Merkmal stärker bzw. weniger stark ausgeprägt ist. Eine Intervallskala ist der Ordinalskala überlegen, weil hier die Größe eines Merkmalsunterschiedes bei zwei Objekten genau quantifiziert wird. Eine Verhältnisskala schließlich gestattet zusätzlich Aussagen, die die Merkmalsausprägungen verschiedener Objekte zueinander ins Verhältnis setzen.

Messungen auf den vier Skalenniveaus werden in Tab. 8.5 durch Fragebogenitems verdeutlicht.

Empirische Sachverhalte werden durch die vier Skalenarten bzw. Skalenniveaus unterschiedlich genau abgebildet. Die hieraus ableitbare Konsequenz für die Planung empirischer Untersuchungen liegt auf der Hand: Bieten sich bei einer Quantifizierung mehrere Skalenarten an, sollte diejenige mit dem **höchsten Skalenniveau** gewählt werden (Bortz & Schuster, 2010. S. 22f.). Erweist sich im Nachhinein, dass die erhobenen Daten dem angestrebten Skalenniveau letztlich nicht genügen, besteht die Möglichkeit, die erhobenen Daten auf ein niedrigeres Skalenniveau zu transformieren. (Beispiel: Zur Operationalisierung des Merkmals „Schulische Reife“ sollten Experten intervallskalierte Punkte vergeben. Im Nachhinein stellte sich heraus, dass die Experten mit dieser Aufgabe überfordert waren, so dass man beschließt, für weitere Auswertungen nur die aus den Punktzahlen ableitbare Rangfolge der Kinder zu verwenden.) Eine nachträgliche Transformation auf ein höheres Skalenniveau ist hingegen nur im Ausnahmefall möglich (▶ Abschn. 8.4.3 „Indirekte Rangordnungen“).

Wie jedoch – so lautet die zentrale Frage – wird in der Forschungspraxis entschieden, auf welchem Skalenniveau ein bestimmtes Merkmal gemessen wird? Ist es erforderlich bzw. üblich, bei jedem Merkmal die gesamte Axiomatik der mit einer Skalenart verbundenen Messstruktur empirisch zu überprüfen? Kann man – um im oben genannten Beispiel zu bleiben – wirklich guten Gewissens behaupten, die Punktzahlen zur „Schulischen Reife“ seien, wenn schon nicht intervallskaliert, so doch zumindest ordinalskaliert?

Sucht man in der Literatur nach einer Antwort auf diese Frage, so wird man feststellen, dass hierzu unterschiedliche Auffassungen vertreten werden (Hand 1996; King, Rosopa, & Minium, 2010). Unproblematisch und im Allgemeinen ungeprüft ist die Annahme, ein Merkmal sei nominalskaliert. Wohnort, Parteizugehörigkeit, Studienfach etc. sind einfache manifeste Merkmale, deren Nominalskalenqualität unstrittig ist.

Weniger eindeutig fällt die Antwort jedoch aus, wenn es darum geht zu entscheiden, ob Schulnoten, Testwerte oder auf Ratingskalen abgegebene Einstellungsmessun-

Tabelle 8.5 Messungen auf allen vier Skalenniveaus am Beispiel von Operationalisierungen des Merkmals „Rauchen"

Fragebogenitem	Messwerte für die Antwortalternativen	Skalenniveau der Variable
Sind Sie Raucher/-in? Ja ☐ Nein ☐	Ja (1) Nein (2) Ja (1) Nein (0)	**2-fach gestufte nominalskalierte (binäre, dichotome) Variable** ➜ welche Zahlen zugeordnet werden, ist egal, es müssen nur zwei unterschiedliche Zahlen sein, üblich sind 0, 1 oder 1, 2
Was rauchen Sie hauptsächlich? (bitte nur *eine* Antwort ankreuzen) Zigaretten mit Filter ☐ Zigaretten ohne Filter ☐ Cigarillos ☐ Zigarren ☐ Pfeife ☐ Anderes ☐	 Zigaretten mit Filter (1) Zigaretten ohne Filter (2) Cigarillos (3) Zigarren (4) Pfeife (5) Anderes (6)	Mehrfach bzw. **6-fach gestufte nominalskalierte (polytome) Variable** ➜ bei Einfachauswahl („forced choice") entsteht eine polytome Variable ➜ die Messwerte repräsentieren unterschiedliche Qualitäten des Rauchens
Was rauchen Sie? (Mehrfachauswahl möglich) Zigaretten mit Filter ☐ Zigaretten ohne Filter ☐ Cigarillos ☐ Zigarren ☐ Pfeife ☐ Anderes ☐	 Zigaretten mit Filter (0/1) Zigaretten ohne Filter (0/1) Cigarillos (0/1) Zigarren (0/1) Pfeife (0/1) Anderes (0/1)	**6 Variablen**, die **jeweils 2-fach gestuft nominalskaliert** sind ➜ bei Mehrfachauswahl („multiple choice") bildet jede einzelne Antwortkategorie eine neue binäre Variable
Welcher Rauchertyp sind Sie? Kettenraucher/-in ☐ Regelmäßiger Raucher/-in ☐ Gelegenheitsraucher/-in ☐ Nichtraucher/-in ☐	 Kettenraucher/-in (4) Regelmäßiger Raucher/-in (3) Gelegenheitsraucher/-in (2) Nichtraucher/-in (1)	**Ordinalskalierte Variable** ➜ die Messwerte repräsentieren eine eindeutige Rangreihe der Intensität des Rauchens: 4 (Kettenraucher) > 3 (regelmäßiger Raucher) > 2 (Gelegenheitsraucher) etc. ➜ die Messwerte können auch in umgekehrter Reihenfolge von Kettenraucher (1) bis Nichtraucher (4) vergeben werden, intuitiv am besten erfassbar ist es, wenn für starke Merkmalsausprägungen hohe Werte vergeben werden
Wie oft rauchen Sie? Nie ☐ Gelegentlich ☐ Oft ☐	 Nie (1) Gelegentlich (2) Oft (3)	**Ordinalskalierte Variable** ➜ vergebene Messwerte sollten intuitiv verständlich sein: höhere Messwerte = stärkere Ausprägung
Wie oft rauchen Sie? Nie ☐ Sehr selten ☐ Selten ☐ Gelegentlich ☐ Oft ☐ Sehr oft ☐ Fast immer ☐	 Nie (1) Sehr selten (2) Selten (3) Gelegentlich (4) Oft (5) Sehr oft (6) Fast immer (7)	**Intervallskalierte Variable** (7stufige, annähernd gleichabständige Häufigkeits-Ratingskala als Antwortformat) ➜ intuitiv verständliche Vergabe der Messwerte: höhere Messwerte = stärkere Ausprägung
Ich werde mir das Rauchen nächstes Jahr abgewöhnen. Keinesfalls ☐ Wahrscheinlich nicht ☐ Vielleicht ☐ Ziemlich wahrscheinlich ☐ Ganz sicher ☐	 Keinesfalls (1) Wahrscheinlich nicht (2) Vielleicht (3) Ziemlich wahrscheinlich (4) Ganz sicher (5)	**Intervallskalierte Variable** (5-stufige, annähernd gleichabständige Wahrscheinlichkeits-Ratingskala als Antwortformat)
Wie viele Zigaretten haben Sie gestern geraucht? ____ Zigaretten	0 Zigaretten (0) 1 Zigarette (1) 2 Zigaretten (2) 3 Zigaretten (3) ... 50 Zigaretten (50) ...	**Verhältnisskalierte Variable** (die Messwerte sind gleichabständig und haben einen absoluten Nullpunkt)

gen ordinal- oder intervallskaliert sind (zu messtheoretischen Problemen ► Abschn. 8.4.4 „Messtheoretische Probleme bei Ratingskalen“). Eine richtige Entscheidung ist insoweit von Bedeutung, als die Berechnung von sinnvoll interpretierbaren Mittelwerten und anderen wichtigen statistischen Maßen nur bei intervallskalierten Merkmalen zu rechtfertigen ist. Das heißt, dass für ordinalskalierte Daten andere statistische Verfahren einzusetzen sind als für intervallskalierte Daten.

Die übliche Forschungspraxis verzichtet auf eine empirische Überprüfung der jeweiligen Skalenaxiomatik. Die meisten Messungen sind **Per-fiat-Messungen** (Messungen „durch Vertrauen“), die auf Erhebungsinstrumenten (Fragebögen, Tests, Ratingskalen etc.) basieren, von denen man annimmt, sie würden das jeweilige Merkmal auf einer Intervallskala messen. Es kann so der gesamte statistische „Apparat“ für Intervallskalen eingesetzt werden, der erheblich differenziertere Auswertungen ermöglicht als die Verfahren für Ordinal- oder Nominaldaten (Rasmussen 1989; Zumbo & Zimmerman, 1993). Hinter dieser „liberalen“ Auffassung steht die Überzeugung, dass die Bestätigung einer Forschungshypothese durch die Annahme eines falschen Skalenniveaus eher **erschwert** wird. Strengere Auffassungen fordern jedoch eine ausdrückliche Überprüfung der messtheoretischen Annahmen, dies ist z. B. im Rahmen der probabilistischen Testtheorie möglich (► Abschn. 10.4.4).

Im Folgenden werden wir das Konzept der „**Messung**“ von sozialwissenschaftlichen Sachverhalten etwas vertiefen (► Abschn. 8.4.1) und anschließend jedes einzelne **Skalenniveau** noch einmal detailliert mit seinen messtheoretischen Eigenschaften, Problemen und Operationalisierungsvarianten erörtern (► Abschn. 8.4.5). Von besonderem praktischem Interesse ist dabei die Intervallskala. Intervallskalierte Daten werden sehr oft erhoben, indem man z. B. im Interview oder Fragebogen **Ratingskalen** (z. B. „stimmt gar nicht – wenig – ziemlich – völlig“) als Antwortvorgaben präsentiert. Die Konstruktion derartiger Ratingskalen muss methodischen Standards folgen, damit die so gewonnenen Daten Intervallskalencharakter beanspruchen können. Schließlich wenden wir uns noch der **Skalentransformation** zu, also der Umwandlung von Daten eines Skalenniveaus auf ein niedrigeres oder höheres Niveau (► Abschn. 8.4.6).

8.4.1 Messung

Das „Messen“ wird in der Alltagssprache meistens mit physikalischen Vorstellungen in Verbindung gebracht. Dabei bezeichnet man als **fundamentale Messung** das Bestimmen einer (Maß-)Zahl als das Vielfache einer Einheit (z. B. Messungen mit einem Zollstock oder einer Balkenwaage). Für derartige Messungen ist der Begriff „**Einheit**“ zentral. Man wählt hierfür eine in der Natur vorgegebene Größe (wie z. B. die Ladung eines Elektrons als Einheit des Merkmals „elektrische Ladung“) oder man legt aus Gründen der Zweckmäßigkeit willkürlich eine Größe als Normeinheit fest (z. B. der in Paris niedergelegte „Archivmeter“ bzw. „Urmeter“). Eine physikalische Messung besteht darin, möglichst genau zu erfassen, wie oft die gewählte Merkmalseinheit in dem zu messenden Objekt enthalten ist.

Eine Übertragung dieser Messvorstellung auf die Sozialwissenschaften scheitert daran, dass „Einheiten“ in diesem Sinne in den Sozialwissenschaften bislang fehlen. Dennoch sind auch hier – allerdings mit einer weiter gefassten Messkonzeption – Messoperationen möglich.

Allgemein formuliert besteht eine Messoperation im aussagekräftigen Zuordnen von Zahlen zu Objekten. Die logisch-mathematische Analyse dieser Zuordnungen und die Spezifizierung von Zuordnungsregeln sind Aufgaben der **Messtheorie** (► Messtheorie).

Messung – Eine Messung („measurement“) meint in der quantitativen Sozialforschung eine Zuordnung von Zahlen zu Objekten oder Ereignissen, sofern diese Zuordnung eine homomorphe (strukturerhaltende) Abbildung eines empirischen Relativs in ein numerisches Relativ ist (Orth 1983, S. 138).

Diese Definition sei kurz erläutert: Ein **empirisches Relativ** ist eine Menge an Objekten, z. B. an Personen, Ereignissen, Medienangeboten. In einem empirischen Relativ gibt es eine oder mehrere Relationen, die die Beziehung zwischen den Objekten charakterisieren (größer als; kleiner als; gleich etc.).

Ein **numerisches Relativ** ist eine Menge an Zahlen (z. B. 1; 2; 15; 17.5 …), die eine mathematische Relation aufweisen (>; <; = etc.).

Eine **homomorphe (strukturerhaltende) Abbildung** ist eine Funktion. Diese ordnet jedem Objekt des empirischen Relativs (z. B. jedem Kind einer Schulklasse) genau eine Zahl des numerischen Relativs zu (z. B. die jeweilige Körpergröße in cm). Voraussetzung für Homomorphismus ist, dass für jedes Objekt im empirischen Relativ genau ein Element im numerischen Relativ gefunden wird (d. h. für dasselbe Kind dürfen nicht mehrere verschiedene Körpergrößen als Messwerte erscheinen). Außerdem müssen die Relationen im empirischen Relativ auch im numerischen Relativ gültig sein (z. B. wenn Kind A größer ist als Kind B: empirisches Relativ, dann muss auch der Körpergrößenmesswert von Kind A größer sein als der Körpergrößenmesswert von Kind B: numerisches Relativ). Sind diese beiden Voraussetzungen erfüllt, spricht man von einer eindeutigen Zuordnung.