# Forschungsmethoden

*Tobias Engelschalk, Martin Daumiller, Marion Reindl und Markus Dresel*




D. Urhahne, M. Dresel, F. Fischer (Hrsg.), *Psychologie für den Lehrberuf*, https://doi.org/10.1007/978-3-662-55754-9_27

## 27.1 Macht Kaugummikauen schlau?

Steigert Kaugummikauen wirklich die Leistungsfähigkeit von Schülerinnen und Schülern? Sollte ich im Unterricht lieber keine Witze machen und wenn doch, welche? Wie können die Möglichkeiten digitaler Medien im Unterricht lerneffektiv genutzt werden? Wie entstehen Disziplinschwierigkeiten und wodurch kann ihnen vorgebeugt werden? Was kennzeichnet effektives selbstgesteuertes Lernen und wie kann dieses gefördert werden?

Vor diesen und ähnlichen Fragen stehen Lehrkräfte tagtäglich in ihrer Arbeit. Zu vielen schulrelevanten Themen gibt es Alltagsannahmen und Alltagstheorien, die manchmal sogar zutreffen. In den meisten Fällen sind diese im Lichte wissenschaftlicher Forschung aber zu undifferenziert, wie etwa die Alltagstheorie, dass Humor im Unterricht keinen Platz hat (vgl. Bieg, Grassinger & Dresel 2017). Viele Fragen, mit denen Lehrkräfte konfrontiert sind, resultieren aus aktuellen gesellschaftlichen Herausforderungen wie der fortschreitenden Digitalisierung. Hier versucht die Forschung schnell tragfähige Erkenntnisse zu liefern (▶ Kap. 19). Zu anderen Fragen, wie zu Disziplinschwierigkeiten oder selbstreguliertem Lernen, existiert dagegen bereits ein umfassender Korpus belastbarer und differenzierter Antworten aus einer Vielzahl von Forschungsstudien (▶ Kap. 4, 18).

Die Ergebnisse der schulbezogenen Psychologie und der empirischen Bildungsforschung bieten Lehrkräften einen reichen Fundus praxistauglichen Wissens, um professionelle Kompetenzen und Lehrtätigkeit kontinuierlich zu entwickeln. Dazu passend legen die Standards für die Lehrerbildung fest, dass Lehrerinnen und Lehrer Forschungsergebnisse rezipieren, bewerten und für die eigene Tätigkeit nutzen können sollen – ganz im Sinne der Auffassung ihres Berufs als ständige Lernaufgabe (KMK 2004). Um diesen Anforderungen mit Sachverstand, Selbstvertrauen und gesunder Skepsis begegnen zu können, ist es notwendig, die Aussagekraft empirischer Studien selbst beurteilen und sich eine eigene Meinung bilden zu können. Zentrales Anliegen dieses Kapitels ist es, genau solche Kompetenzen für den Bereich der empirischen, d. h. auf systematischer Erfahrung beruhenden Forschung zu vermitteln. Dazu werden grundlegende forschungsmethodische Begriffe und Vorgehensweisen vorgestellt, die Verständnis und Interpretation empirischer Arbeiten ermöglichen. Nebenbei soll auch die Lust am Forschen geweckt und zur Durchführung eigener Untersuchungen ermutigt werden.

Im folgenden ▶ Abschn. 27.2 werden dazu Begriffe und Konzepte eingeführt, die für ein Verständnis des Forschungsprozesses notwendig sind. In ▶ Abschn. 27.3 geht es darum, wie nicht beobachtbare, psychische Merkmale erfasst werden können. In ▶ Abschn. 27.4 werden die wichtigsten Gestaltungsmöglichkeiten (Forschungsdesigns) empirischer Studien besprochen. Der darauf folgende ▶ Abschn. 27.5 gibt eine Einführung in grundlegende statistische Methoden, die zur Auswertung gewonnener Daten nötig sind. Am Beispiel einer konkreten Studie (Rost, Wirthwein, Frey & Becker 2010), die den Mythos zur leistungsförderlichen Wirkung des Kaugummikauens entkräftet, illustriert ▶ Abschn. 27.6 schließlich die sachgemäße und gewinnbringende Nutzung publizierter Forschungsergebnisse.

## 27.2 Wie entsteht empirisch gesichertes Wissen?

### 27.2.1 Wechselwirkungen zwischen Praxis, Theorie und Empirie

Ein häufig bemühtes Vorurteil besagt, Wissenschaftlerinnen und Wissenschaftler würden ihren Elfenbeinturm kaum verlassen, Praktikerinnen und Praktiker würden sich dagegen oft den Erkenntnissen der Wissenschaft verschließen. Gerade die bildungswissenschaftliche Forschung ist aber stark von praktischen Fragestellungen inspiriert (▶ Kap. 18). Umgekehrt ist professionelles Unterrichten ohne fundiertes Theoriewissen undenkbar und profitiert von den Ergebnissen der empirischen Forschung.

Besondere Bedeutung im Verhältnis zwischen Praxis und Theorie kommt der Empirie zu. Damit wird die systematische Sammlung von Informationen bezeichnet, die auf methodisch kontrollierten Datenquellen basiert (Döring & Bortz 2016). Empirisches Arbeiten hat zum Ziel, praktische Phänomene zu beschreiben, zu ordnen und zu quantifizieren. Auf diese Weise lassen sich übergeordnete Regeln und Muster finden, die es ermöglichen, Beobachtetes mit bestehendem Wissen zu verknüpfen und Beziehungen zwischen Phänomenen zu verstehen. So entwickeln sich wissenschaftliche Theorien.

Eine Theorie ist ein „allgemeines Prinzip, das aufgestellt wurde, um eine Gruppe von Beziehungen zwischen Ereignissen zu klären. Etwas ausführlicher gesagt: Eine Theorie verknüpft induktiv (vom Speziellen auf das Allgemeine schließend) oder deduktiv (vom Allgemeinen auf das Spezielle schließend) gewonnene Erkenntnisse eines Wissensbereichs systematisch miteinander, wodurch Einzelphänomene regelhaft erklärt werden“ (Rost 2013, S. 17). Eine Theorie ist also alles andere als ohne Bezug zum realen Leben. Vielmehr ermöglichen wissenschaftliche Theorien einen klareren Blick auf die häufig komplexen Vorgänge und Entscheidungsoptionen in der Praxis. Wichtige Aspekte können so leichter identifiziert werden, während sich die Bedeutungen anderer im Lichte theoretischer Überlegungen möglicherweise relativieren. Fragt sich beispielsweise eine Lehrkraft, ob ihr Unterrichtsstil motivierend sei, könnte sie zur Klärung die Selbstbestimmungstheorie der Motivation von Deci und Ryan (1993) heranziehen (▶ Kap. 11). Mit dieser theoretischen Sichtweise würde sie vielleicht feststellen, dass sie zwei wesentliche Aspekte, die Unterstützung des Kompetenzerlebens und der sozialen Eingebundenheit bereits gut realisiert, zukünftig aber noch stärker auf den Aspekt der Autonomieunterstützung achten sollte. Durch die Theorie kann die Lehrkraft ihren Unterricht über ein differenziertes Raster relevan-

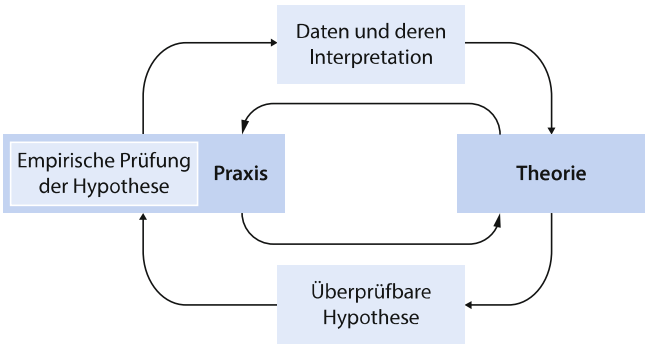



**Abb. 27.1** Zusammenspiel von Praxis, Theorie und Empirie

ter Merkmale bewerten. Dieses Beispiel veranschaulicht die Bedeutung der wissenschaftlichen Theorie für typische Fragestellungen der Praxis. Ist keine gesicherte Theorie vorhanden, wird fehlendes Wissen nicht selten durch Alltagstheorien ersetzt.

**Im Fokus: Alltagstheorien**

Alltagstheorien sind Annahmen, die sich auf Basis persönlicher Erfahrungen herausgebildet haben (Wirtz 2017). Oft scheinen sie auch deshalb plausibel, weil sie von anderen Personen geteilt werden. In Abgrenzung zu wissenschaftlichen Theorien sind Alltagstheorien meist abseits einer systematischen Sammlung von Daten entstanden, enthalten oft subjektive, kaum überprüfbare Einschätzungen und beinhalten häufig unzulässige Verallgemeinerungen oder gar Vorurteile. Alltagstheorien müssen nicht stets falsch sein – sie sind jedoch nicht für wissenschaftliche Aussagen nutzbar und zudem ungeeignet, pädagogisches Handeln zu begründen.

Wissenschaftliche Theorien repräsentieren niemals endgültiges oder abgeschlossenes Wissen. Vielmehr sind Theorien immer nur Momentaufnahmen in einem stetigen Erkenntnisprozess, der sich aus dem Zusammenspiel von Praxis, Theorie und Empirie ergibt (Abb. 27.1).

Der Kreislauf soll verdeutlichen, dass Praxis und Theorie in Wechselwirkung zueinanderstehen: Beobachtungen in der Praxis stoßen die (Weiter-)Entwicklung von wissenschaftlichen Theorien an. Aus diesen ergeben sich konkrete Aussagen oder Vorhersagen (Hypothesen), die in der Praxis empirisch überprüft werden können. Die so gewonnenen Daten lassen sich dann mit Blick auf die zugrunde gelegte Theorie interpretieren. Nicht selten führen die Ergebnisse empirischer Studien dazu, dass die Gültigkeit von Theorien in Frage gestellt oder sie entsprechend neuer Erkenntnisse angepasst werden müssen.

## 27.2.2 Hypothesen und Variablen

Eine wissenschaftliche Hypothese ist eine auf der Basis bisherigen (theoretischen) Wissens gut begründbare Aussage oder Vermutung über Beziehungen zwischen bestimmten Gegebenheiten, wobei gesichertes Wissen oft noch unvollständig ist (Hussy, Schreier & Echterhoff 2010). Wissenschaftliche Hypothesen müssen einigen Anforderungen genügen (siehe Im Fokus), die den Unterschied zur alltagssprachlichen Verwendung im Sinne einer persönlichen Vermutung oder Meinung markieren. Je nach Forschungsanliegen lassen sich verschiedene Arten wissenschaftlicher Hypothesen unterschieden.

**Im Fokus: Wissenschaftliche Hypothesen**

Wissenschaftliche Hypothesen weisen drei Merkmale auf (Hussy et al. 2010):

1. Sie beziehen sich auf reale, theoretisch begründbare und empirisch zugängliche Sachverhalte.
2. Sie stellen eine allgemeingültige, über den Einzelfall oder ein einmaliges Ereignis hinausgehende Behauptung dar.
3. Ihre Aussagen müssen falsifizierbar sein, d. h. auch widersprechende Ereignisse müssen grundsätzlich denkbar und überprüfbar sein.

Einige wichtige Arten von Hypothesen mit Beispielen:

- Zusammenhangshypothese: Je mehr Interesse Schülerinnen und Schüler für ein bestimmtes Fach haben, umso mehr Lernstrategien setzen sie in diesem Fach ein.
- Unterschiedshypothese: Jungen und Mädchen unterscheiden sich im Ausmaß der von ihnen eingesetzten Lernstrategien.
- Veränderungshypothese: Die Qualität eingesetzter Lernstrategien nimmt im Verlauf der Schulzeit zu.

Angenommen eine Hypothese besagt, dass Trainings zum selbstregulierten Lernen im Unterricht zu einer verbesserten Selbstregulation führen. Um diese bewusst allgemeingültig formulierte Hypothese testen zu können, muss sie so konkretisiert (operationalisiert) werden, dass eine Prüfung in der realen Welt möglich wird (Für welchen Personenkreis wird die Hypothese getestet? In welcher Form wird selbstreguliertes Lernen trainiert? Wie wird die Selbstregulationsfähigkeit gemessen? etc.). Dies taten beispielsweise Labuhn, Bögeholz und Hasselhorn (2008) in einer Studie zur Wirkung von gezielten Anregungen zum selbstregulierten Lernen während des Unterrichts zum Thema Ernährung. Eine Gruppe von Gesamtschülerinnen und -schülern erhielt zusätzliche Anregungen zum selbstregulierten Lernen, eine andere keine. Mit einem Fragebogen wurde das selbstregulierte Lernen vor und nach der Teilnahme an den Unterrichtseinheiten festgestellt. Es zeigte sich, dass nur bei der Interventionsgruppe eine Verbesserung im selbstregulierten Lernen festzustellen war. Somit konnte die in der Hypothese geäußerte Vermutung beibehalten werden.

Variablen sind veränderliche Größen, die Objekte anhand von Eigenschaften oder Merkmalen beschreiben (Eid, Gollwitzer & Schmitt 2015). In der Psychologie sind dies häufig Merkmale von Personen wie beispielsweise die Variablen

„Geschlecht“ oder „Intelligenz“. Variablen fungieren also als Platzhalter, die während der Durchführung einer Untersuchung für jede Person mit konkreten Werten gefüllt werden (z. B. weiblich, IQ-Wert 107). In gleicher Weise werden Variablen genutzt, um die Bedingungen der Untersuchungssituation zu beschreiben. So könnte in einer Variablen vermerkt sein, ob das betreffende Kind eine Unterrichtseinheit mit oder ohne Anregung zur Selbstregulation besucht hat. Abhängig von der jeweiligen Hypothese einer Studie kann einzelnen Variablen eine spezielle Bedeutung zukommen (▶ Im Fokus).

**Im Fokus: Wichtige Arten von Variablen**

Eine **unabhängige Variable** (UV) ist eine frei veränderliche Einflussgröße, von der angenommen wird, dass sie andere Variablen der Untersuchung beeinflusst, selbst jedoch von diesen unabhängig ist. In der Studie von Labuhn et al. (2008) ist die unabhängige Variable der Sachverhalt, dass ein Kind die Unterrichtseinheiten mit oder eben ohne das Training zur Selbstregulation besucht hat.

Als **abhängige Variable** (AV) wird diejenige Variable bezeichnet, auf die sich die unabhängige Variable auswirken soll. Im Beispiel wäre dies die per Fragebogen erfasste Selbstregulation des Lernens.

Eine **Mediatorvariable** (auch: intervenierende Variable) ist ein Merkmal, das die Wirkung einer anderen Variablen zumindest teilweise vermittelt. So könnte die Wirkung eines Trainings zum selbstregulierten Lernen nicht direkt zu verbesserter Selbstregulation führen, sondern zunächst die Motivation zum Einsatz von Lernstrategien erhöhen. Führt letztlich die erhöhte Motivation zu einem verstärkten Einsatz von Selbstregulationsstrategien, wäre die Motivation eine Mediatorvariable, die die Wirkung des Trainings auf die Selbstregulationsfähigkeit mediiert.

Eine **Moderatorvariable** beeinflusst die Stärke der Wirkung einer unabhängigen Variablen auf eine abhängige Variable. Würden beispielsweise Mädchen stärker von dem Selbstregulationstraining profitieren als Jungen, wäre das Geschlecht eine Moderatorvariable, die den Trainingseffekt auf die Selbstregulation moderiert.

### 27.2.3 Wichtige Schritte im Forschungsprozess

In aller Regel vollzieht sich empirische Forschung in einer Reihe typischer Schritte. Sie zu kennen ermöglicht es, das in Forschungsarbeiten beschriebene Vorgehen besser zu verstehen und einzuordnen. Dies ist eine wichtige Voraussetzung dafür, die Qualität und Aussagekraft einer empirischen Studie beurteilen zu können. ◘ Tab. 27.1 liefert eine Übersicht über die Schritte im Forschungsprozess. Daran wird ebenso das besprochene Zusammenspiel zwischen Praxis, Empirie und Theorie deutlich.

Im Folgenden werden die aufgeführten Schritte genauer erläutert.

1. Empirische Forschung beginnt immer mit einer Forschungsfrage. Die Quellen einer solchen Frage können vielfältig sein. Sie können aus theoretischen Überlegungen, einer Beobachtung von Ungereimtheiten in der Praxis, dem Ziel ein Forschungsergebnis nachprüfen zu wollen oder auch aus einem spezifischen Forschungsauftrag resultieren. Beispielsweise wird die Förderung des selbstregulierten Lernens als explizites und wichtiges Bildungsziel angesehen. Entsprechend werden Untersuchungen, die sich mit Fragestellungen zum selbstgesteuerten Lernen bei Schülerinnen und Schülern beschäftigen, gezielt gefördert.
2. Nachdem ein Erkenntnisinteresse entstanden ist, beginnt die Suche nach dazu passenden Informationen. Wissenschaftlerinnen und Wissenschaftler sammeln zunächst, was bereits über das zu erklärende Phänomen bekannt ist und welche theoretischen Überlegungen und Modelle für die Fragestellung relevant sind. Ebenfalls bedeutsam sind empirische Studien, die dazu vorliegen. Mit Hilfe von Literaturdatenbanken lassen sich Artikel in Fachzeitschriften und Büchern recherchieren. So führen Labuhn et al. (2008) beispielsweise aus, dass bereits in früheren Studien die Wirkung von Trainingsmaßnahmen zum selbstregulierten Lernen nachgewiesen werden konnte, die durchgeführten Trainings allerdings meist außerhalb des Unterrichts stattgefunden hatten. Aus einer groben, interessant erscheinenden Forschungsidee resultiert schließlich eine präzise formulierte Fragestellung für eine empirische Untersuchung, die den aktuellen Forschungsstand aufgreift und den Anspruch hat, diesen sinnvoll zu erweitern.
3. Auf Basis des theoretischen Wissens und der Forschungsfrage werden nun eine oder mehrere wissenschaftliche Hypothesen formuliert. Bei Labuhn et al. (2008) war dies u.a. eine Hypothese, die sinngemäß lautete: Anregungen zum selbstregulierten Lernen im Unterricht führen zu einer Verbesserung der Selbstregulation.
4. Um eine Hypothese in der realen Welt zu testen, bedarf es einer Konkretisierung ihrer Bestandteile in empirisch zugängliche Aussagen. Anschaulich gesprochen geht es um die „Messbarmachung“ der Hypothese, was auch als Operationalisierung bezeichnet wird. So operationalisierten Labuhn et al. (2008) ihre abhängige Variable (Selbstregulation) als Mittelwert der Antworten auf insgesamt 85 Fragen zum selbstregulierten Lernen, die den Schülerinnen und Schülern vorgelegt worden waren.
5. Im nächsten Schritt wird die gesamte Planung der Untersuchung in den Blick genommen. Je nach Hypothese, Ressourcen und praktischen Gegebenheiten werden Untersuchungsdesign und Stichprobe gewählt. Außerdem ist in dieser Phase zu entscheiden, welche Messinstrumente eingesetzt werden sollen.
6. Der darauf folgende Schritt der Umsetzung beinhaltet die Erstellung der Untersuchungsmaterialien, die Konstruktion von Erhebungsverfahren, die Rekrutierung der Stichprobe sowie die Datenerhebung selbst. Während dieses

**Tabelle 27.1** Schritte im Forschungsprozess

| Schritt | Beschreibung |
|---|---|
| 1. Forschungsfrage | Beobachtungen in der Praxis (induktiv) oder Ableitungen aus einer Theorie (deduktiv) führen zu einer Fragestellung, die empirisch geklärt werden soll |
| 2. Suche nach Informationen | Theoretisches Wissen und bisherige Studien werden gesichtet, was auch der Präzisierung der Forschungsfrage dient |
| 3. Aufstellen von Hypothesen | Die Forschungsfrage wird so formuliert, dass allgemeingültige und falsifizierbare Aussagen (Hypothesen) entstehen |
| 4. Operationalisierung | Die Aussagen der Hypothesen werden so transformiert, dass alle enthaltenen Aspekte einer Messung zugänglich werden |
| 5. Planung der Untersuchung | Stichprobe, Anlage der Untersuchung und Erhebungsmethoden werden festgelegt und begründet |
| 6. Umsetzung | Untersuchungsmaterialien werden erstellt, teilnehmende Personen rekrutiert. Die eigentliche Untersuchung wird durchgeführt |
| 7. Datenauswertung | Die erhobenen Daten werden aufbereitet und mit Bezug zu den Hypothesen analysiert |
| 8. Beantwortung der Fragestellung | Die Ergebnisse der Datenauswertung entscheiden über Annahme oder Ablehnung der Hypothesen und werden im Hinblick auf die Forschungsfragen interpretiert |
| 9. Publikation | Die Studie wird der (Fach-)Öffentlichkeit zugänglich gemacht |

Prozesses ist eine sorgfältige Protokollierung wichtig, sodass im Nachhinein etwaige Probleme analysiert werden können und die Untersuchung für andere Forscherinnen und Forscher nachvollziehbar wird.

7. In der Phase der Auswertung werden die erhobenen Daten zunächst in ein geeignetes Format gebracht. Meist geschieht dies durch numerische Kodierung in Zahlen und Erstellung einer Datenmatrix. Mit Verfahren der beschreibenden und der schließenden Statistik werden die Daten analysiert.
8. Auf Basis der Ergebnisse wird entschieden, ob die Hypothese angenommen werden kann oder abzulehnen ist. Damit kann die ursprüngliche Forschungsfrage differenziert beantwortet und diskutiert werden.
9. Im letzten Schritt erfolgt die Veröffentlichung. Damit werden die Ergebnisse der Untersuchung der Öffentlichkeit mitgeteilt, sei es durch die Präsentation auf einer Tagung oder durch die Veröffentlichung in einem Fachartikel.

## 27.3 Erhebungsmethoden

### 27.3.1 Konstrukte und die Schwierigkeiten ihrer Messung

In psychologischen Untersuchungen spielt die Erfassung der Merkmale, über die Aussagen zu treffen sind, eine zentrale Rolle. Dies ist keine triviale Angelegenheit (▶ Kap. 24). Im Unterschied zu vielen physikalischen Größen (z. B. Größe, Masse, Geschwindigkeit), die zumindest im Alltag mit großer Präzision direkt gemessen werden können, sind psychische Merkmale häufig nicht direkt beobachtbar. So lässt sich etwa selbstreguliertes Lernens nicht direkt beobachten, sondern nur aus äußeren Indikatoren erschließen. Wenn die Schülerin Nora beim Lesen eines Lerntextes spontan Zeichnungen zu den Textinhalten generiert, während des Lesens wiederholt darüber nachdenkt, ob sie den Inhalt versteht, und zudem ihr Smartphone ausschaltet, lässt dies auf eine angemessene Selbstregulation ihres Lernprozesses schließen (▶ Kap. 4; Schwamborn, Mayer, Thillmann, Leopold & Leutner 2010). In ähnlicher Weise lassen sich auch andere Merkmale von Personen (z. B. Prüfungsangst, Intelligenz und Wissen von Schülerinnen und Schülern oder professionelle Kompetenzen und Unterrichtsqualität von Lehrkräften) oder Gruppen (z. B. Klassenklima) ausschließlich indirekt über bestimmte, genau zu definierende Indikatoren erschließen.

> Ein Konstrukt (auch: hypothetisches Konstrukt) ist ein Begriff, der sich auf ein nicht direkt messbares Merkmal von Personen oder Gruppen bezieht (Wirtz 2017). Konstrukte sind keine frei erfundenen Gedankengebäude, sondern werden aus theoretischen Zusammenhängen heraus erschlossen – sie sind Bestandteile von theoretischen Aussagen. Die Ausprägung eines Konstrukts kann nur indirekt aus messbaren Indikatoren erschlossen werden.
>
> Als Operationalisierung wird die Messbarmachung eines Konstrukts mittels empirisch fassbarer und quantifizierbarer Größen bezeichnet (Eid et al. 2015).
>
> Unter Messen wird die Zuordnung von Zahlen zu Objekten (z. B. Ausprägungen von Merkmalen) verstanden; dabei sollen sich in den zugeordneten Zahlen die Relationen, die zwischen den Objekten bestehen, widerspiegeln (Wirtz 2017).

Viele Merkmale, mit denen sich die Psychologie beschäftigt, umfassen mehrere Facetten oder Komponenten – sind also mehrdimensional. Im Beispiel des selbstregulier-

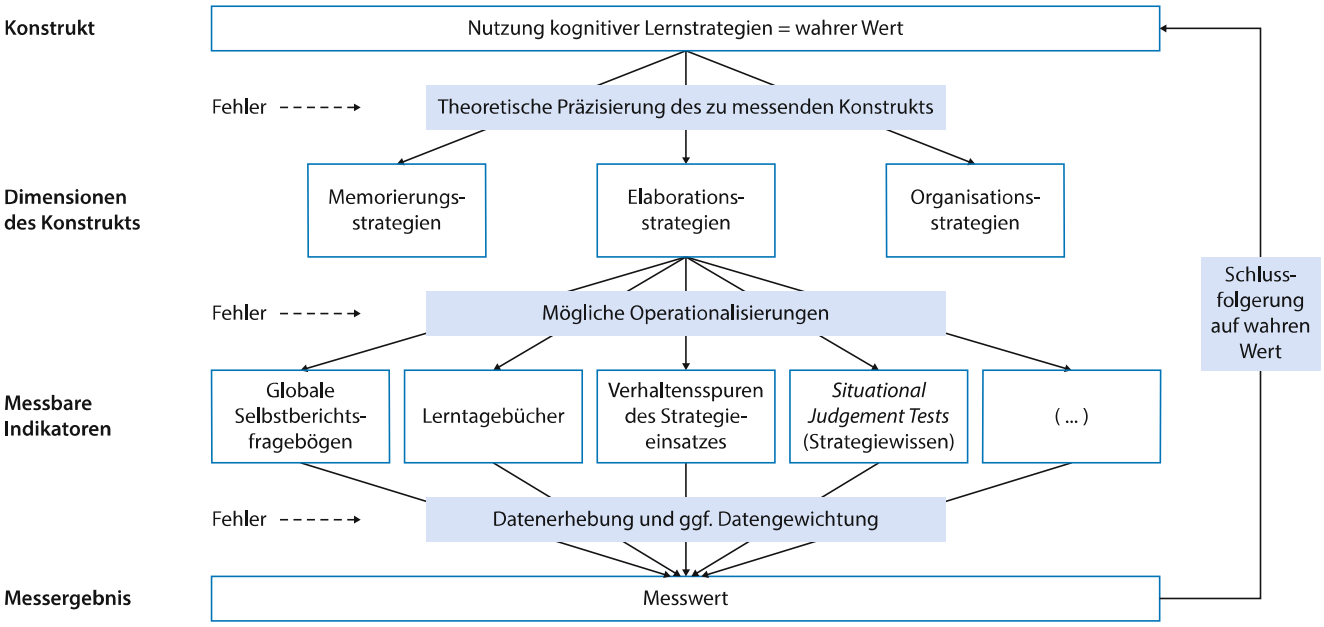



**Abb. 27.2** Konstrukte und deren Messung durch beobachtbare Indikatoren (modifiziert nach Hesse & Latzko 2017)

ten Lernens lassen sich etwa die Nutzung von kognitiven Lernstrategien, metakognitiven Kontrollstrategien und Strategien des Ressourcenmanagements voneinander abgrenzen. Innerhalb der kognitiven Lernstrategien lassen sich nochmals Memorierungs-, Elaborations- und Organisationsstrategien unterscheiden (Boekaerts 1999; Pintrich & Garcia 1994; Weinstein & Hume 1998). An dem Beispiel lässt sich erkennen, dass die Dimensionierung von Konstrukten auch hierarchisch sein kann. Solche theoretischen Dimensionierungen sind nützlich, z. B. um die mit einem Phänomen im Zusammenhang stehenden Konsequenzen besser vorhersagen und daraus praktische Fördermöglichkeiten ableiten zu können. Hier kann die Theorie über die Struktur von Lernstrategien zur Unterstützung des Wissenserwerbs genutzt werden: Klar förderliche Effekte haben Elaborations- und Organisationsstrategien (vor allem, wenn sie metakognitiv überwacht und kontrolliert werden), während Memorierungsstrategien oftmals weniger effektiv sind – erstere sollten dementsprechend vorrangig gefördert werden (Artelt 1999). Der praktische Nutzen ist umso größer, je angemessener die theoretische Präzisierung und Dimensionierung des Konstrukts ist (Abb. 27.2).

Die tatsächliche Operationalisierung der auf diese Weise präzisierten Konstrukte erfolgt anhand von messbaren Indikatoren, die eine ausreichende Messgüte aufweisen müssen. Darunter sind Messinstrumente zu verstehen, die die betreffenden Konstruktdimensionen erfassen und dabei die Gütekriterien psychologisch-diagnostischer Verfahren möglichst gut erfüllen – je mehr, desto besser. Die drei Hauptgütekriterien sind (▶ Kap. 24):

- Objektivität: Das Messinstrument erbringt Messungen, die möglichst unabhängig von den Personen sind, die die Messung durchführen (z. B. Beobachterinnen und Beobachter).
- Reliabilität: Das Messinstrument erfasst das Konstrukt möglichst messgenau (präzise).
- Validität: Das Messinstrument misst möglichst das, was es zu messen vorgibt – d. h. es erfasst das Konstrukt in Übereinstimmung mit dessen Inhalt, Struktur und Beziehungen zu anderen Merkmalen.

Neben den drei Hauptgütekriterien, die ein Messinstrument notwendigerweise in hinreichendem Maße erfüllen muss, um belastbare Schlussfolgerungen zu ermöglichen, gibt es eine Reihe von Nebengütekriterien, die die Qualität eines Messinstruments verbessern (z. B. Normierung, Ökonomie, Fairness, Zumutbarkeit, Unverfälschbarkeit). Einige von ihnen sind im Forschungsprozess weniger wichtig als in der Individualdiagnostik – dies gilt insbesondere für die Normierung, die für die Interpretation der Ergebnisse einzelner Personen essentiell, in der Forschung mit größeren Gruppen aber meist nicht nötig sind. Deshalb existieren in der psychologischen Forschung weit mehr Erhebungsansätze und -verfahren als für individualdiagnostische Zwecke zur Verfügung stehen.

### 27.3.2 Messansätze in der Psychologie

Typischerweise lassen sich bei der Operationalisierung eines bestimmten Konstrukts verschiedene Messinstrumente nutzen, die auf unterschiedlichen Indikatoren und Messansätzen beruhen (Abb. 27.2). So könnte die Nutzung von Elaborationsstrategien durch globale Selbstberichtsfragebögen (z. B. Baumert, Heyn & Köller 1992), standardisierte Lerntagebücher (Schmitz & Wiese 2006), die Analyse von Verhaltensspuren wie die spontane Generierung von Visualisierungen (Schwamborn et al. 2010), psychologische Tests zur Erfassung

des Wissens über kognitive Lernstrategien (z. B. Schlagmüller & Schneider 2007) oder durch weitere Messansätze operationalisiert werden. Jedes Messinstrument hat dabei sowohl spezifische Stärken als auch blinde Flecken – keines wird das Konstrukt vollständig erfassen. Deshalb sollte nach Möglichkeit ein multimethodales Vorgehen gewählt werden, bei dem Konstrukte mit verschiedenen Erhebungsmethoden operationalisiert werden (Eid et al. 2015).

Aufgrund der konzeptuellen Fehler bei der theoretischen Präzisierung des Konstrukts, der systematischen und unsystematischen Messfehler sowie der Fehler, die auch bei der Erfassung und Weiterverarbeitung der Daten auftreten können und sich allesamt niemals vollständig vermeiden lassen, ist jeder Messwert grundsätzlich fehlerbehaftet. Wichtig ist, ein Bewusstsein dafür zu entwickeln. Ein wichtiges Ziel bei der Entwicklung von Messinstrumenten und der Planung von Untersuchungen ist es aber, die genannten Fehler so klein wie irgend möglich zu halten. Nur wenn dies nachweislich gelungen ist und die Mindestanforderungen an Objektivität, Reliabilität und Validität erfüllt sind, sind belastbare Aussagen möglich (▶ Kap. 24). Werden aus Untersuchungen mit Messinstrumenten, die diese Mindestanforderungen nicht erfüllen, dennoch Aussagen abgeleitet, kann dies zu schwerwiegenden Problemen für Theorie und pädagogische Praxis führen (z. B. unzutreffendes Verständnis praktischer Phänomene, Behandlung von Schülerinnen und Schülern mit unwirksamen oder gar schädlichen Mitteln).

### 27.3.3 Überblick über Erhebungsmethoden

Die psychologische Forschung nutzt eine große Zahl verschiedener Methoden, um Erleben, Kognition und Verhalten zu erfassen. Im Folgenden wird ein Überblick über die wichtigsten Erhebungsmethoden der auf die Schule bezogenen Psychologie gegeben (ausführlichere Erläuterungen zu vielen Methoden ▶ Kap. 24, 25).

#### ■ Verhaltens- und Unterrichtsbeobachtung

Eine Beobachtung ist oftmals das Erhebungsverfahren der Wahl, wenn Verhalten im engeren Sinne erfasst werden soll – also z. B. Lern- oder Sozialverhalten von Schülerinnen und Schülern oder Lehrverhalten von Lehrpersonen. Meist ist eine systematische Beobachtung angebracht, die theoriegeleitet einen engen Ausschnitt des Verhaltensstroms fokussiert und dazu bestimmte Beobachtungsinstrumente und Protokollsysteme nutzt. Eine unsystematische Beobachtung, die den gesamten Verhaltensstrom einbezieht, ist durch subjektive Wahrnehmungen und Beobachtungsfehler dominiert – sie kann keine hinreichende Testgüte erzielen, entspricht weitgehend der Alltagsbeobachtung und ist allenfalls zur Exploration sinnvoll.

Die Beobachtung kann u. a. nach der Beteiligung der beobachtenden Personen (teilnehmende vs. nicht-teilnehmende Beobachtung), dem Grad der Standardisierung des Beobachtungssettings (Labor- vs. Feldbeobachtung) sowie der Art der Protokollierung variieren (unvermittelte vs. videobasierte Beobachtung; letztere liefert gute Möglichkeiten gerade für die Unterrichtsbeobachtung). Das zu beobachtende Verhalten kann in verschiedenen Einheiten analysiert werden (Zeit- oder Ereignisstichproben) und es können verschiedene Arten von Beobachtungs- bzw. Protokollsystemen zum Einsatz kommen (Zeichen-, Kategorien- und Ratingsysteme; ▶ Kap. 24).

Im Forschungsprozess sollte möglichst auf etablierte Beobachtungsinstrumente zurückgegriffen werden. Ein Beispiel ist das „Beobachtungssystem zur Analyse aggressiven Verhaltens im schulischen Setting“ (BASYS; Wettstein 2008), mit dem verschiedene Formen aggressiven Verhaltens differenziert beobachtet werden können. Ein weiteres Beispiel ist das Unterrichtsbeobachtungssystem „Einblicknahme in die Lehr- und Lernsituation“ (ELL; Helmke 2010). Damit lassen sich auf der Sichtebene des Unterrichts u. a. die Sozialform, der Einsatz neuer Medien sowie die Realisierung offener Unterrichtsformen mit Hilfe von Kategorien registrieren. Auf der Tiefenebene können verschiedene Dimensionen der Unterrichtsqualität anhand von Schätzskalen beobachtet werden (▶ Kap. 18). Müssen Beobachtungsinstrumente neu konstruiert werden, erfordert dies einen aufwändigen Prozess (Seidel & Prenzel 2010).

Ob etabliertes oder neu-konstruiertes Instrument: In jedem Fall ist die Übereinstimmung der Beobachtungen über mehrere beobachtende Personen hinweg zu prüfen, da die Sicherstellung von Objektivität und Reliabilität anspruchsvoll ist. Neben der präzisen und konkreten Definition von Beobachtungskategorien ist dazu auch ein ausführliches Beobachtertraining, der Einsatz von mehreren Beobachterinnen und Beobachtern sowie die Beobachtung von nicht zu kurzen Zeiträumen nötig (z. B. Praetorius, Lenske & Helmke 2012; Praetorius, Pauli, Reusser, Rakoczy & Klieme 2014). Der hohe Aufwand wird durch die Vorteile von Beobachtungsverfahren aufgewogen: Sie sind weniger anfällig für Selbstdarstellungstendenzen und unabhängig davon, ob Probandinnen und Probanden Auskunft geben können oder wollen.

#### ■ Fragebögen

Fragebögen können zur Erfassung einer Vielzahl psychischer Merkmale eingesetzt werden. Motivationale Konstrukte wie Interesse, das Erleben von Emotionen, Einstellungen, Persönlichkeitseigenschaften, aber auch selbstberichtetes Lernverhalten sind typische Beispiele. Daneben lassen sich mit ihnen Fremdbeurteilungen sowie Umwelt- und Situationswahrnehmungen erfassen (z. B. Schülerwahrnehmungen des Lehrerverhaltens, Unterrichts- oder Klassenklimas).

In der Psychologie versteht man unter Fragebögen schriftliche Befragungen, die hoch strukturiert und standardisiert sind: Fragen werden in vorab festgelegter Reihenfolge vorgegeben und die Antwortmöglichkeiten sind ebenfalls vollständig oder weitgehend festgelegt (Eid et al. 2015). Wenn sie für individualdiagnostische Zwecke entwickelt und normiert werden, werden sie auch als psychometrische Persönlichkeitstests bezeichnet (im Gegensatz zu Leistungstests gibt es keine richtigen oder falschen Antworten; ▶ Kap. 24).

Vorteile von Fragebogenverfahren sind ihre Ökonomie (Möglichkeiten zur Gruppentestung und computer-/internetgestützten Durchführung, geringer Schulungsaufwand für Testleiterinnen und Testleiter), ihre hochgradige Objektivität, die durch den hohen Standardisierungsgrad erreicht wird, ihre durch bestimmte Konstruktionsprinzipien relativ einfach erzielbare Reliabilität und Validität sowie die gute Vergleichbarkeit der Messungen über verschiedene Messzeitpunkte, Situationen und Personengruppen hinweg („Messinvarianz"). Fragebogenverfahren setzen voraus, dass die Befragten zutreffend Auskunft geben können und wollen – hier liegen potenzielle Grenzen und Nachteile dieser Erhebungsmethode: Psychische Merkmale, die der Introspektion nicht gut zugänglich sind, können nicht valide erfasst werden (z. B. Arbeitsgedächtniskapazität). Zudem bergen Selbstdarstellungstendenzen wie sozial erwünschtes Antwortverhalten und andere Antwortstile, wie die Tendenz zu mittleren Urteilen, Gefahren für die Validität, denen aber durch eine sorgfältige Fragebogenkonstruktion entgegnet werden kann (z. B. Zusicherung von Anonymität, geeignete Instruktionen, Einsatz von Kontrollskalen, Wahl geeigneter Antwortkategorien). Schließlich sind die geringe Flexibilität und thematische Offenheit von Fragebogenverfahren als Nachteile anzuführen.

In Fragebögen können unterschiedliche Varianten von sog. Items (Fragen oder Aussagen, auf die geantwortet werden soll) sowie unterschiedliche Antwortformate zum Einsatz kommen. Abb. 27.3 illustriert einige typische Varianten, die einen Eindruck vom Spektrum, aber auch des hohen Aufwands seriös konstruierter Fragebögen geben sollen (für umfassende Darstellungen zur Fragebogenkonstruktion vgl. Döring & Bortz 2016; Moosbrugger & Kelava 2011; Mummendey & Grau 2014).

Items können als Fragen (Skalen 3 und 4 in Abb. 27.3) oder als Aussagen (Skalen 5 und 6) formuliert sein. Häufig werden sie in einen Itemstamm, der für alle Items identisch ist und möglichst kurze Itemendungen aufgeteilt (Skala 5). Items können positiv oder negativ im Sinne des zu erfassenden Konstrukts formuliert sein (z. B. Skala 6: Items 1, 3 und 5 vs. Items 2 und 4). Antwortformate können offen (Item 1) oder geschlossen sein (alle anderen Items). Geschlossene Antwortformate können entweder zwei Antwortalternativen wie ja/nein vorsehen (Item 2) oder mehrere Optionen vorgeben; im ersten Fall spricht man von dichotomen, im zweiten Fall von polytomen Items. Die Antwortoptionen können sich qualitativ voneinander unterscheiden („kategoriale Variablen", Item 2, Skalen 3 und 4) und dabei geordnet sein (z. B. Skala 4) oder nicht (Item 2, Skala 3). Sehr häufig kommen abgestufte, mehrstufige Antwortskalen zum Einsatz, die eine differenzierte Erfassung von kontinuierlich ausgeprägten Merkmalen ermöglichen. Sie werden als Likert-Skalen bezeichnet (Skalen 5 und 6). Gelegentlich werden auch bipolare Antwortskalen genutzt, bei denen zwei verschiedene, gegensätzliche Merkmale an den Enden der Antwortskalen dargeboten werden (Skala 7).

Die Erfassung psychologischer Konstrukte erfolgt typischerweise nach dem Prinzip der Messwiederholung: Mehrere Items einer Skala thematisieren das Konstrukt in all seinen Facetten. Durch die unterschiedlichen Formulierungen werden itemspezifische Fehlereinflüsse „herausgemittelt" und eine hohe Reliabilität gewährleistet. Beispiele sind die Skalen 5 und 6, die mit 6 bzw. 5 Items jeweils einen einzigen Faktor erfassen.

### Interview

Interviews sind mündliche Befragungen, die in unterschiedlichen Strukturierungs- und Standardisierungsgraden realisiert werden können. Gering strukturierte und standardisierte Interviews machen wenige Vorgaben zu Durchführung und Auswertung und ermöglichen dadurch eine größere Flexibilität und thematische Offenheit als etwa schriftliche Fragebogenverfahren. Dies ist sinnvoll zur Exploration neuer Forschungsfragen, aber nicht für das Testen von Hypothesen, da hier in der Regel keine ausreichend hohe Testgüte gegeben ist. Stärker strukturierte und standardisierte Interviews geben mit Hilfe eines Leitfadens genau vor, welche Fragen in welcher Reihenfolge gestellt werden (▶ Kap. 24). Zudem kann auch eine begrenzte Anzahl an Antwortmöglichkeiten vorgegeben werden; falls nicht, sind die offenen Antworten mit Hilfe von Kategoriensystemen zu kodieren. Interviewleitfäden erhöhen die Objektivität (die durch eine hinreichende Beurteilerübereinstimmung nachgewiesen werden muss), reduzieren aber den Vorteil der thematischen Offenheit. Da Interviews im Vergleich zu Fragebogenverfahren eine geringere Ökonomie aufweisen (oft nur Einzeltestung möglich, Interviewerschulung nötig), ist genau abzuwägen, wann ihr Einsatz sinnvoll ist.

Beispielsweise nutzten Engelschalk, Steuer und Dresel (2015) Leitfadeninterviews mit anschließender Kategorisierung der Antworten, um zu erfassen, welche Strategien Lernende zur Regulation ihrer Motivation bei sechs unterschiedlichen Motivationsproblemen einsetzen. Die Wahl fiel hier auch deshalb auf die Methode der Leitfadeninterviews, da die Probandinnen und Probanden wiederholt ausführliche Antworten geben sollten – was mit schriftlichen Fragenbögen nur schwer sichergestellt werden kann.

### Leistungstests

Leistungstests sind standardisierte Verfahren, bei denen eine größere Menge an Aufgaben zu bearbeiten ist, für die es richtige und falsche Antworten gibt – bei den meisten anderen Arten an Erhebungsverfahren ist dies nicht der Fall. Leistungstests erfassen damit, wie gut eine Person etwas tut oder tun kann, während die bisher besprochenen Verfahren erfassen, was eine Person tatsächlich tut, warum sie dies tut und was sie dabei erlebt (Eid et al. 2015, S. 64). Leistungstests existieren für verschiedene Leistungsbereiche (ausführliche Erläuterungen ▶ Kap. 24):

- Allgemeine Leistungstests erfassen Konzentration und Aufmerksamkeit als allgemeine Leistungsvoraussetzungen.
- Intelligenztests erfassen je nach zugrundeliegender theoretischer Konzeption die allgemeine Intelligenz oder bestimmte Intelligenzkomponenten (▶ Kap. 8) und sind da-

1) Dein Alter: _____ Jahre

2) Dein Geschlecht: ☐ weiblich ☐ männlich

3) In welchem Land wurdest du geboren? In welchem Land wurden deine Eltern geboren?

| Du | Mutter | Vater | |
|---|---|---|---|
| ☐ | ☐ | ☐ | in Deutschland |
| ☐ | ☐ | ☐ | in Griechenland |
| ☐ | ☐ | ☐ | in Italien |
| ☐ | ☐ | ☐ | im ehemaligen Jugoslawien (Bosnien-Herzegowina, Kroatien, Mazedonien, Montenegro, Serbien und Slowenien) |
| ☐ | ☐ | ☐ | in Polen |
| ☐ | ☐ | ☐ | in Russland, Kasachstan oder einer anderen ehemaligen Sowjetrepublik |
| ☐ | ☐ | ☐ | in der Türkei |
| ☐ | ☐ | ☐ | in einem anderen Land |

4) Welches ist der höchste Schulabschluss deiner Eltern?

| Mutter | Vater | |
|---|---|---|
| ☐ | ☐ | Abitur |
| ☐ | ☐ | Realschulabschluss |
| ☐ | ☐ | Hauptschulabschluss |
| ☐ | ☐ | Kein Schulabschluss |

5) Denke bitte an deine Deutschlehrerin/deinen Deutschlehrer.

| Unsere Deutschlehrerin/ Unser Deutschlehrer ... | nie | selten | gele-gentlich | oft | sehr oft |
|---|---|---|---|---|---|
| ... verwendet Humor im Zusammenhang des Unterrichtsstoffs. | ☐ | ☐ | ☐ | ☐ | ☐ |
| ... verwendet witzige Dinge zur Veranschaulichung oder als Beispiel. | ☐ | ☐ | ☐ | ☐ | ☐ |
| ... erzählt uns Witze, die sich auf den Unterrichtsinhalt beziehen. | ☐ | ☐ | ☐ | ☐ | ☐ |
| ... erzählt uns witzige Geschichten, die zum Unterricht passen . | ☐ | ☐ | ☐ | ☐ | ☐ |
| ... bringt uns den Stoff auf humorvolle Art bei. | ☐ | ☐ | ☐ | ☐ | ☐ |
| ... verwendet lustige Beispiele im Unterricht. | ☐ | ☐ | ☐ | ☐ | ☐ |

6) Wie ist das bei dir?

Im Fach Deutsch werde ich in Zukunft bestimmt gute Leistungen bringen.

| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
|---|---|---|---|---|---|
| stimmt gar nicht | stimmt nicht | stimmt eher nicht | stimmt eher | stimmt | stimmt völlig |

Ich werde im Fach Deutsch in Zukunft bestimmt schlechter abschneiden als die meisten anderen.

| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
|---|---|---|---|---|---|
| stimmt gar nicht | stimmt nicht | stimmt eher nicht | stimmt eher | stimmt | stimmt völlig |

Bestimmt werde ich im Fach Deutsch in Zukunft viele neue Dinge lernen.

| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
|---|---|---|---|---|---|
| stimmt gar nicht | stimmt nicht | stimmt eher nicht | stimmt eher | stimmt | stimmt völlig |

Bestimmt werde ich im Fach Deutsch in Zukunft schlechte Noten bekommen.

| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
|---|---|---|---|---|---|
| stimmt gar nicht | stimmt nicht | stimmt eher nicht | stimmt eher | stimmt | stimmt völlig |

Im Fach Deutsch werde ich in Zukunft bestimmt immer mehr können.

| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
|---|---|---|---|---|---|
| stimmt gar nicht | stimmt nicht | stimmt eher nicht | stimmt eher | stimmt | stimmt völlig |

7) Für den Unterricht im Fach Deutsch halte ich die Unterschiedlichkeit der Schülerinnen und Schüler in Bezug auf ihren Migrationshintergrund für:

| | | |
|---|---|---|
| schlecht | ○ ○ ○ ○ ○ ○ | gut |
| unangenehm | ○ ○ ○ ○ ○ ○ | angenehm |
| förderlich für das Lernen der Schüler(innen) | ○ ○ ○ ○ ○ ○ | hinderlich für das Lernen der Schüler(innen) |
| nicht belastend für die Lehrkraft | ○ ○ ○ ○ ○ ○ | belastend für die Lehrkraft |

**Abb. 27.3** Einige Varianten von Item- und Antwortformaten in Fragebögen. Item 1 und Item 2 leisten eine einfache Erfassung von Lebensalter und biologischem Geschlecht. Skala 3 ist eine typische Operationalisierung des Migrationshintergrunds von Schülerinnen und Schülern (z. B. Kunter et al. 2003). Skala 4 illustriert, stark vereinfacht, die Erfassung der Schul- und Berufsbildung der Eltern (vgl. Kunter et al. 2003). Skala 5 misst Schülerwahrnehmungen des Lehrkrafthumors (Faktor „Lerngegenstandsbezogener Humor", der mit sechs Items erfasst wird; Bieg & Dresel 2016). Skala 6 erfasst die Erfolgserwartung von Schülerinnen und Schülern (5-Item-Skala von Dresel, Schober & Ziegler 2005). Skala 7 bildet Beispielitems von Skalen zur Erfassung der Einstellungen von Lehrpersonen gegenüber Heterogenität ab (Lehmann-Grube et al. 2017)

mit erklärungsstarke Prädiktoren für gelingende Lernprozesse sowie für Schul-, Ausbildungs- und Berufserfolg insgesamt.

- Entwicklungstests erfassen den Entwicklungsstand von Kindern und Jugendlichen übergreifend oder in Bezug auf einzelne Funktionsbereiche (z. B. Psychomotorik, Sprache, Kognition, Motivation).
- Standardisierte Schulleistungstests dienen dazu, Erträge von Lehr-Lernprozessen zu erfassen – z. B. in Abhängigkeit von bestimmten Unterrichts- oder Fördermaßnahmen (▶ Kap. 25). Ihnen kommt eine tragende Rolle in der schulbezogenen Forschung zu. Steht konzeptuelles Wissen im Fokus, werden diese Tests häufig Wissenstests genannt; werden stattdessen Fertigkeiten und die Umsetzung erworbenen Wissens in bestimmten Anwendungs- oder Problemlösesituationen erfasst, werden sie oft als Kompetenztests bezeichnet. Nicht selten sind in Forschungsstudien standardisierte Schulleistungstests nicht verfügbar und es finden Zensuren, also Lehrkraftbewertungen von Schülerleistungen Verwendung. Aufgrund

der oftmals nicht befriedigenden Testgüte von Zensuren (▶ Kap. 25) hat dies jedoch in den meisten Fällen deutliche Einschränkungen in der Aussagekraft dieser Studien zur Folge.

- Schulfähigkeitstests und Tests zur Erfassung von Vorläuferfähigkeiten (z. B. phonologische Bewusstheit als Vorläuferfähigkeit des Schriftspracherwerbs; Jansen, Mannhaupt, Marx & Skowronek 2002) können im Kontext der Einschulung genutzt werden.

Die für individualdiagnostische Zwecke entwickelten Leistungstests finden häufig auch in Forschungsstudien Verwendung – insbesondere auch aufgrund ihrer nachweislich hohen Testgüte und Ökonomie (oft sind Gruppentestungen möglich). Oft werden Leistungstests für empirische Studien aber auch eigens entwickelt. Da die Verfügbarkeit von Normwerten im Forschungsprozess weniger von Bedeutung ist, wird hierbei meist auf eine aufwändige Normierung verzichtet. Weil es nicht einfach ist, einen Test von hinreichender Testgüte zu konstruieren, sollte der Nachweis geführt werden, dass Reliabilität und Validität gegeben sind. Aufgrund der vollständigen Standardisierung von Durchführung, Auswertung und Interpretation kann meist perfekte Objektivität angenommen werden.

### Ambulantes Assessment

Viele Erhebungsverfahren erfassen psychologische Daten punktuell und in unnatürlichen Erhebungssituationen. Ambulantes Assessment versucht dies zu überwinden und den kontinuierlichen Strom von Erleben und Verhalten in authentischen Alltagssituationen zu erfassen – etwa um Lern- und Entwicklungsprozesse ökologisch valide abzubilden (Fahrenberg 2010). Dies wird oft mit technischer Unterstützung realisiert, z. B. mit internetbasierten Erhebungssystemen oder Smartphones. Ein wichtiger Ansatz des ambulanten Assessments ist die *Experience-Sampling*-Methode. Probandinnen und Probanden werden hier zu vielen, zufällig ausgewählten Zeitpunkten durch einen Signalgeber aufgefordert, Auskunft über ihr momentanes Erleben oder Verhalten zu geben. Beispielsweise haben Nett, Götz und Hall (2011) so untersucht, wie Schülerinnen und Schüler die zeitlich variable Emotion Langeweile im Unterricht regulieren. Ein weiterer Ansatz ist die Tagebuchbuchmethode, bei der Individuen gebeten werden, regelmäßig zu festen Zeitpunkten (z. B. täglich abends über einen Zeitraum von zwei Wochen) ihr Erleben und Verhalten zu berichten. Zum Beispiel untersuchten Schmitz und Wiese (2006) damit selbstreguliertes Lernen über fünf Wochen hinweg und wiesen Trainingseffekte auf der Prozessebene nach.

Nachteilig am ambulanten Assessment sind ein oft hoher Aufwand für die Versuchspersonen und damit einhergehende Selbstselektionseffekte sowie mögliche Urteilsverzerrungen, die durch eine erhöhte Selbstaufmerksamkeit entstehen können.

### Soziometrie

Soziometrische Verfahren dienen dazu, interpersonelle Beziehungen in Schulklassen und anderen Gruppen zu erfassen (Dollase & Koch 2010). Meist werden dazu positive Wahlpräferenzen erfragt (z. B. „Neben wem möchtest Du sitzen?“, „Mit wem möchtest Du in einer Lerngruppe sein?“). Um unterscheiden zu können, ob ein Gruppenmitglied abgelehnt oder ignoriert wird, wenn es bei Positivwahlen nicht genannt wird, werden daneben häufig auch negative Wahlen erfragt (z. B. „Neben wem möchtest Du nicht sitzen?“). Die Wahlen liefern zunächst Daten auf Ebene der einzelnen Gruppenmitglieder, insbesondere den positiven und negativen Status (z. B. relative Häufigkeit von An- und Abwahlen), sowie die Mitgliedschaft in Statusgruppen (z. B. „Stars“, „Außenseiter“, „Abgelehnte“). Diese Daten sind weniger anfällig für Selbstdarstellungstendenzen als Selbstberichte der eigenen Beliebtheit. Darüber hinaus können Freundschaftsbeziehungen und Cliquen innerhalb der Gruppe identifiziert und Daten auf Ebene der gesamten Gruppe generiert werden (z. B. Kohäsion, Ablehnungsbereitschaft). Soziometrische Daten sind – trotz ihrer relativ einfachen Gewinnbarkeit – sehr sensibel und sollten mit großer Sorgfalt genutzt werden. Beispielsweise sollten die Ergebnisse keinesfalls auf Gruppenebene zurückgemeldet werden (▶ Kap. 22).

### Psychophysiologische Verfahren

Psychophysiologische Verfahren machen sich zu Nutze, dass psychische Prozesse oft mit biologischen Vorgängen einhergehen. So lassen Aktivitätsmuster in bestimmten Gehirnarealen, die mittels Elektroenzephalogramm (EEG) oder bildgebenden Verfahren wie der funktionellen Magnetresonanztomografie (fMRT) erfasst werden können, u. a. auf die emotionale Bedeutung von Reizen schließen oder ermöglichen, die beim Lernen beteiligten anatomischen Strukturen zu rekonstruieren (▶ Kap. 5).

Neben diesen relativ aufwändigen Verfahren lassen sich einfacher zu erfassende Indikatoren wie die Herzschlagrate, die Hautleitfähigkeit, die Spannungsmuster bestimmter Muskeln oder die Konzentration bestimmter Hormone (z. B. Cortisol, Adrenalin, Oxytocin) nutzen, um psychische Zustände wie Stress oder spezifisches emotionales Erleben zu erfassen. Beispielsweise untersuchten Tulis und Dresel (2018) emotionale Reaktionen nach Fehlern bei einer Lernaufgabe. Sie analysierten dazu in 3-Sekunden-Intervallen nach einer Fehlerrückmeldung die elektrische Aktivität zweier Gesichtsmuskeln (über den Augenbrauen als Indikator negativer Emotionen und entlang der Wangen als Indikator positiver Emotionen) sowie Herzschlagrate und Hautleitfähigkeit (als Indikatoren für aktivierende versus desaktivierende Emotionen).

Ein weiteres Verfahren, das zunehmend Bedeutung in der Lehr-Lernforschung erlangt, ist die Analyse von Blickbewegungen, Pupillenweite und Lidschlagrate (*Eye-Tracking*). Damit kann die spezifische Aufmerksamkeit auf bestimmte Informationen sowie die dabei entstehende kognitive Belastung erfasst werden. Beispielsweise untersuchte Tobisch (2017) mit Hilfe dieser Verfahren kognitive Prozesse bei der Leistungsbeurteilung von Schülerinnen und Schülern unterschiedlicher Herkunft (vgl. Tobisch & Dresel 2017).

Vorteile psychophysiologischer Verfahren liegen u. a. in ihrem hohen zeitlichen Auflösungsgrad, der feinmaschige Verlaufsanalysen ermöglicht, und in ihrer geringen Ver-

fälschbarkeit. Andererseits sind sie vergleichsweise aufwändig und die Untersuchungssituation ist oftmals recht künstlich, was die Validität einschränken kann.

#### Reaktionszeitgestützte Verfahren

Reaktionszeitgestützte Verfahren können genutzt werden, um insbesondere Einstellungen, Stereotype und Persönlichkeitsmerkmale implizit zu erfassen – d. h. ohne dass vorausgesetzt wird, dass die Probandinnen und Probanden darüber zutreffend Auskunft geben können oder wollen. Sie können den Zweck der Messungen nicht leicht durchschauen und entsprechend weniger leicht verfälschen.

Ein typisches Beispiel ist der Implizite Assoziationstest (IAT; Greenwald, McGhee & Schwartz 1998), mit dem z. B. implizite Stereotype als Stärke der Assoziation von Objekten (z. B. Ethnien, Geschlechter) und Attributen (z. B. Wertigkeit, Eigenschaften) erfasst werden können – beispielsweise bei Lehrpersonen (z. B. Glock & Kleen 2017). Er umfasst eine Reihe von computerbasierten Diskriminationsaufgaben, bei der möglichst schnell entschieden werden soll, welcher von zwei Antwortkategorien (z. B. deutscher vs. türkischer Name; positives vs. negatives Attribut) ein Wort zugeordnet werden kann. Dabei ist die zentrale Annahme, dass implizite Assoziationen (z. B. positive Assoziationen zu türkischen Namen) eine schnelle und fehlerfreie Bearbeitung begünstigen.

#### Projektive Verfahren

Projektive Verfahren, die auch als Persönlichkeitsentfaltungsverfahren bezeichnet werden (▶ Kap. 23), wurden zur Erfassung unbewusster, verdrängter oder latenter Eigenschaften entwickelt. Sie basieren auf der Annahme, dass innere Empfindungen der Außenwelt zugeschrieben werden können – insbesondere wenn es sich um unerwünschte oder selbstbedrohliche Empfindungen handelt, die aus tiefenpsychologischer Perspektive dadurch unbewusst abgewehrt werden (Konzept der Projektion; Freud 1936). Viele projektive Verfahren präsentieren Szenen, die von anderen Personen handeln und uneindeutig oder offen im Ergebnisausgang sind. Die Versuchspersonen sollen sich in diese Personen hineinversetzen und beschreiben, was in ihnen vorgeht oder wie sie handeln werden. Ein Beispiel ist der Rosenzweig *Picture Frustration Test* (deutsche Version von Rauchfleisch, Battegay & Rosenzweig 1979), der insgesamt 24 gezeichnete Szenen enthält, bei denen eine Person durch eine verbale Äußerung eine andere Person frustriert. Erfragt wird die Antwort der frustrierten Person, die den Probandinnen und Probanden spontan einfällt.

Die großen Erwartungen, die in projektive Verfahren gesetzt wurden, erfüllten diese letztlich nicht (Rauchfleisch 2006). Die meisten Verfahren entziehen sich aufgrund ihrer nicht-quantitativen Testkonzeption einer Überprüfung ihrer Testgüte. Eine Ausnahme ist das genannte Beispiel, das eine befriedigende Testgüte erzielt; daneben lässt sich auf semiprojektive Verfahren wie das Multi-Motiv-Gitter (Schmalt, Sokolowski & Langens 2000) verweisen, die szenische Stimuli mit gebundenen Antwortformaten kombinieren. Abgesehen von diesen Ausnahmen nehmen projektive Verfahren deshalb im Forschungsprozess allenfalls eine hypothesengenerierende Funktion im Sinne von Explorationshilfen ein – ihre Ergebnisse sollten jedenfalls mit Vorsicht interpretiert werden.

Neben den genannten Erhebungsmethoden, die messmethodische Grundtypen der Psychologie reflektieren, existiert noch eine größere Anzahl weiterer Messverfahren (Döring & Bortz 2016; Petermann & Eid 2006). Dazu zählen u. a.

- Analyse von Verhaltensspuren (z. B. Notizen, spontane Zeichnungen oder Markierungen von Textstellen als Verhaltensspuren des selbstregulierten Lernens; Hadwin, Nesbit, Jamieson-Noel, Code & Winne 2007),
- Methode des lauten Denkens zur Erfassung kognitiver Prozesse (Ericsson & Simon 1993),
- Verfahren zur Analyse un- oder halbstrukturiert erfasster Texte (z. B. Essays, Chat-Protokolle) auf quantitativer oder qualitativer Ebene (Mayring 2010; Mehl 2006),
- apparative Verfahren (z. B. zur Erfassung psychomotorischer Leistungen oder des kognitiven Entwicklungsstands) und
- Auswertung von Archivdaten (z. B. archivierte Zensuren).

### 27.3.4 Skalenniveaus

Psychologische Erhebungsverfahren liefern Daten mit unterschiedlichen Eigenschaften, was Konsequenzen für ihre Analyse und Interpretation hat. Beispielsweise weisen Angaben dazu, (a) welche der drei angebotenen Wahlkurse von wie vielen Schülerinnen und Schülern einer Schulklasse belegt werden, (b) wie viele von ihnen bei einem sportlichen Wettbewerb eine Ehrenurkunde, eine Siegerurkunde oder eine Teilnahmeurkunde erhielten und (c) wie viele Rechtschreibfehler sie in einem Deutschdiktat gemacht haben, jeweils andere Eigenschaften auf. Im ersten Fall sind die Daten nicht geordnet; hier kommt ausschließlich eine Analyse der Häufigkeiten in Betracht. Im zweiten Fall weisen die Daten bereits eine sinnvolle Ordnung auf; hier werden Analysen von Ranginformationen möglich (z. B. „56 % der Schülerinnen und Schüler erhielten mindestens eine Siegerurkunde“). Im dritten Fall sind die Abstände zwischen den einzelnen Ausprägungen definiert und konstant; erst hier macht die Berechnung des arithmetischen Mittels Sinn.

Die gewonnenen Daten eröffnen also aufgrund der Eigenschaften der zugrundeliegenden Skalen mehr oder weniger Möglichkeiten zur Datenanalyse und damit den Weg zu unterschiedlichen Aussagen. Dies wird in sog. Skalenniveaus zusammengefasst, die stufenweise mehr Eigenschaften erfüllen und mehr Möglichkeiten der Analyse verfügbar machen. Unterschieden werden insbesondere drei Skalenniveaus:

#### Nominalskalierte Variablen

Eine nominalskalierte Variable bildet lediglich die Gleichheit (bzw. Äquivalenz) und Ungleichheit der Ausprägungen von Merkmalen ab: „Eine Nominalskala ordnet den Objekten eines empirischen Relativs Zahlen zu, die so geartet sind, dass Objekte mit gleicher Merkmalsausprägung

**Tabelle 27.2** Wichtige Untersuchungsdesigns

| Untersuchungsdesign | Herangehensweise und Potential |
|---|---|
| Experiment | Systematische Beeinflussung einer oder mehrerer unabhängiger Variablen (UV) in kontrolliert (mittels zufälliger Zuweisung) zusammengesetzten Untersuchungsgruppen. Untersucht werden die Auswirkungen der UV auf eine oder mehrere abhängige Variablen (AV) unter maximaler Kontrolle weiterer Einflussfaktoren. Ermöglicht den Nachweis kausaler Zusammenhänge |
| Quasiexperiment | Systematische Beeinflussung einer oder mehrerer UV in natürlich vorgefundener Untersuchungsgruppen (z. B. Schulklassen), bei denen keine Zufallsaufteilung möglich ist. Wie beim Experiment werden die Auswirkungen der UV auf eine oder mehrere AV untersucht. Bestehende Unterschiedlichkeiten zwischen den Gruppen werden berücksichtigt und kontrolliert. Ermöglicht den Nachweis starker, aber nicht zweifelsfreier Aussagen über kausale Zusammenhänge |
| Querschnittuntersuchung | Einmalige Messung von Merkmalen ohne systematische Beeinflussung von Variablen. Ermöglicht den Nachweis von Zusammenhängen zwischen Merkmalen |
| Längsschnittuntersuchung | Wiederholte Messung von Merkmalen ohne systematische Beeinflussung von Variablen. Ermöglicht den Nachweis von Veränderungen der Merkmale über die Zeit |
| Metaanalyse | Zusammenführung der Ergebnisse mehrerer bereits vorliegender Studien zu einem Forschungsthema. Zielt auf die zusammenfassende Einschätzung von Effekten ab |

gleiche Zahlen und Objekte mit verschiedener Merkmalsausprägung verschiedene Zahlen erhalten" (Döring & Bortz 2016, S. 238). Nominalskalierte Variablen sind beispielsweise Themen-, Kurs- und Fachwahlen, biologisches Geschlecht und Migrationshintergrund. In Abb. 27.3 sind Item 2 und Skala 3 auf Nominalskalenniveau.

### Ordinalskalierte Variablen

Gegenüber nominalskalierten bilden ordinalskalierte Variablen zusätzlich die Ordnung der Merkmalsausprägungen ab – etwa nach deren Intensität, Wertigkeit oder Güte: „Eine Ordinalskala ordnet den Objekten eines empirischen Relativs Zahlen (Rangzahlen) zu, die so geartet sind, dass von jeweils zwei Objekten das dominierende Objekt die größere Zahl erhält. Bei Äquivalenz sind die Zahlen identisch" (Döring & Bortz 2016, S. 240). Typische ordinalskalierte Variablen sind Auszeichnungen verschiedenen Grades, Schul- oder Berufsabschlüsse, der sozioökonomische Status und Schulnoten. In Abb. 27.3 ist Skala 4 auf Ordinalskalenniveau.

### Intervallskalierte Variablen

Die Intervallskala bildet zusätzlich die Distanzen zwischen den Merkmalsausprägungen ab – damit reflektiert sie anders als die Ordinalskala die Größe der Unterschiede zwischen ihnen: „Eine Intervallskala ordnet den Objekten eines empirischen Relativs Zahlen zu, die so geartet sind, dass die Rangordnung der Zahlendifferenzen zwischen je zwei Objekten der Rangordnung der Merkmalsunterschiede zwischen je zwei Objekten entspricht. Die Intervallskala zeichnet sich durch Äquidistanz bzw. Gleichabständigkeit der Messwerte aus" (Döring & Bortz 2016, S. 244). Die meisten psychischen Merkmale können und sollten mindestens per Intervallskala gemessen werden. In Befragungsverfahren und Beobachtungssystemen bieten sich dazu die erwähnten Likert-Skalen an. In Abb. 27.3 sind die Skalen 5 bis 7 auf Intervallskalenniveau.

Über die drei genannten Skalenniveaus hinaus werden häufig noch verhältnisskalierte Variablen (die zusätzlich einen absoluten Nullpunkt besitzen, z. B. Zeit, Fehlerzahl) definiert. Da sich für viele psychische Merkmale ein Nullpunkt nur schwer definieren lässt, spielt sie keine entscheidende Rolle. Intervall- und verhältnisskalierte Variablen werden deshalb oft als metrische oder kardinalskalierte Variablen zusammengefasst. Welches Skalenniveau bei einem bestimmten Erhebungsverfahren erreicht wird, hängt von dem zu messenden Merkmal sowie dem eingesetzten Messverfahren ab. Für eine gute Datenqualität sollte das für ein gegebenes Merkmal höchstmögliche Skalenniveau realisiert werden.

## 27.4 Untersuchungsdesigns

Um eine Forschungsfrage unter gegebenen Rahmenbedingungen möglichst eindeutig beantworten zu können, muss ein geeignetes Vorgehen gewählt werden. Häufig anzutreffende Grundmuster der psychologischen Forschung, sogenannte Forschungs- oder Untersuchungsdesigns, sind in Tab. 27.2 zusammengestellt.

Die einzelnen Untersuchungsdesigns stehen für sehr unterschiedliche Herangehensweisen. Mit der Wahl eines Designs sind auch Festlegungen verbunden, welche Aussagen die Ergebnisse einer Studie letztlich erlauben. Beispielsweise ermöglicht es das Experiment, Kausalhypothesen zu testen, was in der Querschnittuntersuchung nicht möglich ist. Welches Untersuchungsdesign zu wählen ist, hängt somit wesentlich von der Hypothese ab, die geprüft werden soll. Soll untersucht werden, ob ursächliche Wirkungen einer Variablen (z. B. bestimmtes Lehrerhandeln) auf eine andere Variable (z. B. Schülerverhalten) bestehen, so wäre es naheliegend, ein Experiment oder Quasiexperiment zu realisieren. Zielen die Forschungshypothesen in erster Linie auf die Aufklärung von Zusammenhängen zwischen einzelnen Merkmalen ab, sind Korrelationsstudien das Vorgehen der Wahl. Dazu zählen Querschnitt- und Längsschnittuntersuchungen, wobei letztere den Faktor Zeit systematisch einbeziehen und dadurch

für die Untersuchung von Veränderungshypothesen geeignet sind und durchaus Hinweise zur Kausalität liefern können. Werden zur Prüfung einer Forschungsfrage keine eigenen Daten erhoben, sondern eine größere Zahl bereits vorliegender Studienergebnisse zusammengetragen und integrierend analysiert, handelt es sich um eine Metaanalyse. Im Folgenden werden die Untersuchungsdesigns ausführlich erläutert.

### 27.4.1 Experiment und Quasiexperiment

Um Hypothesen zu kausalen Wirkungen oder zur Wirksamkeit von Maßnahmen prüfen zu können, werden experimentelle Forschungsdesigns benötigt. Dazu zählen Experiment und Quasiexperiment.

Nach Klauer (2006) ist ein Experiment ein „planmäßig ausgelöster und wiederholbarer Vorgang, bei dem beobachtet wird, in welcher Weise sich unter Konstanthaltung anderer Bedingungen mindestens eine abhängige Variable ändert, nachdem mindestens eine unabhängige Variable geändert worden ist“ (S. 77). Die Änderung der unabhängigen Variablen erfolgt gezielt, was als experimentelle Manipulation bezeichnet wird. Die Zuweisung der teilnehmenden Personen zu den verschiedenen Stufen der unabhängigen Variablen erfolgt randomisiert, d. h. zufällig. Soll beispielsweise untersucht werden, ob Ergänzungen des herkömmlichen schriftlichen Feedbacks auf Hefteinträge mit bunten Stickern ursächlich zu höherer Lernmotivation am Ende des ersten Grundschuljahres führt, wäre folgendes fiktive Experiment das ideale Design: Zwei Untersuchungsgruppen würden sich ausschließlich darin unterscheiden, dass die unabhängige Variable (Feedback mit oder ohne Sticker) variiert. Welches Kind in welche Gruppe kommt, würde per Zufall entschieden. Am Ende des Schuljahres ließe sich mit Hilfe eines Motivationsfragebogens prüfen, ob die experimentelle Manipulation der unabhängigen Variablen mit Unterschieden in der abhängigen Variablen (Lernmotivation) zwischen den beiden Untersuchungsgruppen einhergeht. Ist dies der Fall, sind zwei Voraussetzungen einer Ursache-Wirkungs-Beziehung bereits erfüllt: Die vermutete Ursache geht der angenommenen Wirkung zeitlich voraus und mit der Veränderung der unabhängigen Variable geht auch eine Veränderung der abhängigen Variable einher. Von einem Kausalzusammenhang kann allerdings erst dann ausgegangen werden, wenn zusätzlich als sicher gelten kann, dass allein die Manipulation der unabhängigen Variablen zu den beobachteten Ergebnissen geführt hat (Eid et al. 2015). Lassen sich alternative Erklärungen für den gefundenen Effekt ausschließen, wird die Untersuchung als intern valide bezeichnet (▶ Definition).

> Mit interner Validität ist die Eindeutigkeit gemeint, mit der die kausale Wirkung einer unabhängigen Variable auf eine abhängige Variable belegt werden kann (Döring & Bortz 2016). Je weniger Alternativerklärungen für ein Untersuchungsergebnis denkbar sind, desto intern valider ist eine Untersuchung.

Interne Validität, die meint, dass ein Untersuchungsergebnis eindeutig erklärbar ist, sollte nicht mit dem Begriff der Validität als Gütekriterium von Erhebungsinstrumenten verwechselt werden (▶ Abschn. 27.3).

Störeinflüsse bzw. Alternativerklärungen können die interne Validität von Untersuchungen gefährden. Damit sind alle Einflüsse gemeint, die systematisch auf die abhängige Variable einwirken (Eid et al. 2015): Solche können sich direkt aus der Untersuchungssituation heraus ergeben. Im fiktiven Experiment wäre beispielsweise denkbar, dass den Schülerinnen und Schülern die unterschiedliche Behandlung auffällt. Dann wäre zu befürchten, dass sich die Gruppe ohne Sticker benachteiligt und allein aus diesem Grund weniger motiviert fühlt. Störeinflüsse können aber auch ungewollt mit der Variation der Experimentalbedingungen einhergehen. Dies wäre beispielsweise der Fall, wenn die Lehrkraft durch das Anbringen des Stickers dazu verleitet wird, unwissentlich ein stärker motivierendes schriftliches Feedback zu geben. Schließlich könnten sich Störeinflüsse auch aus den Eigenheiten der untersuchten Personen ergeben, wie es der Fall bei einer ungleichen Verteilung besonders motivierter Kinder in den unterschiedenen Untersuchungsgruppen wäre.

Im Experiment können – im Vergleich zu anderen Untersuchungsdesigns – solche Störeinflüsse weitestgehend ausgeschlossen bzw. kontrolliert werden. Störvariablen, die sich aus der Zusammensetzung der Untersuchungsgruppen ergeben, werden in einem Experiment dadurch eliminiert, dass per Zufall entschieden wird, welcher Untersuchungsgruppe die einzelnen Personen zugeteilt werden. Diese Randomisierung führt bei ausreichend großen Stichproben dazu, dass sich etwaige Unterschiede zwischen den Untersuchungsgruppen von selbst ausgleichen.

Von einem Quasiexperiment wird gesprochen, wenn natürlich vorgefundene Gruppen genutzt werden, um die unabhängige Variable zu variieren. Ein solches Vorgehen bietet sich an, wenn eine randomisierte Zuteilung der einzelnen Versuchspersonen zu Untersuchungsgruppen nicht möglich oder nicht sinnvoll ist. So wäre es wenig praktikabel, Kinder per Zufall neuen Klassen zuzuteilen. Um auch im Quasiexperiment eine möglichst gute interne Validität zu erreichen, werden etwaige Unterschiede zwischen den Untersuchungsgruppen in Bezug auf die abhängigen Variablen (z. B. Lernmotivation) und weitere relevante Variablen (z. B. Intelligenz) noch vor der experimentellen Manipulation erfasst und bei der späteren Datenanalyse berücksichtigt (statistisch kontrolliert). Abhängige Variablen werden also nicht nur nach, sondern auch bereits vor der experimentellen Manipulation gemessen. ◘ Abb. 27.4 veranschaulicht das quasiexperimentelle Vorgehen in einer einfachen Variante. Je nach Fragestellung können auch mehr als zwei Versuchsbedingungen, wie beispielsweise weitere Treatmentgruppen und eine Placebogruppe (zur Kontrolle von etwaigen Zuwendungseffekten) vorgesehen werden. Auch weitere Messzeitpunkte sind denkbar, insbesondere ein zusätzlicher *Follow-up* zur Überprüfung von Langzeiteffekten.

Trotz der aufwendigeren Datenerfassung und -analyse sowie den gewissen Abstrichen in der internen Validität, die

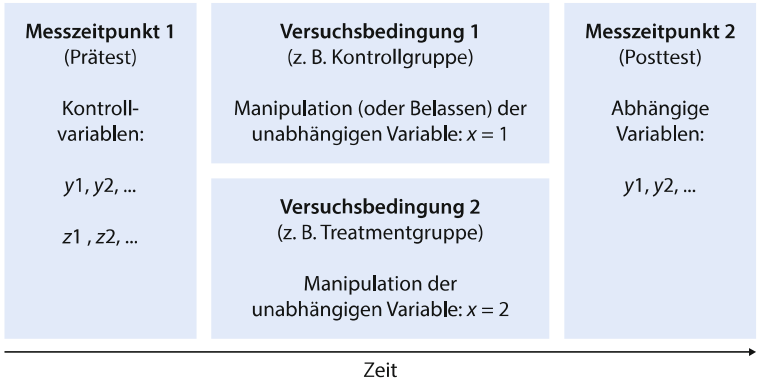



**Abb. 27.4** Einfache Variante eines quasiexperimentellen Designs

gegenüber dem klassischen Experiment in Kauf genommen werden müssen, bietet das quasiexperimentelle Design einen sinnvollen, effektiven und praktikablen Weg, Fragestellungen zu Wirkungen von Maßnahmen und kausalen Beziehungen zwischen Merkmalen im Bildungsbereich zu untersuchen. Wenn in der pädagogischen Praxis ein klassisches Experiment mit vollständiger Randomisierung aufgrund der vorgefundenen Gegebenheiten (insb. bestehende Lerngruppen und Schulklassen) nicht realisierbar ist, bietet ein Quasiexperiment die beste interne Validität und belastbarsten Ergebnisse. Das Quasiexperiment ist deshalb beispielsweise in der Unterrichtsforschung das Untersuchungsdesign der Wahl.

### 27.4.2 Querschnittuntersuchung

Das Prinzip einer Querschnittstudie besteht im Gegensatz zu experimentellen Designs darin, an den natürlicherweise bestehenden Ausprägungen von Merkmalen von Personen anzusetzen und zu untersuchen, wie diese mit anderen Variablen in Zusammenhang stehen. Alle Untersuchungsvariablen werden bei einer querschnittlichen Studie gleichzeitig erfasst. Deshalb ist die Unterscheidung zwischen unabhängigen und abhängigen Variablen hier in gewisser Weise artifiziell und auch umkehrbar. Stärke und Richtung der gefundenen Zusammenhänge werden typischerweise mit Korrelationen quantifiziert (▶ Abschn. 27.5.1). Beispielsweise fanden Leopold und Leutner (2002) in einer Querschnittuntersuchung von Schülerinnen und Schülern positive Zusammenhänge zwischen der Nutzung von Elaborationsstrategien und den Leistungen in einem Wortschatztest. Kinder, die Elaborationsstrategien seltener nutzen, schnitten also auch beim Wortschatztest mit höherer Wahrscheinlichkeit schlechter ab. Bei korrelativen Ergebnissen ist es wichtig, nicht der (menschlichen) Versuchung zu erliegen, diese in einer bestimmten Kausalrichtung zu interpretieren. So gibt es stets mehrere theoretisch mögliche Kausalmechanismen, die einem gefundenen korrelativen Zusammenhang zugrunde liegen können (Abb. 27.5). Auch aus den Ergebnissen von Leopold und Leutner (2002) kann nicht geschlossen werden, welcher dies ist: Es bleibt unklar, ob der geringe Einsatz von Elaborationsstrategien zu einem geringeren Wortschatz führt oder ob andersherum geringere verbale Fähigkeiten dazu führen, dass Kinder beim Lernen seltener Elaborationsstrategien anwenden. Weiterhin könnte auch eine wechselseitige (reziproke) Abhängigkeit vorliegen. Möglicherweise liegt sogar eine Scheinkorrelation vor, die sich aus dem Einfluss einer unberücksichtigten oder gar unbekannten Drittvariablen ergibt.

Die jeder korrelativen Studie innewohnende Mehrdeutigkeit im Hinblick auf Wirkrichtungen und die damit verbundenen Einschränkungen der internen Validität stellen den Nutzen von Querschnittuntersuchungen keineswegs komplett in Frage. So kann in der pädagogischen Praxis das bloße Wissen um einen bestehenden Zusammenhang bereits hilfreich sein, selbst wenn die jeweilige Ursache-Wirkungs-Be-

**Abb. 27.5** Auswahl denkbarer Zusammenhänge der Variablen x und y mit fiktiven Beispielen

| Wirkrichtung | Kausalmodell | Beispiel |
|---|---|---|
| $x \longrightarrow y$ | *x* beeinflusst *y*. | Der Einsatz von Elaborationsstrategien fördert den Wortschatz. |
| $x \longleftarrow y$ | *y* beeinflusst *x*. | Ein großer Wortschatz begünstigt den Einsatz von Elaborationsstrategien. |
| $x \rightleftarrows y$ | *x* und *y* beeinflussen sich wechselseitig. | Der Einsatz von Elaborationsstrategien und der Umfang des Wortschatzes bedingen sich gegenseitig. |
| $x \quad y$ <br> $\nwarrow \nearrow$ <br> $z$ | *x* und *y* werden von einer weiteren Variable *z* beeinflusst. | Intelligente Eltern befördern sowohl die Lemstrategieentwicklung als auch den Wortschatz ihrer Kinder. |

ziehung (noch) nicht vollständig aufgeklärt ist. Beispielsweise mag es in der Elternberatung durchaus sinnvoll sein, warnend auf den belegten Zusammenhang zwischen übermäßigem Medienkonsum und geringeren Schulleistungen (z. B. Pfeiffer, Mößle, Kleimann & Rehbein 2008) hinzuweisen, obwohl zu den möglichen Wirkungszusammenhängen noch viele Fragen offen sind. Besondere Vorteile von Querschnittuntersuchungen ergeben sich aus ihrer Praktikabilität und Effizienz, weshalb sie trotz ihrer Einschränkungen häufig ein erster Schritt sind, dem intern validere Untersuchungen folgen.

### 27.4.3 Längsschnittuntersuchung

Auch Längsschnittuntersuchungen zählen zu den korrelativen Ansätzen, bei denen Merkmale in ihren natürlichen Ausprägungen untersucht werden, ohne sie gezielt zu beeinflussen. In Längsschnittuntersuchungen werden die interessierenden Merkmale bei denselben Personen zu mindestens zwei unterschiedlichen Zeitpunkten erhoben. Ein solches Vorgehen wird gewählt, wenn Entwicklungen über definierte Zeitspannen hinweg untersucht werden sollen. Diese können von wenigen Tagen bis hin zu vielen Jahren reichen. Da der Beschreibung und Erklärung von Veränderungen gerade in Bildungskontexten besondere Bedeutung zukommt, sind längsschnittliche Designs in diesem Feld trotz ihres hohen Aufwands häufig angezeigt. Eine prototypische Forschungsfrage könnte lauten, ob sich die Lesefähigkeit in den vier Grundschuljahren kontinuierlich aufbaut, oder ob die Entwicklung eher durch Leistungssprünge charakterisiert ist (▶ Kap. 12). Diese Frage lässt sich nur mit einem längsschnittlichen Forschungsdesign klären: Über die gesamte Grundschulzeit hinweg würden in regelmäßigen Abständen Lesetests mit den Kindern durchgeführt. Aus den Verläufen der Leistungen ließen sich die gewünschten Aussagen zur Entwicklung der Lesefähigkeit ableiten.

Neben der reinen Beschreibung von Entwicklungen und Veränderungen ermöglichen Längsschnittuntersuchungen auch Aussagen dazu, welche Faktoren mit den beobachteten Entwicklungen und Veränderungen in Zusammenhang stehen oder gar als ursächlich dafür in Frage kommen. So ließe sich – um im obigen Beispiel zu bleiben – untersuchen, ob die phonologische Bewusstheit im Alter von 5 Jahren die Entwicklung der Leseleistung der Grundschulkinder vorhersagen kann (▶ Kap. 28). Würde sich zeigen, dass die Testergebnisse aus dem Vorschulalter mit dem über die verschiedenen Messzeitpunkte festgestellten Anstieg der Leseleistung in Zusammenhang stehen, würde sich schließen lassen, dass Kinder mit einer guten phonologischen Bewusstheit ihre Lesekompetenzen schneller entwickeln und am Ende der Grundschulzeit höhere Niveaus erreichen. Hier legt die zeitliche Ordnung der Messungen nahe, dass tatsächlich die im Vorschulalter erfassten Vorläuferfähigkeiten die spätere Leistung beeinflusst haben. Insbesondere kann – im Gegensatz zu querschnittlichen Designs – die gegenläufige Wirkrichtung ausgeschlossen werden. Allerdings sind auch hier mögliche Einflüsse von Drittvariablen zu berücksichtigen: So könnte etwa die elterliche Förderung sowohl für hohe Werte in der phonologischen Bewusstheit im Alter von 5 Jahren als auch für die positive Entwicklung der Leseleistung ursächlich sein. Längsschnittliche Untersuchungsdesigns sind somit nicht nur unverzichtbar, um Veränderungen über die Zeit untersuchen zu können, sondern liefern auch Indizien zu den kausalen Bedingungen von Veränderungen.

### 27.4.4 Metaanalyse

Metaanalysen zeichnen sich dadurch aus, dass keine eigenen Daten erhoben, sondern die Ergebnisse bereits vorliegender Studien systematisch gesichtet, bewertet und mit Hilfe statistischer Methoden integriert werden (Rost 2013). Angenommen eine Grundschullehrerin stellt sich die Frage, ob sie den Lernerfolg ihrer Schülerinnen und Schüler durch ein in den Unterricht integriertes Training von kognitiven Lernstrategien fördern kann. Bei der Suche nach entsprechenden empirischen Studien, würde unsere Lehrerin schnell feststellen, dass zur Frage nach den Effekten von Maßnahmen zur Förderung des selbstregulierten Lernens bereits Dutzende von Untersuchungen durchgeführt wurden. Was den Plan, sich einen raschen Überblick zu verschaffen zusätzlich erschwert, ist die Tatsache, dass sich die einzelnen Arbeiten teils deutlich in den konkreten Trainingsmaßnahmen und den untersuchten Merkmalen unterscheiden. Da die Lehrerin kaum die Zeit aufbringen kann, die Vielzahl an Forschungsarbeiten zu sichten und zu bewerten, wäre ihr zu raten, eine Metaanalyse heranzuziehen. Zur Frage nach den Effekten des selbstregulierten Lernens existieren bereits mehrere solche Studien (▶ Kap. 4). Beispielsweise integrierten Dignath und Büttner (2008) 74 Studien, in denen die Wirkung unterschiedlicher Interventionen zum Gebrauch von Lernstrategien untersucht worden ist. Sie gelangten zu der übergreifenden Aussage, dass die Förderung von kognitiven, metakognitiven und motivationalen Strategien im Durchschnitt einen positiven Effekt auf die akademische Leistung zeitigt. Maßnahmen zur Stärkung von Selbstlernkompetenzen wirken also. Aber ist diese Wirkung auch stark genug, um praktische Relevanz im Schulalltag beanspruchen zu können? Um neben einer statistischen Absicherung gegenüber zufälligen Effekten (▶ Abschn. 27.5.2) auch Aussagen zur Stärke von Effekten machen zu können, arbeiten Metaanalysen meist mit Hilfe sogenannter Effektstärken (▶ Im Fokus).

**Im Fokus: Maße der Effektstärke**

Als Effektstärken werden statistische Maße bezeichnet, die Aussagen über die Größe und damit die inhaltliche Bedeutsamkeit von Effekten (z. B. Unterschieden, Zusammenhängen) erlauben. Geht es um die Bedeutsamkeit von Unterschieden zwischen Gruppen, wird häufig das Effektstärkemaß d herangezogen (Cohen 1988). Dieses

ergibt sich aus der Differenz der Mittelwerte der verglichenen Gruppen, die an der vorgefundenen Streuung der Messwerte innerhalb der Gruppen relativiert wurde. Ein wichtiges Effektstärkemaß zur Beurteilung der Enge eines Zusammenhangs zwischen zwei Merkmalen ist der Korrelationskoeffizient $r$ (zu statistischen Begriffen ▶ Abschn. 27.5.1).
Effektstärken werden so berechnet, dass sie unabhängig von der Einheit (Skala) der gemessenen Konstrukte interpretiert werden können. Wenn etwa das Durchschnittsergebnis einer Treatmentgruppe mit Selbstregulationstraining im Deutschtest fünf Punkte über jenem der Kontrollgruppe liegt, bleibt die Frage offen, ob dieser Abstand als klein, mittel oder groß zu bewerten ist. Hingegen ergeben sich beispielsweise aus der Angabe einer Effektstärke von $d = 0.40$ maßstabsunabhängige und damit klare Anhaltspunkte für die Bedeutsamkeit des gefundenen Gruppenunterschieds: Nach ◘ Tab. 27.3, die gängige Konventionen zur Interpretation von Effekten auflistet, lässt sich von einem mittleren Effekt sprechen. Dies ist gleichbedeutend damit, dass ein durchschnittlich leistendes Kind aus der Treatmentgruppe im Deutschtest besser abschneidet als 66 % der Kinder aus der Kontrollgruppe. Angenommen beide Gruppen bestünden aus je 25 Kindern, so würde das durchschnittlich leistende Kind aus der Treatmentgruppe (Rangplatz 13) in der Kontrollgruppe schon den Rangplatz 9 einnehmen.

Was unsere Grundschullehrerin betrifft, würde sie in der Metaanalyse von Dignath und Büttner (2008) lesen, dass über alle Studien hinweg bei Trainings des selbstregulierten Lernens eine Effektstärke von $d = 0.69$ über verschiedene abhängige Variablen (z. B. Strategieeinsatz, Leistung) gefunden wurde. Auf Basis der Interpretationshilfe in ◘ Tab. 27.3 könnte die Lehrerin also durchaus auf mittlere bis große Effekte hoffen, wenn sie ein Training von Lernstrategien in ihren Unterricht integriert. Zudem könnte sie aus den Befunden von Dignath und Büttner (2008) Bedingungen ableiten, die eine besonders große Wirksamkeit des Trainings begünstigen. So zeigten sich beispielsweise Moderatoreffekte dahingehend, dass bei Einbezug metakognitiver Reflexion beim Einsatz von Lernstrategien stärkere Effekte bestanden als bei Trainings ohne Fokus auf die metakognitive Ebene.

Um die praktische Bedeutung von Effektstärken noch besser einschätzen zu können, ist es hilfreich, typische Effekte von unterschiedlichen pädagogischen Maßnahmen zu kennen. Einen großen Fundus hierzu bietet die Zusammenstellung von Hattie (2009). Diese stellt eine übergreifende Metaanalyse dar, die viele auf einzelne Maßnahmen gerichtete Metaanalysen integriert.

Metaanalysen ermöglichen zwar einen schnellen Überblick über eine Vielzahl von Befunden und können damit sehr unterschiedliche und große Stichproben einbeziehen. Jedoch darf nicht außer Acht gelassen werden, dass die Aussagekraft einer Metaanalyse untrennbar mit der Qualität jeder einzelnen der integrierten Primärstudien verknüpft ist. Ferner hängt sie von den Kriterien ab, die zur Auswahl der Studien geführt haben, von der Art und Weise der Verrechnung und Gewichtung der Einzelergebnisse und vom Umgang mit der Problematik, dass Studien, die keine Effekte nachweisen konnten, oft nicht veröffentlicht werden.

Neben den behandelten quantitativen Forschungsansätzen sei auch auf die Vielfalt qualitativer Ansätze wie z. B. Fallstudien oder die beschreibende Feldforschung verwiesen. Eine ausführliche Darstellung qualitativer Forschungsansätze findet sich bei Hussy et al. (2010).

### 27.4.5 Generalisierbarkeit von Untersuchungsergebnissen

Mit dem Begriff der Generalisierbarkeit ist die Frage angesprochen, inwieweit Untersuchungsbefunde tatsächlich auf die (oft facettenreiche) Lebenswirklichkeit, beispielsweise von Lehrenden und Lernenden, übertragbar sind. Zwar gilt unabhängig vom gewählten Forschungsdesign: Je besser die Kontrolle von Störeinflüssen und der Ausschluss von Alternativerklärungen für gefundene Ergebnisse gelingt, desto höher ist die interne Validität der Untersuchung und damit die Eindeutigkeit der Schlussfolgerungen. Hierin liegen auch die Vorteile experimenteller Ansätze mit einem hohen Maß an Kontrolle gegenüber korrelativen Ansätzen mit natürlicher Variation der Variablen. Allerdings birgt eine starke Kontrolle von möglichen Einflussfaktoren die Gefahr, dass die Versuchsbedingungen die Realität nur noch eingeschränkt abbilden können, wie es zum Beispiel beim Nachstellen von Lehr- und Lernsituationen im Labor mitunter der Fall sein kann. Der Begriff der externen Validität dient dazu, Studien in dieser Hinsicht beurteilen zu können.

Eine Untersuchung ist extern valide, wenn ihr Ergebnis über die besonderen Bedingungen der Untersuchungssituation und über die untersuchten Personen hinausgehend generalisierbar ist. Die externe Validität sinkt mit wachsender Unnatürlichkeit der Untersuchungsbedingungen bzw. mit abnehmender Repräsentativität der untersuchten Stichproben (Bortz & Schuster 2010).

Externe Validität bezieht sich also darauf, inwieweit die Ergebnisse einer Studie auf reale zukünftige Gegebenheiten übertragbar sind und auch für solche Personen, Kontexte und Sachverhalte Gültigkeit beanspruchen können, die in der Untersuchung nicht direkt abgedeckt waren. In ▶ Abschn. 27.2.2 wurde über die quasiexperimentelle Interventionsstudie von Labuhn et al. (2008) berichtet, bei der Siebtklässlerinnen und Siebtklässler einer nordrhein-westfälischen Gesamtschule untersucht wurden. Was die externe Validität dieser Studie betrifft, wäre beispielsweise zu fragen, ob der Befund, dass Anregungen zum selbstregulierten Lernen zu Vorteilen

**Tabelle 27.3** Interpretation von Effektstärken (nach Coe 2002)

| Klassifikation der Effektstärke | Effektstärke *d* | Prozentanteil an Personen der Kontrollgruppe mit Werten unterhalb des Werts einer „Durchschnittsperson" in der Treatmentgruppe | Rang einer Person in der Kontrollgruppe von 25 Personen, die der „Durchschnittsperson" in der Treatmentgruppe entspricht |
|---|---|---|---|
| | 0.0 | 50 % | 13 |
| | 0.1 | 54 % | 12 |
| klein | 0.2 | 58 % | 11 |
| | 0.3 | 62 % | 10 |
| | 0.4 | 66 % | 9 |
| mittel | 0.5 | 69 % | 8 |
| | 0.6 | 73 % | 7 |
| | 0.7 | 76 % | 6 |
| groß | 0.8 | 79 % | 6 |
| | 0.9 | 82 % | 5 |
| | 1.0 | 84 % | 4 |
| | ≥1.2 | ≥88 % | ≤3 |

beim Lernerfolg führen, auch für Kinder anderer Schularten oder Jahrgangsstufen Gültigkeit beanspruchen kann. Für die externe Validität der Studie spricht unter anderem, dass die Anregungen zum selbstregulierten Lernen in den realen naturwissenschaftlichen Fachunterricht integriert waren und von Lehrkräften dargeboten wurden. Ob die Befunde allerdings auch für Kinder in der Grundschule gelten, bleibt fraglich, da nur Kinder der siebten Klassen an der Untersuchung teilgenommen hatten. Wie das Beispiel zeigt, sind Informationen zur Auswahl der Teilnehmenden an einer Studie, also zur Zusammensetzung der Stichprobe, ein wichtiges Kriterium, um beurteilen zu können, wie generalisierbar die jeweiligen Befunde sind.

### Im Fokus: Grundgesamtheit und Stichprobe

Als Grundgesamtheit oder Population werden all jene Personen bezeichnet, denen ein umschriebenes Erkenntnisinteresse gilt und für die etwaige Befunde gelten sollen. Geht es etwa um die Bedingungen der Wahl verschiedener Ausbildungsberufe, könnten alle Schülerinnen und Schüler in Deutschland, die nach der Schule eine Berufsausbildung aufnehmen, die Grundgesamtheit darstellen.

Die Stichprobe ist die für eine Untersuchung ausgewählte Teilmenge der Grundgesamtheit, welche diese möglichst gut repräsentieren sollte.

Mit dem Begriff der Repräsentativität einer Stichprobe ist gemeint, dass alle für die Fragestellung relevanten Merkmale der ausgewählten Personen ähnlich wie in der Grundgesamtheit verteilt sind. Repräsentativ oder nicht ist eine Stichprobe also in Bezug auf bestimmte Merkmale, weshalb diese auch angegeben sein sollten.

Eine Zufallsstichprobe entsteht durch zufällige Auswahl von Personen der Grundgesamtheit, die jeweils identische Chancen auf Aufnahme haben. Dies ist die beste und einfachste Methode, um repräsentative Stichproben zu gewinnen, lässt sich in der Forschungspraxis jedoch oft nicht realisieren. Ist eine Zufallsstichprobe ausreichend groß, kann angenommen werden, dass die Untersuchungsergebnisse in jeder Hinsicht auch für die Grundgesamtheit gelten.

Eine Quotenstichprobe wird bewusst so zusammengesetzt, dass als besonders wichtig erachtete Merkmale (z. B. sozioökonomischer Status) in einem ähnlichen Verhältnis vorkommen wie aus der Grundgesamtheit bekannt. Bereits bei diesem Vorgehen ist mit Einschränkungen der Generalisierbarkeit zu rechnen, da niemals alle Personenmerkmale kontrolliert werden können.

Gelegenheitsstichproben sind solche, die ohne weitere Vorkehrungen zur Sicherung der Repräsentativität gewonnen werden; beispielsweise wenn Schülerinnen und Schüler untersucht werden, die in Schulen vor Ort und damit gut erreichbar sind. In diesem Fall muss die Frage, inwieweit die Befunde generalisierbar sind, besonders genau beleuchtet werden.

Der Stichprobenfehler (oder Standardfehler) bezeichnet die zufällige Abweichung der Gegebenheiten in der Stichprobe von denen in der Grundgesamtheit. Der Stichprobenfehler ist umso kleiner, je größer die Stichprobe ist – weshalb größere Stichproben das Potential für präzisere Aussagen und den Nachweis bereits kleiner Effekte haben. Er ist entscheidend für die statistische Beurteilung der Frage, ob aus einem in einer Stichprobe gewonnenen Befund gefolgert werden kann, dass dieser auch in der Grundgesamtheit gilt (Inferenzstatistik, ▶ Abschn. 27.5.2).

## 27.5 Analysemethoden

Bei der Analyse quantitativer Daten lassen sich zwei Vorgehensweisen unterscheiden: Die deskriptive Statistik (beschreibende Statistik) stellt Werkzeuge und Maße zur Verfügung, die ein zusammenfassendes Bild der Daten in der Stichprobe vermitteln. Bei der Inferenzstatistik (schließende Statistik) geht es um die Frage, ob aus den Befunden der untersuchten Stichprobe mit hinreichend kleiner Irrtumswahrscheinlichkeit auf die Verhältnisse in der Grundgesamtheit geschlossen werden kann.

### 27.5.1 Deskriptive Statistik

Um eine erste Übersicht über die gewonnenen Daten zu bekommen, sind einfache Tabellen oder Diagramme geeignet.

#### Häufigkeitsverteilung

Sehr anschaulich lässt sich die Häufigkeitsverteilung eines Merkmals mit Hilfe eines Säulendiagramms darstellen. Die Säulen auf der x-Achse stehen dabei für die einzelnen Merkmalsausprägungen, während die Höhe einer Säule (y-Achse) anzeigt, wie oft das jeweilige Merkmal in der Stichprobe beobachtet wurde (Kasten).

#### Lagemaße

Lagemaße dienen dazu, die Tendenz der Verteilung eines Merkmals in einer Stichprobe mit Hilfe von Kennzahlen zusammenfassend zu charakterisieren. Solche Größen werden entsprechend auch als Maße der zentralen Tendenz bezeichnet. Im Kasten findet sich ein Säulendiagramm für eine fiktive Verteilung von Werten in einem Kompetenztest. An diesem Beispiel werden die wichtigsten Lagemaße erklärt.

**Im Fokus: Häufigkeitsverteilung mit Lagemaßen**

Angenommen eine Stichprobe von 25 Schülerinnen und Schülern hat an einem Kompetenztest zum selbstregulierten Lernen teilgenommen. Der Test soll intervallskalierte Kompetenzwerte zwischen 1 (keine Kompetenz) und 7 (maximale Kompetenz) liefern. Die Verteilung des Merkmals „Selbstregulationskompetenz" zeigt sich in einem Histogramm (Abb. 27.6). Auf der x-Achse finden sich die sieben möglichen Ausprägungen des Testergebnisses. Die Höhe eines Balkens repräsentiert die Häufigkeit des jeweiligen Ergebnisses – also die Anzahl der Personen, die dieses Ergebnis erreicht haben.

**Mittelwert**

Um die Durchschnittsleistung $M$ im Kompetenztest zu berechnen, werden die Kompetenzwerte der Schülerinnen und Schüler aufsummiert und durch deren Anzahl n geteilt:

$$M_x = \frac{(x_1 + x_2 + \ldots)}{n}$$

Im Beispiel liegt der mittlere Kompetenzwert bei 2.84.

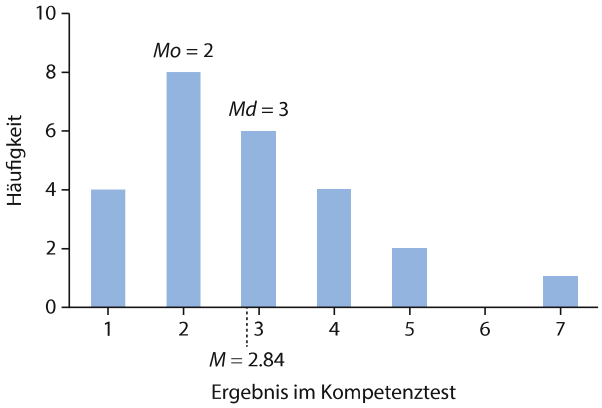



**Abb. 27.6** Säulendiagramm der Häufigkeitsverteilung des Merkmals „Selbstregulationskompetenz" in einer fiktiven Stichprobe von 25 Kindern. Eingezeichnet sind Mittelwert (*M*), Median (*Md*) und Modus (*Mo*)

**Median**

Der Median ist derjenige tatsächlich beobachtete Wert, der eine Verteilung in zwei gleich große Hälften teilt. Bei den 25 Beobachtungen des obigen Beispiels (1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,5,5,7) ist dies der Wert von $Md = 3$.

**Modus**

Als Modus (Modalwert) wird der am häufigsten beobachtete Wert einer Verteilung bezeichnet. Im Beispiel ist dies der Wert von $Mo = 2$, den acht Personen erreichten.

#### Streuungsmaße

Wenn es darum geht, Unterschiede zwischen Personen und Gruppen angemessen zu beschreiben und zu quantifizieren, sind Streuungsmaße hilfreich. Aus diesen geht hervor, wie stark die beobachteten Werte voneinander abweichen. Die gebräuchlichsten Streuungsmaße werden im Kasten erklärt.

**Im Fokus: Streuungsmaße**

Angenommen eine Lehrerin hat zu Beginn des Schuljahres ihren beiden achten Klassen einen Kompetenztest zum selbstregulierten Lernen vorgelegt, der intervallskalierte Testwerte liefert. Bei Klasse 8a hat sie einen Mittelwert von $M = 3.96$ errechnet, für Klasse 8b den sehr ähnlichen Wert von $M = 3.88$. Wie die Darstellungen der beiden Häufigkeitsverteilungen zeigen (Abb. 27.7), muss sie

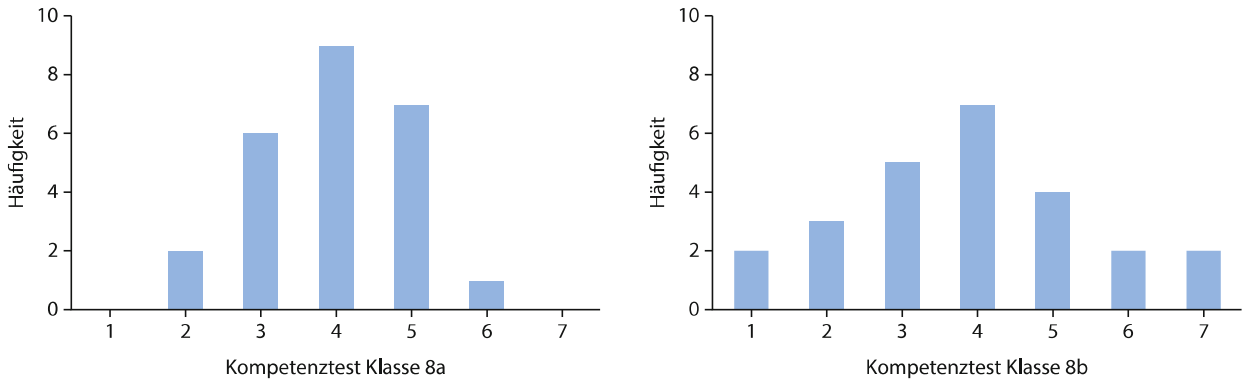



**Abb. 27.7** Häufigkeitsverteilungen des Merkmals „Selbstregulationskompetenz" in den Klassen 8a und 8b

sich trotzdem auf sehr unterschiedliche Vorbedingungen in den beiden Klassen einstellen. So hätte sie in Klasse 8b mit einigen Kindern zu tun, die bereits über sehr gute Kompetenzen verfügen, während gleichzeitig auch viele Kinder zu fördern sind – hier besteht also eine große Heterogenität hinsichtlich der Lernvoraussetzungen der Schülerinnen und Schüler. In Klasse 8a hingegen streuen die beobachteten Kompetenzwerte deutlich weniger um den Mittelwert.

**Streuungsbreite**

Der Abstand zwischen der geringsten und der größten Ausprägung des betrachteten Merkmals ist die Streuungsbreite (Spannweite). Diese liegt bei Klasse 8a bei dem Wert 4, bei Klasse 8b bei dem Wert 6. Zu beachten ist, dass die Streuungsbreite stark von Extremwerten beeinflusst ist und erst ab Ordinalskalenniveau zur Anwendung kommen kann.

**Perzentile**

Der Median wird auch als 50. Perzentil bezeichnet, da 50 % der Stichprobe gleiche oder geringere Merkmalsausprägungen haben. Nach derselben Logik lassen sich weitere Perzentile angeben. So entspräche in der Klasse 8b der Kompetenzwert von 2 dem 20. Perzentil, da 20 % der Kinder (fünf von insgesamt 25) einen Testwert von 2 oder geringer erzielten.

**Varianz**

Um die Variabilität in einer Stichprobe zu quantifizieren, eignet sich die Varianz, deren Berechnung Intervallskalenniveau voraussetzt. Dafür wird für jeden Wert die Differenz zum Mittelwert berechnet. Damit sich positive und negative Abweichungen vom Mittelwert nicht gegenseitig aufheben und größere Abweichungen stärker gewichtet werden, werden diese Differenzen quadriert. Der Durchschnitt dieser quadratischen Abweichungen ist die Varianz:

$$Var_x = \left[(M_x - x_1)^2 + (M_x - x_2)^2 + \ldots\right] / n$$

Die Varianz in Klasse 8a ergibt sich damit zu $Var = 1.00$. Für Klasse 8b fällt die Varianz mit $Var = 2.59$ deutlich höher aus.

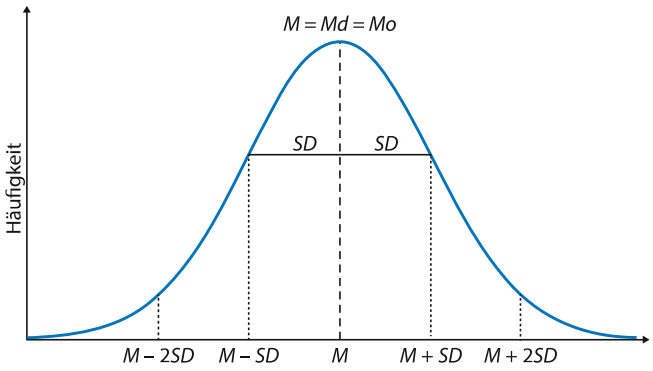



**Abb. 27.8** Normalverteilungskurve mit Mittelwert ($M$) und Standardabweichung ($SD$)

**Standardabweichung**

Häufig wird statt der Varianz die Standardabweichung angegeben, um die in einer Stichprobe beobachtete Streuung eines Merkmals zu quantifizieren. Diese ergibt sich als Quadratwurzel der Varianz, wodurch ein Streuungsmaß in der ursprünglichen Metrik resultiert, das gut interpretierbar ist:

$$SD_x = \sqrt{Var_x}$$

Die Standardabweichungen der Kompetenzwerte ergeben sich zu $SD = 1.00$ für Klasse 8a und zu $SD = 1.61$ für Klasse 8b.

Die Verteilungen vieler psychischer Merkmale sind näherungsweise normalverteilt. Abb. 27.8 zeigt eine Normalverteilungskurve, die sich ergeben könnte, wenn eine große Zahl von Kindern beispielsweise einem Selbstregulationstest unterzogen würde. Analog zu Säulendiagrammen repräsentiert die Abszisse die Merkmalsausprägung und die Ordinate die Häufigkeit, mit der die jeweiligen Werte auftreten.

Anhand der Normalverteilungskurve soll die Bedeutung von Lage- und Streuungsmaßen nochmals illustriert werden: Ist ein Merkmal normalverteilt, sind Mittelwert, Median und Modus deckungsgleich (bei schiefen Verteilungen wie

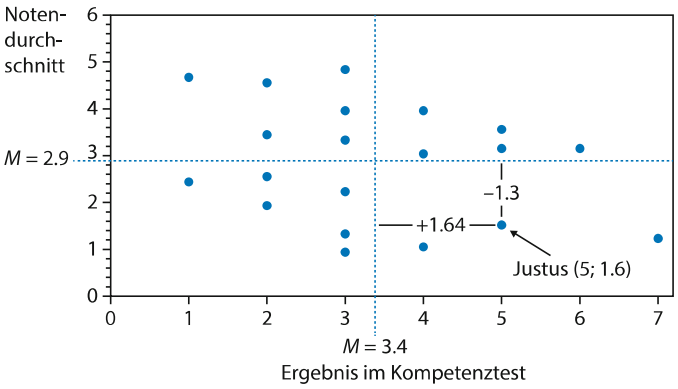



**Abb. 27.9** Zusammenhang des Merkmals „Selbstregulationskompetenz" mit dem Merkmal „Schulleistung" in einer fiktiven Stichprobe von 20 Kindern

in Abb. 27.6 ist dies nicht der Fall). Die Standardabweichung spiegelt sich in der Breite der Verteilungskurve wider. Bei gegebener Normalverteilung werden Merkmalsausprägungen von einer Standardabweichung unter dem Mittelwert ($M - SD$) bis hin zu einer Standardabweichung über dem Mittelwert ($M + SD$) bei 68.3 % der untersuchten Personen beobachtet. Im Intervall von zwei Standardabweichungen vom Mittelwert finden sich 95.5 % aller Personen.

## Zusammenhangsmaße

Um zu analysieren, ob und wie stark zwei oder auch mehrere Merkmale miteinander zusammenhängen, steht eine große Zahl teils sehr fortgeschrittener Verfahren zur Verfügung (z. B. Regressionsanalysen, Strukturgleichungsmodelle; Übersicht bei Eid et al. 2015). Im Folgenden wird die Berechnung der für alle Verfahren grundlegenden Korrelation besprochen. Dabei handelt es sich um ein Maß, das die gemeinsame Variation zweier Merkmalen quantifiziert. Dieses wird auch als Korrelationskoeffizient $r$ bezeichnet und dient dazu, über Stärke und Richtung von Zusammenhängen zwischen zwei Variablen Auskunft zu geben. Anhand eines Beispiels ist im Kasten dargestellt, wie der Korrelationskoeffizient gebildet wird.

### Im Fokus: Korrelation

Angenommen, in einer Stichprobe von 20 Kindern soll der Zusammenhang zwischen den Ergebnissen eines Tests zur Selbstregulationskompetenz und Zeugnisschnitten untersucht werden. Abb. 27.9 zeigt 20 Datenpunkte, die entstehen, wenn für jedes Kind das Ergebnis im Kompetenztest auf der Abszisse und der Notendurchschnitt auf der Ordinate abgetragen werden. Bei genauer Betrachtung zeigt sich, dass Kinder mit geringen Kompetenzwerten tendenziell schlechtere Schulleistungen erzielt haben, als solche mit höheren Werten.

Um den Zusammenhang von zwei Variablen zu quantifizieren wird zunächst die Kovarianz bestimmt. Diese drückt Ausmaß und Richtung der gemeinsamen Variation („Ko-Variation") zweier Merkmale als Zahlenwert aus. Sie wird bestimmt, indem zunächst pro Person für jeden der beiden Messwerte die Abweichung vom jeweiligen Mittelwert bestimmt und diese beiden Abweichungen dann miteinander multipliziert werden. Für den Schüler Justus in Abb. 27.9 ergibt sich dieses Produkt der einzelnen Abweichungen zu $-1.30 \cdot +1.60 = -2.08$. Das negative Vorzeichen verrät, dass bei Justus eine gegenläufige Beziehung von Notendurchschnitt und Selbstregulationskompetenz besteht. Die Stärke des Zusammenhangs findet im Absolutbetrag des Produkts der beiden Abweichungen ihren Niederschlag. Der Mittelwert dieser Abweichungsprodukte ist die Kovarianz zwischen den beiden Variablen x und y:

$$Cov(x, y) = \left[(M_x - x_1) \cdot (M_y - y_1) + (M_x - x_2) \cdot (M_y - y_2) + \ldots\right] / n$$

Die Kovarianz hat den Nachteil, dass sie von der zugrundeliegenden Skala abhängt und deshalb schlecht über verschiedene Variablen hinweg zu vergleichen ist. Benötigt wird deshalb eine Standardisierung.

Die Korrelation $r$ ist nun nichts anderes als die standardisierte Kovarianz. Die Standardisierung wird erreicht, indem die Kovarianz an den Standardabweichungen beider Merkmale relativiert wird:

$$r(x, y) = \frac{Cov(x, y)}{SD_x \cdot SD_x}$$

In unserem Beispiel beträgt die Korrelation $r = -.27$. Sie bedeutet, dass mit höherer Selbstregulationskompetenz tendenziell niedrige Notenschnitte verbunden sind.

Mit der Korrelation steht eine Maßzahl zur Verfügung, die es erlaubt, den Zusammenhang zweier Merkmale in einer Stichprobe unabhängig von Skalenmaßstäben zu quantifizieren. Der Korrelationskoeffizient kann theoretisch Werte zwischen $+1.00$ und $-1.00$ annehmen. Bei $r = +1.00$ wird von einem perfekt positiven, $r = -1.00$ von einem perfekt negativen Zusammenhang gesprochen. Ist $r = 0$, besteht kein Zusammenhang. In Abb. 27.10 sind Punktewolken für eine hohe negative, eine hohe positive und eine nicht vorhandene Korrelation der Merkmale $x$ und $y$ dargestellt.

In ▶ Abschn. 27.4.4 wurde die Effektgröße $d$ zur Beurteilung von Mittelwertsunterschieden besprochen. Analog da-

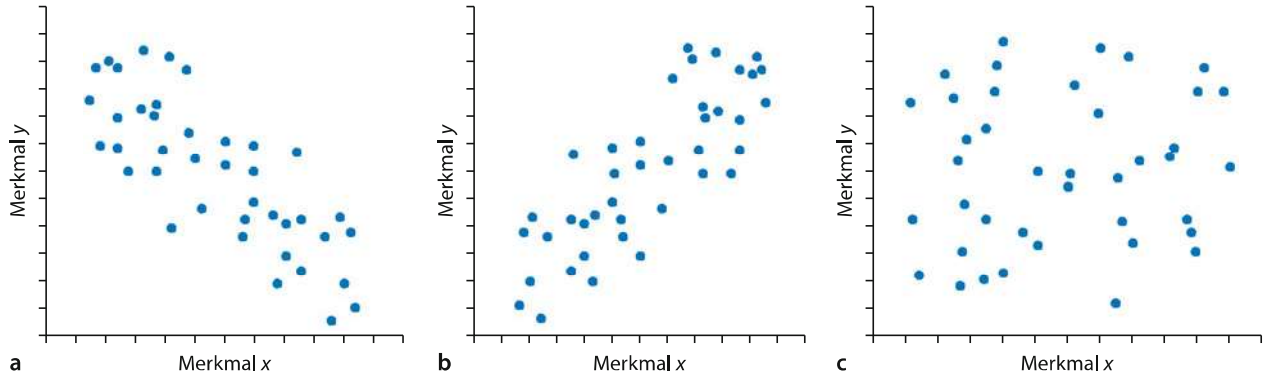



**Abb. 27.10** Datenpunkte mit Ausprägungen von zwei Merkmalen $x$ und $y$. **a** hohe negative Korrelation. **b** hohe positive Korrelation. **c** Nullkorrelation

zu bietet der Zahlenwert des Korrelationskoeffizienten einen guten Anhaltspunkt, um die Bedeutsamkeit und praktische Relevanz von Zusammenhängen zwischen Merkmalen einschätzen zu können. Zurückgehend auf Cohen (1988) wird meist bei $r = \pm.10$ von einem schwachen Zusammenhang gesprochen, bei $r = \pm.30$ von einem mittleren und bei $r = \pm.50$ von einem starken Zusammenhang. Hilfreich für die Interpretation von Korrelationen ist es außerdem, regelmäßig gefundene Werte für typische Zusammenhänge zu kennen. Sehr robust ist beispielsweise die Korrelation um $r = .50$ zwischen Intelligenz und Schulleistung (Maltby, Day & Macaskill 2011).

**Im Fokus: Nullhypothese und Alternativhypothese**

Zunächst wird eine Nullhypothese $H_0$ formuliert, die keinen Zusammenhang, Unterschied oder Effekt annimmt. Komplementär dazu wird eine Alternativhypothese $H_1$ formuliert, die das gegenteilige Ergebnis postuliert. Sie ist in der Regel identisch mit der zu prüfenden Aussage.
In unserem Beispiel:
$H_0$: Bei Schülerinnen und Schülern existiert kein Zusammenhang zwischen Selbstregulationskompetenz und Schulleistung.
$H_1$: Bei Schülerinnen und Schülern hängen Selbstregulationskompetenz und Schulleistung zusammen.

## 27.5.2 Inferenzstatistik

Das Ziel der Inferenzstatistik ist es, zur Überprüfung von Forschungshypothesen von den in einer Stichprobe vorgefundenen Bedingungen auf die Gegebenheiten in der Grundgesamtheit zu schließen. Entsprechend wird dieses Vorgehen auch als schließende oder hypothesenprüfende Statistik bezeichnet. Angenommen, die obige Fragestellung hätte gelautet, ob in der Stichprobe von 20 Kindern die Werte des Kompetenztests mit der Durchschnittsnote zusammenhängen: Die Antwort wäre ein klares „Ja" gewesen, denn schließlich wurde eine Korrelation von $r = -.27$ ermittelt. Der Anspruch wissenschaftlicher Hypothesen liegt aber gerade darin, möglichst allgemeingültige Aussagen zu treffen, die über den untersuchten Personenkreis hinausgehen. Es stellt sich also die Frage, ob aus der Korrelation in der Stichprobe geschlossen werden kann, dass ein Zusammenhang auch in der Grundgesamtheit besteht. Schließlich könnten Kinder mit sowohl guten Kompetenzwerten als auch guten Schulleistungen allein durch eine Laune des Zufalls in die Stichprobe geraten sein (Stichprobenfehler, ▶ Abschn. 27.4.5). Im Folgenden wird skizziert, wie eine Beantwortung dieser Frage möglich wird (ausführlich in Döring & Bortz 2016).

In einem ersten Schritt wird zur prüfenden Forschungshypothese eine Nullhypothese und eine Alternativhypothese formuliert (▶ Im Fokus).

Ausgangspunkt inferenzstatistischer Hypothesentestungen ist die Vermutung der Gültigkeit der Nullhypothese in der Grundgesamtheit. Nur wenn die Daten der Stichprobe hinreichend Anlass geben, die Nullhypothese anzuzweifeln, wird sie zugunsten der Alternativhypothese verworfen. Ob dies gerechtfertigt ist, wird mit Hilfe eines statistischen Tests entschieden, der auch als Nullhypothesen-Signifikanztest bezeichnet wird. Dabei wird ermittelt, wie groß die Wahrscheinlichkeit ist, das nachgewiesene Datenmuster in einer Stichprobe zu finden, obwohl die Nullhypothese zutrifft. Nur wenn diese Wahrscheinlichkeit sehr gering ist und unterhalb einer zuvor festgelegten Grenze liegt, gehen Forscherinnen und Forscher vom Vorliegen desselben Musters auch in der Grundgesamtheit aus. Übertragen auf unser Beispiel wäre also zu berechnen, wie groß in einer 20-köpfigen Stichprobe die Wahrscheinlichkeit ist, einen Zusammenhang zwischen Kompetenzwerten und Noten von $r = -.27$ oder enger zu finden, obwohl Selbstregulationskompetenz und Schulleistung in der Grundgesamtheit nicht zusammenhängen, also die Nullhypothese gilt. Diese sogenannte Irrtumswahrscheinlichkeit wird auch als Signifikanzwert p bezeichnet und lässt sich in Tabellen nachschlagen oder direkt von der verwendeten Statistiksoftware ausgeben. Sie beträgt in unserem Beispiel $p = .23$, also 23 %. Statistisch gesehen wären somit unter 100 gezogenen Stichproben immerhin 23, die rein zufällig einen Zusammenhang von $r = -.27$ oder enger auf-

weisen. Forscherinnen und Forschern wäre dieses Risiko zu groß und sie würden die Nullhypothese beibehalten.

Bereits vor der Untersuchung ist festzulegen, welche Irrtumswahrscheinlichkeit als gerade noch ausreichend gering gelten soll, um die Alternativhypothese anzunehmen. Diese Höchstgrenze wird als Signifikanzniveau oder auch als α-Niveau bezeichnet. Typischerweise werden Signifikanzniveaus von $p < .05$ (5 %-Niveau), $p < .01$ (1 %-Niveau) oder $p < .001$ (0.1 %-Niveau) angesetzt. Liegt der ermittelte Signifikanzwert unter dem festgelegten Niveau werden die Befunde als „signifikant" bezeichnet; tut er es nicht, wird von rein zufälligen Effekten ausgegangen, die keine inhaltliche Bedeutung haben und nicht interpretiert werden können.

Angemerkt sei noch, dass im Beispiel mit einer „zweiseitigen" Hypothese gearbeitet wurde. Das heißt, die Alternativhypothese enthielt keine Annahme zur Richtung des Zusammenhangs der thematisierten Variablen. Behauptet wurde lediglich ein von Null verschiedener Zusammenhang. Denkbar wäre auch gewesen, die Hypothese „einseitig" als Vermutung eines positiven Zusammenhangs zwischen Selbstregulationskompetenz und Schulleistung zu formulieren. Tatsächlich werden Alternativhypothesen in der Forschungspraxis häufig im Hinblick auf die erwartete Richtung formuliert. Ein solches Vorgehen ist jedoch nur zu rechtfertigen, wenn der gegenläufige Zusammenhang aufgrund solider theoretischer Überlegungen und empirischer Befunde aus früheren Untersuchungen sicher ausgeschlossen werden kann (vgl. Rost 2013).

Insgesamt ermöglicht es die Inferenzstatistik, Aussagen zu großen Gruppen zu treffen, obwohl nur ein Teil davon tatsächlich untersucht wurde. Der Preis dafür ist die Möglichkeit sich zu irren, die jedoch durch Berechnung des Signifikanzwerts kalkulierbar wird. Im Kasten werden die zwei grundlegenden Arten von Fehlern näher beschrieben, die mit der inferenzstatistischen Hypothesentestung einhergehen können.

**Im Fokus: Mögliche Fehlentscheidungen bei der Testung**

Als Fehler 1. Art oder auch als α-Fehler wird der Fall bezeichnet, dass die Alternativhypothese angenommen wird, obwohl die Nullhypothese zutrifft. Komplementär zum Fehler 1. Art bezeichnet der Fehler 2. Art, der auch β-Fehler genannt wird, den Fall, dass die Nullhypothese trotz gültiger Alternativhypothese beibehalten wird.
Während der höchstens zu akzeptierende α-Fehler mit dem Signifikanzniveau schon vor der Untersuchung festgelegt wird, ist die Wahrscheinlichkeit für den β-Fehler von vielfältigen Faktoren wie beispielsweise Stichprobengröße und Qualitätsmerkmalen der Studie abhängig (Rost 2013). Zudem sind die beiden Fehlerarten miteinander verknüpft. So sinkt mit der Wahl eines sehr strengen Signifikanzniveaus zwar das Risiko, die Alternativhypothese fälschlicherweise anzunehmen. Damit wird es aber auch wahrscheinlicher, den Fehler 2. Art zu begehen und so einen tatsächlich vorhandenen Effekt zu übersehen (▶ Kap. 24).

Unabhängig von den Möglichkeiten der Inferenzstatistik, mit statistischen Unabwägbarkeiten umzugehen, ist es unabdingbar, Fehlerquellen im Forschungsprozess so klein wie möglich zu halten, um am Ende verlässliche Aussagen treffen zu können. Dies reicht von einer ausreichend großen Stichprobe, die mit angemessenem Auswahlverfahren gewonnen wird, über ein geeignetes Untersuchungsdesign bis hin zur Wahl passender Messinstrumente, die den Gütekriterien möglichst gut entsprechen.

## 27.6 Finden, Lesen und Bewerten von psychologischen Forschungsstudien

Die bisherigen Abschnitte dienten dazu, grundlegendes forschungsmethodisches Wissen zu vermitteln, um empirische Forschungsergebnisse verstehen und bewerten zu können. Die zentralen Bausteine (Hypothesen, Schritte im Forschungsprozess, Erhebungsmethoden, Untersuchungsdesigns und Analysemethoden) werden nun in diesem abschließenden Abschnitt zusammengeführt. Dies soll zur Klärung zweier praktischer Fragen beitragen, die sich bei der Nutzung des reichhaltigen Fundus an praxistauglichen Forschungsergebnissen stellen, den die Pädagogische Psychologie bereitstellt: (1) Wie finde ich belastbare Forschungsergebnisse zu einem praktischen Phänomen? (2) Wie lese ich einen psychologischen Originalartikel?

### 27.6.1 Wie finde ich belastbare Forschungsergebnisse zu einem praktischen Phänomen?

Die erste Wahl für die Veröffentlichung und Suche hochwertiger Forschungsergebnisse der Psychologie sind Publikationen in Fachzeitschriften – im Gegensatz zu manch anderen Fächern, in denen Forschungsergebnisse vorrangig in Monographien oder Herausgeberwerken veröffentlicht werden. Gründe für diese Präferenz für Fachzeitschriften liegen darin, dass sich empirische Studien gut im Format eines Fachartikels darstellen lassen, diese sehr einfach (elektronisch) einer breiten Leserschaft verfügbar gemacht werden können und Fachzeitschriften eine strenge Qualitätskontrolle vornehmen.

Eine solche Qualitätssicherung ist notwendig, um zu gewährleisten, dass nur Studien veröffentlicht werden, die einen innovativen Beitrag zum Forschungsstand liefern und methodische Standards erfüllen (z. B. Güte der verwendeten Messinstrumente, statistische Absicherung von Schlussfolgerungen). Ganz besonders ist dies bei einer anspruchsvollen methodischen Anlage nötig, weil dann Leserinnen und Leser oft gar nicht in der Lage sind, die Methodik nachvollziehen zu können. Die Qualitätskontrolle wird in erster Linie durch strenge und doppelblinde Begutachtungsverfahren erreicht. Dabei werden Forschungsarbeiten von mehreren Expertinnen und Experten aus dem gleichen Fachge-

**Tabelle 27.4** Wichtige Fachzeitschriften aus dem Bereich der schulbezogenen Psychologie

| Deutschsprachige Fachzeitschriften | Englischsprachige Fachzeitschriften |
|---|---|
| Kindheit und Entwicklung | American Educational Research Journal |
| Psychologie in Erziehung und Unterricht | British Journal of Educational Psychology |
| Unterrichtswissenschaft | Contemporary Educational Psychology |
| Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie | Educational Psychologist* |
| Zeitschrift für Erziehungswissenschaft | Educational Psychology |
| Zeitschrift für Pädagogische Psychologie | Educational Psychology Review* |
| | European Journal of Psychology of Education |
| | International Journal of Educational Research |
| | Journal of Educational Psychology |
| | Journal of Educational Research |
| | Journal of Experimental Education |
| | Journal of School Psychology |
| | Journal of the Learning Sciences |
| | Learning and Individual Differences |
| | Learning and Instruction |
| | Metacognition and Learning |
| | Review of Educational Research* |
| | Teaching and Teacher Education |

Mit * gekennzeichnete Zeitschriften veröffentlichen vorrangig systematische Überblicksdarstellungen („Reviews") und Metaanalysen.

biet beurteilt, wobei sowohl die Autorinnen und Autoren als auch die Gutachterinnen und Gutachter anonym bleiben. Tab. 27.4 listet wichtige deutschsprachige und internationale Fachzeitschriften auf, die eine solche Qualitätskontrolle realisieren.

Die einzelnen Fachartikel sind nicht in den OPAC-Systemen (Online Public Access Catalogue) der Bibliotheken verzeichnet. Eine bessere Suchstrategie ist die Recherche in einer psychologischen Literaturdatenbank, die über die Datenbanken-Informationssysteme (DBIS) der Universitätsbibliotheken aufgerufen werden kann. Zu nennen sind insbesondere die Datenbanken PSYNDEX (die psychologische Forschungsarbeiten aus den deutschsprachigen Ländern katalogisiert und mehr als eine viertel Million Publikationen, Tests, audiovisuelle Medien und Interventionsprogramme enthält) sowie PsycINFO (internationale Datenbank, die mehr als vier Millionen Einträge umfasst). In diesen Datenbanken ist eine gezielte Suche nach Begriffen in Titeln, Zusammenfassungen, Schlagwörtern und anderen Feldern möglich. Daneben können auch Quellenhinweise in Lehrbüchern und Inhaltsverzeichnisse von Fachzeitschriften brauchbare Ergebnisse liefern.

Wenn es das Ziel ist, sich einen systematischen Überblick über ein Forschungsthema zu verschaffen, kann gezielt nach Überblicksdarstellungen („Reviews") und Metaanalysen recherchiert werden (z. B. in den in Tab. 27.4 gekennzeichneten Zeitschriften).[1]

Sind passende Fachartikel identifiziert, können die Volltexte meist über die Elektronische Zeitschriftenbibliothek (EZB) der Universitätsbibliotheken, aber auch über einschlägige Internet-Suchmaschinen abgerufen werden.

## 27.6.2 Wie lese ich einen psychologischen Originalartikel?

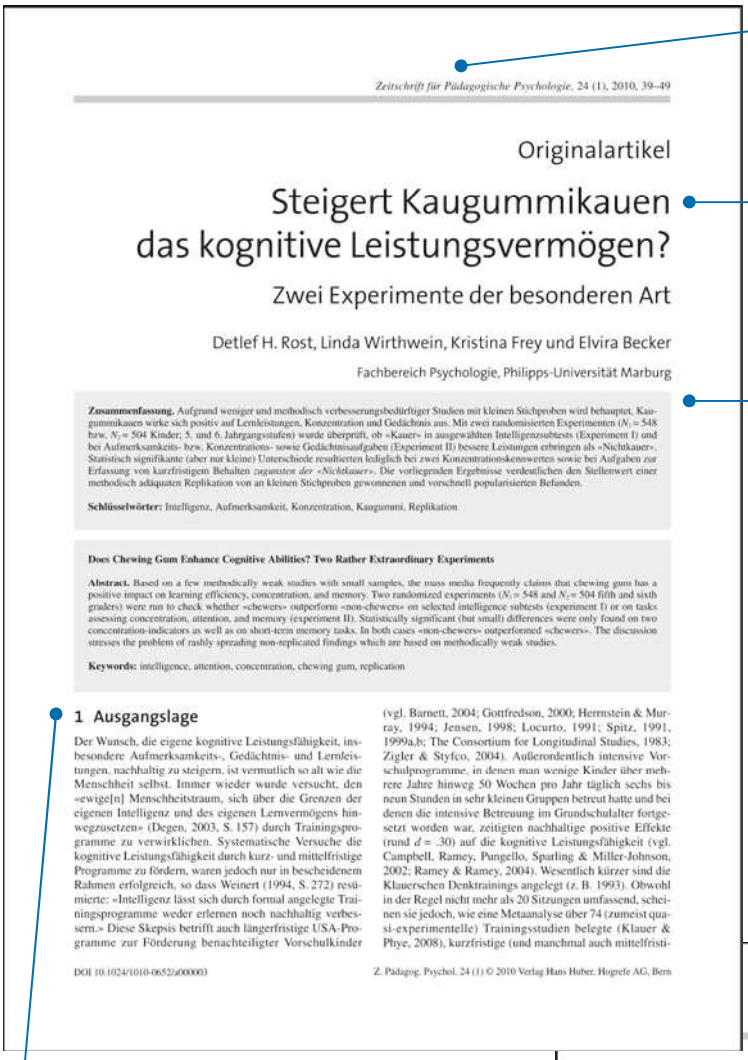
So leicht es klingt, so schwer fällt es vielen (nicht nur Studierenden), psychologische Fachartikel zu lesen – und dabei die zentralen Aussagen zu verstehen sowie die Belastbarkeit der empirischen Ergebnisse zu beurteilen (von der Mühlen, Richter, Schmid, Schmidt & Berthold 2016). Originalarbeiten – also Fachartikel, die neue („originäre") Erkenntnisse präsentieren – sind grundsätzlich schwieriger zu lesen als Lehrbuchtexte und andere Überblicksarbeiten.

Um Sie bei der Lektüre psychologischer Originalarbeiten zu unterstützen, finden Sie auf den folgenden Seiten eine Lesehilfe (Abb. 27.11). Diese soll Sie Schritt für Schritt durch die Lektüre leiten, auf wichtige Punkte aufmerksam machen und Ihnen helfen, etwaige Verständnisschwierigkeiten zu überwinden (die oft im forschungsmethodischen Bereich liegen). Die Originalstudie, anhand derer diese Lesehilfe gestaltet ist, befasst sich mit dem bereits erwähnten populären Mythos, dass Kaugummikauen das kognitive Leistungsvermögen steigern könne (Rost et al. 2010). Es ist empfehlenswert, den Artikel parallel zur Lektüre dieses Abschnitts zu lesen.[2]

Ausführlichere Hinweise zur Interpretation und Bewertung psychologischer Studien finden sich bei Rost (2013).

[1] Das Ziel, Forschungsergebnisse zu einzelnen Themengebieten zusammenzutragen und aufzubereiten, verfolgt auch das „Clearing House Unterricht", das sich als Vermittler zwischen Forschung und schulischer Praxis versteht (▶ https://www.clearinghouse.edu.tum.de/).

[2] Der Artikel lässt sich kostenfrei aus den Netzwerken der meisten Universitäten abrufen (z. B. unter https://doi.org/10.1024/1010-0652/a000003).



Zeitschrift für Pädagogische Psychologie, 24 (1), 2010, 39–49

Originalartikel

# Steigert Kaugummikauen das kognitive Leistungsvermögen?

## Zwei Experimente der besonderen Art

Detlef H. Rost, Linda Wirthwein, Kristina Frey und Elvira Becker

Fachbereich Psychologie, Philipps-Universität Marburg

**Zusammenfassung.** Aufgrund weniger und methodisch verbesserungsbedürftiger Studien mit kleinen Stichproben wird behauptet, Kaugummikauen wirke sich positiv auf Lernleistungen, Konzentration und Gedächtnis aus. Mit zwei randomisierten Experimenten ($N_1$ = 548 bzw. $N_2$ = 504 Kinder; 5. und 6. Jahrgangsstufen) wurde überprüft, ob «Kauer» in ausgewählten Intelligenzsubtests (Experiment I) und bei Aufmerksamkeits- bzw. Konzentrations- sowie Gedächtnisaufgaben (Experiment II) bessere Leistungen erbringen als «Nichtkauer». Statistisch signifikante (aber nur kleine) Unterschiede resultierten lediglich bei zwei Konzentrationskennwerten sowie bei Aufgaben zur Erfassung von kurzfristigem Behalten zugunsten der «Nichtkauer». Die vorliegenden Ergebnisse verdeutlichen den Stellenwert einer methodisch adäquaten Replikation von an kleinen Stichproben gewonnenen und vorschnell popularisierten Befunden.

**Schlüsselwörter:** Intelligenz, Aufmerksamkeit, Konzentration, Kaugummi, Replikation

**Does Chewing Gum Enhance Cognitive Abilities? Two Rather Extraordinary Experiments**

**Abstract.** Based on a few methodically weak studies with small samples, the mass media frequently claims that chewing gum has a positive impact on learning efficiency, concentration, and memory. Two randomized experiments ($N_1$ = 548 and $N_2$ = 504 fifth and sixth graders) were run to check whether «chewers» outperform «non-chewers» on selected intelligence subtests (experiment I) or on tasks assessing concentration, attention, and memory (experiment II). Statistically significant (but small) differences were only found on two concentration-indicators as well as on short-term memory tasks. In both cases «non-chewers» outperformed «chewers». The discussion stresses the problem of rashly spreading non-replicated findings which are based on methodically weak studies.

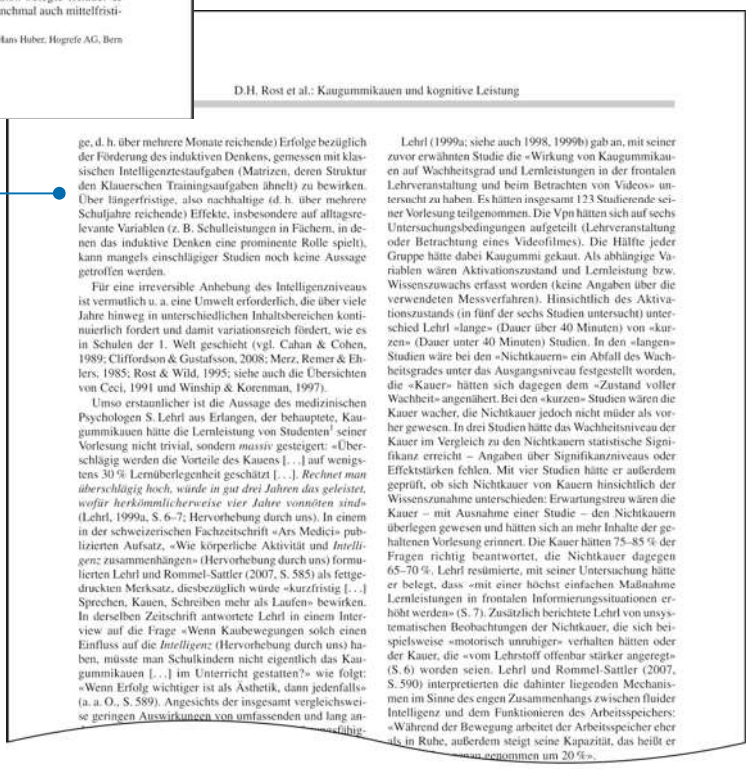
**Keywords:** intelligence, attention, concentration, chewing gum, replication

### 1 Ausgangslage

Der Wunsch, die eigene kognitive Leistungsfähigkeit, insbesondere Aufmerksamkeits-, Gedächtnis- und Lernleistungen, nachhaltig zu steigern, ist vermutlich so alt wie die Menschheit selbst. Immer wieder wurde versucht, den «ewige[n] Menschheitstraum, sich über die Grenzen der eigenen Intelligenz und des eigenen Lernvermögens hinwegzusetzen» (Degen, 2003, S. 157) durch Trainingsprogramme zu verwirklichen. Systematische Versuche die kognitive Leistungsfähigkeit durch kurz- und mittelfristige Programme zu fördern, waren jedoch nur in bescheidenem Rahmen erfolgreich, so dass Weinert (1994, S. 272) resümierte: «Intelligenz lässt sich durch formal angelegte Trainingsprogramme weder erlernen noch nachhaltig verbessern.» Diese Skepsis betrifft auch längerfristige USA-Programme zur Förderung benachteiligter Vorschulkinder (vgl. Barnett, 2004; Gottfredson, 2000; Herrnstein & Murray, 1994; Jensen, 1998; Locurto, 1991; Spitz, 1991, 1999a,b; The Consortium for Longitudinal Studies, 1983; Zigler & Styfco, 2004). Außerordentlich intensive Vorschulprogramme, in denen man wenige Kinder über mehrere Jahre hinweg 50 Wochen pro Jahr täglich sechs bis neun Stunden in sehr kleinen Gruppen betreut hatte und bei denen die intensive Betreuung im Grundschulalter fortgesetzt worden war, zeitigten nachhaltige positive Effekte (rund $d$ = .30) auf die kognitive Leistungsfähigkeit (vgl. Campbell, Ramey, Pungello, Sparling & Miller-Johnson, 2002; Ramey & Ramey, 2004). Wesentlich kürzer sind die Klauerschen Denktrainings angelegt (z. B. 1993). Obwohl in der Regel nicht mehr als 20 Sitzungen umfassend, scheinen sie jedoch, wie eine Metaanalyse über 74 (zumeist quasi-experimentelle) Trainingsstudien belegte (Klauer & Phye, 2008), kurzfristige (und manchmal auch mittelfristi-

DOI 10.1024/1010-0652/a000003







ge, d. h. über mehrere Monate reichende) Erfolge bezüglich der Förderung des induktiven Denkens, gemessen mit klassischen Intelligenztestaufgaben (Matrizen, deren Struktur den Klauerschen Trainingsaufgaben ähnelt) zu bewirken. Über längerfristige, also nachhaltige (d. h. über mehrere Schuljahre reichende) Effekte, insbesondere auf alltagsrelevante Variablen (z. B. Schulleistungen in Fächern, in denen das induktive Denken eine prominente Rolle spielt), kann mangels einschlägiger Studien noch keine Aussage getroffen werden.

Für eine irreversible Anhebung des Intelligenzniveaus ist vermutlich u. a. eine Umwelt erforderlich, die über viele Jahre hinweg in unterschiedlichen Inhaltsbereichen kontinuierlich fordert und damit variationsreich fördert, wie es in Schulen der 1. Welt geschieht (vgl. Cahan & Cohen, 1989; Cliffordson & Gustafsson, 2008; Merz, Remer & Ehlers, 1985; Rost & Wild, 1995; siehe auch die Übersichten von Ceci, 1991 und Winship & Korenman, 1997).

Umso erstaunlicher ist die Aussage des medizinischen Psychologen S. Lehrl aus Erlangen, der behauptete, Kaugummikauen hätte die Lernleistung von Studenten[1] seiner Vorlesung nicht trivial, sondern *massiv* gesteigert: «Überschlägig werden die Vorteile des Kauens [. . .] auf wenigstens 30 % Lernüberlegenheit geschätzt [. . .]. *Rechnet man überschlägig hoch, würde in gut drei Jahren das geleistet, wofür herkömmlicherweise vier Jahre vonnöten sind*» (Lehrl, 1999a, S. 6–7; Hervorhebung durch uns). In einem in der schweizerischen Fachzeitschrift «Ars Medici» publizierten Aufsatz, «Wie körperliche Aktivität und *Intelligenz* zusammenhängen» (Hervorhebung durch uns) formulierten Lehrl und Rommel-Sattler (2007, S. 585) als fettgedruckten Merksatz, diesbezüglich würde «kurzfristig [. . .] Sprechen, Kauen, Schreiben mehr als Laufen» bewirken. In derselben Zeitschrift antwortete Lehrl in einem Interview auf die Frage «Wenn Kaubewegungen solch einen Einfluss auf die *Intelligenz* (Hervorhebung durch uns) haben, müsste man Schulkindern nicht eigentlich das Kaugummikauen [. . .] im Unterricht gestatten?» wie folgt: «Wenn Erfolg wichtiger ist als Ästhetik, dann jedenfalls» (a. a. O., S. 589). Angesichts der insgesamt vergleichsweise geringen Auswirkungen von umfassenden und lang an-

Lehrl (1999a; siehe auch 1998, 1999b) gab an, mit seiner zuvor erwähnten Studie die «Wirkung von Kaugummikauen auf Wachheitsgrad und Lernleistungen in der frontalen Lehrveranstaltung und beim Betrachten von Videos» untersucht zu haben. Es hätten insgesamt 123 Studierende seiner Vorlesung teilgenommen. Die Vpn hätten sich auf sechs Untersuchungsbedingungen aufgeteilt (Lehrveranstaltung oder Betrachtung eines Videofilmes). Die Hälfte jeder Gruppe hätte dabei Kaugummi gekaut. Als abhängige Variablen wären Aktivationszustand und Lernleistung bzw. Wissenszuwachs erfasst worden (keine Angaben über die verwendeten Messverfahren). Hinsichtlich des Aktivationszustands (in fünf der sechs Studien untersucht) unterschied Lehrl «lange» (Dauer über 40 Minuten) von «kurzen» (Dauer unter 40 Minuten) Studien. In den «langen» Studien wäre bei den «Nichtkauern» ein Abfall des Wachheitsgrades unter das Ausgangsniveau festgestellt worden, die «Kauer» hätten sich dagegen dem «Zustand voller Wachheit» angenähert. Bei den «kurzen» Studien wären die Kauer wacher, die Nichtkauer jedoch nicht müder als vorher gewesen. In drei Studien hätte das Wachheitsniveau der Kauer im Vergleich zu den Nichtkauern statistische Signifikanz erreicht – Angaben über Signifikanzniveaus oder Effektstärken fehlen. Mit vier Studien hätte er außerdem geprüft, ob sich Nichtkauer von Kauern hinsichtlich der Wissenszunahme unterschieden: Erwartungstreu wären die Kauer – mit Ausnahme einer Studie – den Nichtkauern überlegen gewesen und hätten sich an mehr Inhalte der gehaltenen Vorlesung erinnert. Die Kauer hätten 75–85 % der Fragen richtig beantwortet, die Nichtkauer dagegen 65–70 %. Lehrl resümierte, mit seiner Untersuchung hätte er belegt, dass «mit einer höchst einfachen Maßnahme Lernleistungen in frontalen Informierungssituationen erhöht werden» (S. 7). Zusätzlich berichtete Lehrl von unsystematischen Beobachtungen der Nichtkauer, die sich beispielsweise «motorisch unruhiger» verhalten hätten oder der Kauer, die «vom Lehrstoff offenbar stärker angeregt» (S. 6) worden seien. Lehrl und Rommel-Sattler (2007, S. 590) interpretierten die dahinter liegenden Mechanismen im Sinne des engen Zusammenhangs zwischen fluider Intelligenz und dem Funktionieren des Arbeitsspeichers: «Während der Bewegung arbeitet der Arbeitsspeicher eher als in Ruhe, außerdem steigt seine Kapazität, das heißt er

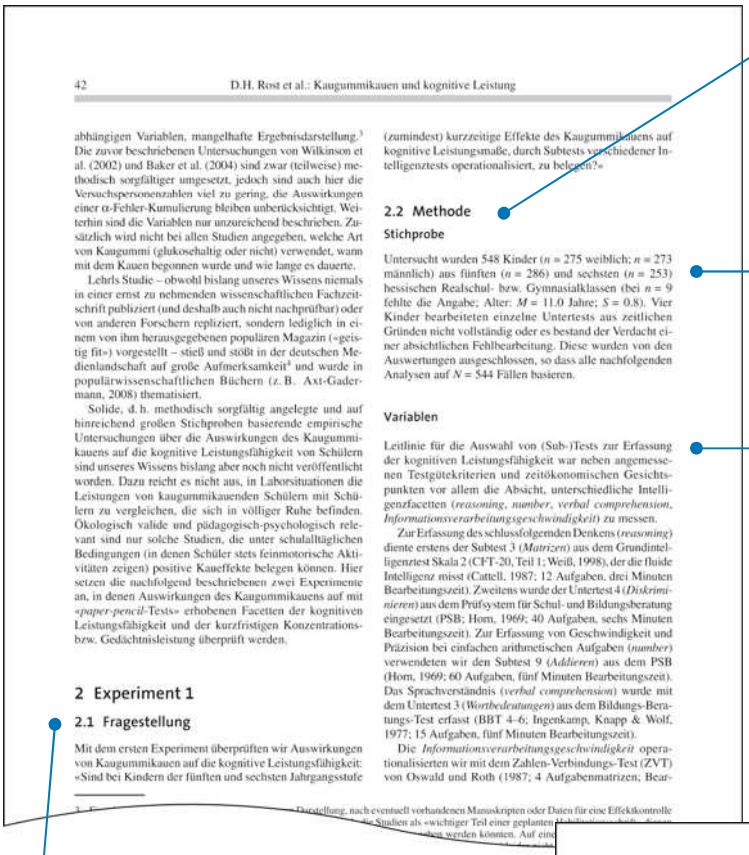
**Titel der Fachzeitschrift**
Renommierte Fachzeitschriften mit Peerreview haben die strengste Qualitätskontrolle und sind daher erste Wahl bei der Suche nach Forschungsergebnissen. Daher sollte beachtet werden, in welcher Form und wo eine Arbeit erschienen ist (vgl. ◘ Tab. 27.4).

**Titel des Artikels**
Der Titel sagt in verdichteter Form, um was es geht; ihn mehrfach genau zu lesen empfiehlt sich.
Bei dieser Arbeit lässt sich z. B. bereits dem Titel entnehmen, dass die Fragestellung mit einem experimentellen (also intern sehr validen) Untersuchungsdesign geklärt wurde.

**Zusammenfassung (auch „Abstract")**
Dieser Textteil enthält – sehr verdichtet – die wichtigsten Informationen über die vorliegende Studie. Dazu zählen Fragestellung, Stichprobe, Untersuchungsdesign, erfasste Konstrukte und Ergebnisse. Da diese Textstelle sehr komprimiert die wesentlichen Aspekte zusammenfasst, ist sie nicht ganz einfach zu verstehen. Sie sollten sie sehr genau (evtl. mehrfach) lesen, da Sie daraus bereits wesentliche Informationen gewinnen können.
Der Zusammenfassung dieser Arbeit lässt sich z. B. bereits das zentrale Ergebnis entnehmen, nämlich dass Kaugummikauen – entgegen dem populären Mythos – keine nachweisbaren Vorteile für Konzentrations- und Gedächtnisleistungen brachte.
Liegen mehrere Studien zu einem Thema vor, ist die Lektüre ihrer Zusammenfassungen eine effektive Strategie, sich einen Überblick über ein Forschungsgebiet zu verschaffen.

**Theoretischer Teil (auch „Einleitung", „Theoretischer Hintergrund", „Ausgangslage")**
Bevor eine Studie in Angriff genommen werden kann, muss das zu untersuchende Phänomen theoretisch präzisiert und die dazu bereits publizierte Forschungsliteratur (der „Forschungsstand") aufgearbeitet werden. Von großer Bedeutung für die Qualität der Arbeit ist, dass präzise theoretische Überlegungen zu dem untersuchten Phänomen angestellt und daraus genaue Hypothesen abgeleitet werden. Alle wichtigen Argumente sollten dabei durch Quellen belegt werden.
Dies erfolgt im ersten großen Textteil (meist mehrere Seiten). Hier stößt man oft auf unbekannte Begriffe, deren Bedeutung nicht auf Anhieb verstanden wird. Davon sollten Sie sich nicht entmutigen lassen, Sie werden die Studie in aller Regel trotzdem verstehen.
Die Autorinnen und Autoren dieser Studie haben über den Zusammenhang von Kaugummikauen und kognitiven Leistungen recherchiert. Sie vergleichen und bewerten die Studien dazu. Damit dokumentieren sie den Stand der Forschung. Durch die sorgfältige Einschätzung der Forschungssituation werden offene Forschungsfragen deutlich („Forschungsdesiderata"). Diesen kann sich nun eine eigene Untersuchung widmen.

**◘ Abb. 27.11** Wie liest man einen psychologischen Originalartikel? Eine Lesehilfe



42 D.H. Rost et al.: Kaugummikauen und kognitive Leistung

abhängigen Variablen, mangelhafte Ergebnisdarstellung.[3] Die zuvor beschriebenen Untersuchungen von Wilkinson et al. (2002) und Baker et al. (2004) sind zwar (teilweise) methodisch sorgfältiger umgesetzt, jedoch sind auch hier die Versuchspersonenzahlen viel zu gering, die Auswirkungen einer α-Fehler-Kumulierung bleiben unberücksichtigt. Weiterhin sind die Variablen nur unzureichend beschrieben. Zusätzlich wird nicht bei allen Studien angegeben, welche Art von Kaugummi (glukosehaltig oder nicht) verwendet, wann mit dem Kauen begonnen wurde und wie lange es dauerte.

Lehrls Studie – obwohl bislang unseres Wissens niemals in einer ernst zu nehmenden wissenschaftlichen Fachzeitschrift publiziert (und deshalb auch nicht nachprüfbar) oder von anderen Forschern repliziert, sondern lediglich in einem von ihm herausgegebenen populären Magazin («geistig fit») vorgestellt – stieß und stößt in der deutschen Medienlandschaft auf große Aufmerksamkeit[4] und wurde in populärwissenschaftlichen Büchern (z. B. Axt-Gadermann, 2008) thematisiert.

Solide, d. h. methodisch sorgfältig angelegte und auf hinreichend großen Stichproben basierende empirische Untersuchungen über die Auswirkungen des Kaugummikauens auf die kognitive Leistungsfähigkeit von Schülern sind unseres Wissens bislang aber noch nicht veröffentlicht worden. Dazu reicht es nicht aus, in Laborsituationen die Leistungen von kaugummikauenden Schülern mit Schülern zu vergleichen, die sich in völliger Ruhe befinden. Ökologisch valide und pädagogisch-psychologisch relevant sind nur solche Studien, die unter schulalltäglichen Bedingungen (in denen Schüler stets feinmotorische Aktivitäten zeigen) positive Kaueffekte belegen können. Hier setzen die nachfolgend beschriebenen zwei Experimente an, in denen Auswirkungen des Kaugummikauens auf mit «*paper-pencil*-Tests» erhobenen Facetten der kognitiven Leistungsfähigkeit und der kurzfristigen Konzentrations- bzw. Gedächtnisleistung überprüft werden.

## 2 Experiment 1

### 2.1 Fragestellung

Mit dem ersten Experiment überprüften wir Auswirkungen von Kaugummikauen auf die kognitive Leistungsfähigkeit: «Sind bei Kindern der fünften und sechsten Jahrgangsstufe (zumindest) kurzzeitige Effekte des Kaugummikauens auf kognitive Leistungsmaße, durch Subtests verschiedener Intelligenztests operationalisiert, zu belegen?»

### 2.2 Methode

#### Stichprobe

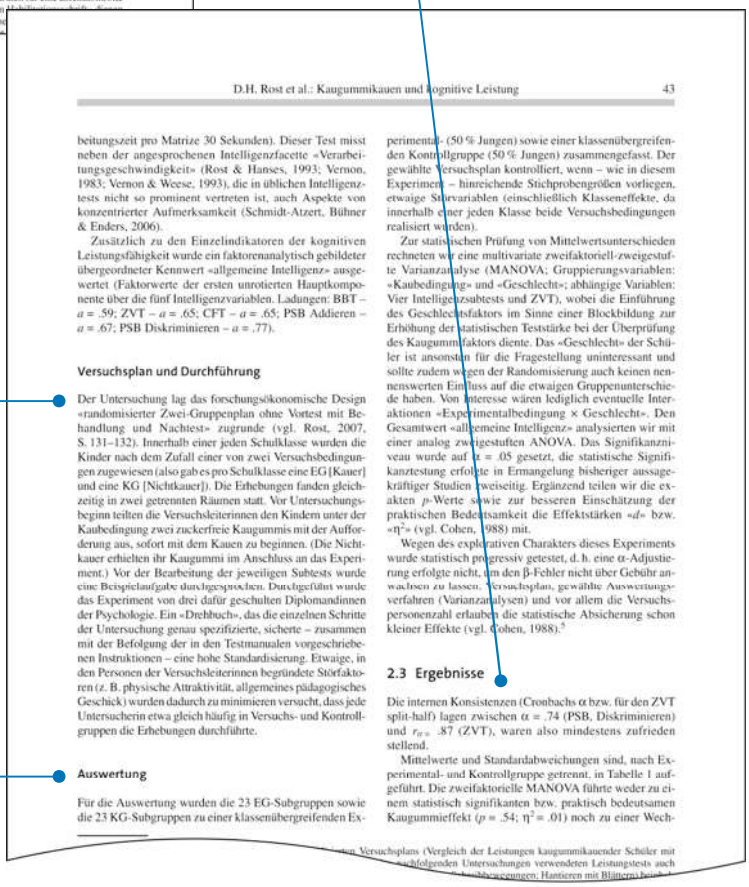
Untersucht wurden 548 Kinder ($n$ = 275 weiblich; $n$ = 273 männlich) aus fünften ($n$ = 286) und sechsten ($n$ = 253) hessischen Realschul- bzw. Gymnasialklassen (bei $n$ = 9 fehlte die Angabe; Alter: $M$ = 11.0 Jahre; $S$ = 0.8). Vier Kinder bearbeiteten einzelne Untertests aus zeitlichen Gründen nicht vollständig oder es bestand der Verdacht einer absichtlichen Fehlbearbeitung. Diese wurden von den Auswertungen ausgeschlossen, so dass alle nachfolgenden Analysen auf $N$ = 544 Fällen basieren.

#### Variablen

Leitlinie für die Auswahl von (Sub-)Tests zur Erfassung der kognitiven Leistungsfähigkeit war neben angemessenen Testgütekriterien und zeitökonomischen Gesichtspunkten vor allem die Absicht, unterschiedliche Intelligenzfacetten (*reasoning, number, verbal comprehension, Informationsverarbeitungsgeschwindigkeit*) zu messen.

Zur Erfassung des schlussfolgernden Denkens (*reasoning*) diente erstens der Subtest 3 (*Matrizen*) aus dem Grundintelligenztest Skala 2 (CFT-20, Teil 1; Weiß, 1998), der die fluide Intelligenz misst (Cattell, 1987; 12 Aufgaben, drei Minuten Bearbeitungszeit). Zweitens wurde der Untertest 4 (*Diskriminieren*) aus dem Prüfsystem für Schul- und Bildungsberatung eingesetzt (PSB; Horn, 1969; 40 Aufgaben, sechs Minuten Bearbeitungszeit). Zur Erfassung von Geschwindigkeit und Präzision bei einfachen arithmetischen Aufgaben (*number*) verwendeten wir den Subtest 9 (*Addieren*) aus dem PSB (Horn, 1969; 60 Aufgaben, fünf Minuten Bearbeitungszeit). Das Sprachverständnis (*verbal comprehension*) wurde mit dem Untertest 3 (*Wortbedeutungen*) aus dem Bildungs-Beratungs-Test erfasst (BBT 4–6; Ingenkamp, Knapp & Wolf, 1977; 15 Aufgaben, fünf Minuten Bearbeitungszeit).

Die *Informationsverarbeitungsgeschwindigkeit* operationalisierten wir mit dem Zahlen-Verbindungs-Test (ZVT) von Oswald und Roth (1987; 4 Aufgabenmatrizen; Bear-



D.H. Rost et al.: Kaugummikauen und kognitive Leistung 43

beitungszeit pro Matrize 30 Sekunden). Dieser Test misst neben der angesprochenen Intelligenzfacette «Verarbeitungsgeschwindigkeit» (Rost & Hanses, 1993; Vernon, 1983; Vernon & Weese, 1993), die in üblichen Intelligenztests nicht so prominent vertreten ist, auch Aspekte von konzentrierter Aufmerksamkeit (Schmidt-Atzert, Bühner & Enders, 2006).

Zusätzlich zu den Einzelindikatoren der kognitiven Leistungsfähigkeit wurde ein faktorenanalytisch gebildeter übergeordneter Kennwert «allgemeine Intelligenz» ausgewertet (Faktorwerte der ersten unrotierten Hauptkomponente über die fünf Intelligenzvariablen. Ladungen: BBT – $a$ = .59; ZVT – $a$ = .65; CFT – $a$ = .65; PSB Addieren – $a$ = .67; PSB Diskriminieren – $a$ = .77).

#### Versuchsplan und Durchführung

Der Untersuchung lag das forschungsökonomische Design «randomisierter Zwei-Gruppenplan ohne Vortest mit Behandlung und Nachtest» zugrunde (vgl. Rost, 2007, S. 131–132). Innerhalb einer jeden Schulklasse wurden die Kinder nach dem Zufall einer von zwei Versuchsbedingungen zugewiesen (also gab es pro Schulklasse eine EG [Kauer] und eine KG [Nichtkauer]). Die Erhebungen fanden gleichzeitig in zwei getrennten Räumen statt. Vor Untersuchungsbeginn teilten die Versuchsleiterinnen den Kindern unter der Kaubedingung zwei zuckerfreie Kaugummis mit der Aufforderung aus, sofort mit dem Kauen zu beginnen. (Die Nichtkauer erhielten ihr Kaugummi im Anschluss an das Experiment.) Vor der Bearbeitung der jeweiligen Subtests wurde eine Beispielaufgabe durchgesprochen. Durchgeführt wurde das Experiment von drei dafür geschulten Diplomandinnen der Psychologie. Ein «Drehbuch», das die einzelnen Schritte der Untersuchung genau spezifizierte, sicherte – zusammen mit der Befolgung der in den Testmanualen vorgeschriebenen Instruktionen – eine hohe Standardisierung. Etwaige, in den Personen der Versuchsleiterinnen begründete Störfaktoren (z. B. physische Attraktivität, allgemeines pädagogisches Geschick) wurden dadurch zu minimieren versucht, dass jede Untersucherin etwa gleich häufig in Versuchs- und Kontrollgruppen die Erhebungen durchführte.

#### Auswertung

Für die Auswertung wurden die 23 EG-Subgruppen sowie die 23 KG-Subgruppen zu einer klassenübergreifenden Experimental- (50 % Jungen) sowie einer klassenübergreifenden Kontrollgruppe (50 % Jungen) zusammengefasst. Der gewählte Versuchsplan kontrolliert, wenn – wie in diesem Experiment – hinreichende Stichprobengrößen vorliegen, etwaige Störvariablen (einschließlich Klasseneffekte, da innerhalb einer jeden Klasse beide Versuchsbedingungen realisiert wurden).

Zur statistischen Prüfung von Mittelwertsunterschieden rechneten wir eine multivariate zweifaktoriell-zweigestufte Varianzanalyse (MANOVA; Gruppierungsvariablen: «Kaubedingung» und «Geschlecht»; abhängige Variablen: Vier Intelligenzsubtests und ZVT), wobei die Einführung des Geschlechtsfaktors im Sinne einer Blockbildung zur Erhöhung der statistischen Teststärke bei der Überprüfung des Kaugummifaktors diente. Das «Geschlecht» der Schüler ist ansonsten für die Fragestellung uninteressant und sollte zudem wegen der Randomisierung auch keinen nennenswerten Einfluss auf die etwaigen Gruppenunterschiede haben. Von Interesse wären lediglich eventuelle Interaktionen «Experimentalbedingung × Geschlecht». Den Gesamtwert «allgemeine Intelligenz» analysierten wir mit einer analog zweigestuften ANOVA. Das Signifikanzniveau wurde auf $\alpha$ = .05 gesetzt, die statistische Signifikanztestung erfolgte in Ermangelung bisheriger aussagekräftiger Studien zweiseitig. Ergänzend teilen wir die exakten $p$-Werte sowie zur besseren Einschätzung der praktischen Bedeutsamkeit die Effektstärken «$d$» bzw. «$\eta^2$» (vgl. Cohen, 1988) mit.

Wegen des explorativen Charakters dieses Experiments wurde statistisch progressiv getestet, d. h. eine α-Adjustierung erfolgte nicht, um den β-Fehler nicht über Gebühr anwachsen zu lassen. Versuchsplan, gewählte Auswertungsverfahren (Varianzanalysen) und vor allem die Versuchspersonenzahl erlauben die statistische Absicherung schon kleiner Effekte (vgl. Cohen, 1988).[5]

### 2.3 Ergebnisse

Die internen Konsistenzen (Cronbachs α bzw. für den ZVT split-half) lagen zwischen α = .74 (PSB, Diskriminieren) und $r_{tt}$ = .87 (ZVT), waren also mindestens zufrieden stellend.

Mittelwerte und Standardabweichungen sind, nach Experimental- und Kontrollgruppe getrennt, in Tabelle 1 aufgeführt. Die zweifaktorielle MANOVA führte weder zu einem statistisch signifikanten bzw. praktisch bedeutsamen Kaugummieffekt ($p$ = .54; $\eta^2$ = .01) noch zu einer Wech-

---

**Methodenteil**
Im Methodenteil finden sich Informationen zu Stichprobe, erfassten Variablen, Untersuchungsdesign (Versuchsplan) und Datenauswertung. Ein Qualitätsmerkmal ist die detaillierte Beschreibung des methodischen Vorgehens samt der eingesetzten Materialien. Nur so kann die Studie problemlos nachvollzogen und hinsichtlich interner und externer Validität sowie der Güte der Messinstrumente beurteilt werden.

**Stichprobenbeschreibung**
Anhand von Größe und Zusammensetzung der Stichprobe lässt sich beurteilen, wie spezifisch die Ergebnisse für die untersuchte Gruppe sind und wie gut sie sich auf andere Gruppen generalisieren lassen.

**Messinstrumente**
In dieser Studie werden die genutzten Erhebungsinstrumente unter der Überschrift „Variablen" vorgestellt. Die Entscheidung fiel hier auf etablierte Verfahren, was weniger Fragen zur Messgüte aufwirft als ad-hoc konstruierte Instrumente.

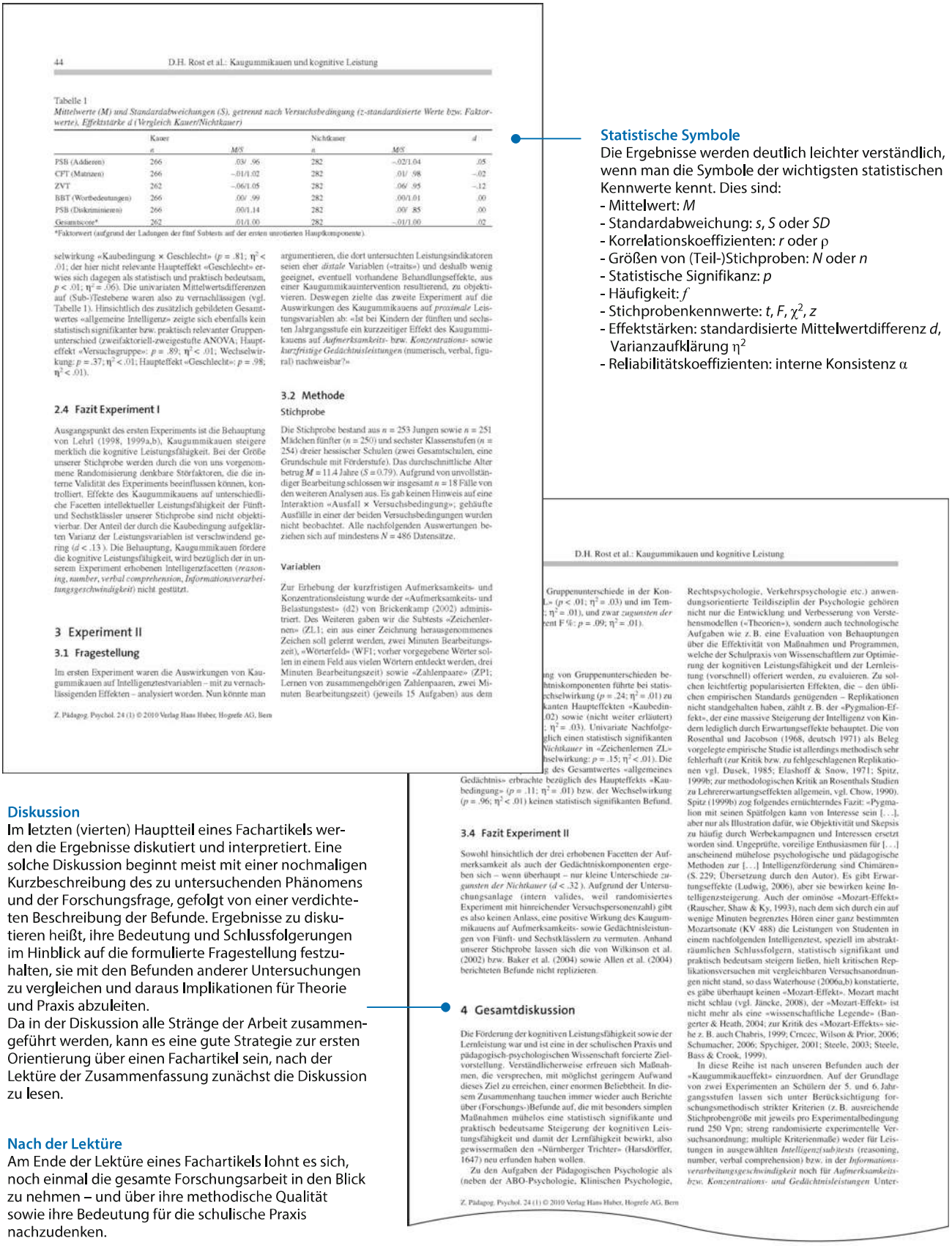
**Ergebnisteil**
Im dritten großen Teil werden die empirischen Ergebnisse referiert. Um sie zu verstehen ist es nicht zentral, dass Sie jedes statistische Verfahren kennen. Hilfreich ist hier eine gewisse „Unempfindlichkeit", die es ermöglicht, zunächst die zentralen Aussagen nachvollziehen zu können. Offene Fragen können dann im Anschluss nachgeschlagen werden.
In dieser Studie werden zunächst Maße für die Reliabilität der Messverfahren berichtet, anschließend wird auf deskriptive statistische Kennwerte verwiesen.

**Fragestellung/Hypothesen (oft am Ende des theoretischen Teils)**
Dreh- und Angelpunkt jeder empirischen Studie ist die Fragestellung mit den zu testenden Hypothesen. Die Anlage der Untersuchung – d. h. ihr Design, die untersuchten Konstrukte sowie deren Operationalisierung – muss darauf abgestimmt sein.
In Experiment 1 dieser Arbeit ist die Hypothese so weit präzisiert, dass sie empirisch überprüft werden kann.

**Untersuchungsdesign**
Rost et al. (2010) haben sich für ein experimentelles Design entschieden, da die kausale Wirkung des Kaugummikauens (unabhängige Variable) auf die kognitive Leistungsfähigkeit (abhängige Variable) untersucht werden sollte.

**Analysemethodik**
Oft schließt der Methodenteil mit der Beschreibung der Datenauswertung ab (häufig ist dies auch in die Ergebnisdarstellung integriert). Um diesen Teil der Studie kritisch überprüfen zu können, ist wohl sehr umfangreiches statistisches Wissen nötig – hier z. B. über Varianzanalysen (ANOVA = Analysis of Variance, MANOVA = multivariate ANOVA). Zum Verständnis der Ergebnisse einer Studie ist dies jedoch nicht zwingend erforderlich.

**Abb. 27.11** (Fortsetzung)



44 D.H. Rost et al.: Kaugummikauen und kognitive Leistung

Tabelle 1
*Mittelwerte (M) und Standardabweichungen (S), getrennt nach Versuchsbedingung (z-standardisierte Werte bzw. Faktorwerte), Effektstärke d (Vergleich Kauer/Nichtkauer)*

| | Kauer | | Nichtkauer | | d |
|---|---|---|---|---|---|
| | n | M/S | n | M/S | |
| PSB (Addieren) | 266 | .03/ .96 | 282 | –.02/1.04 | .05 |
| CFT (Matrizen) | 266 | –.01/1.02 | 282 | .01/ .98 | –.02 |
| ZVT | 262 | –.06/1.05 | 282 | .06/ .95 | –.12 |
| BBT (Wortbedeutungen) | 266 | .00/ .99 | 282 | .00/1.01 | .00 |
| PSB (Diskriminieren) | 266 | .00/1.14 | 282 | .00/ .85 | .00 |
| Gesamtscore* | 262 | .01/1.00 | 282 | –.01/1.00 | .02 |

*Faktorwert (aufgrund der Ladungen der fünf Subtests auf der ersten unrotierten Hauptkomponente).

selwirkung «Kaubedingung × Geschlecht» ($p = .81$; $\eta^2 < .01$; der hier nicht relevante Haupteffekt «Geschlecht» erwies sich dagegen als statistisch und praktisch bedeutsam, $p < .01$; $\eta^2 = .06$). Die univariaten Mittelwertsdifferenzen auf (Sub-)Testebene waren also zu vernachlässigen (vgl. Tabelle 1). Hinsichtlich des zusätzlich gebildeten Gesamtwertes «allgemeine Intelligenz» zeigte sich ebenfalls kein statistisch signifikanter bzw. praktisch relevanter Gruppenunterschied (zweifaktoriell-zweigestufte ANOVA; Haupteffekt «Versuchsgruppe»: $p = .89$; $\eta^2 < .01$; Wechselwirkung: $p = .37$; $\eta^2 < .01$; Haupteffekt «Geschlecht»: $p = .98$; $\eta^2 < .01$).

### 2.4 Fazit Experiment I

Ausgangspunkt des ersten Experiments ist die Behauptung von Lehrl (1998, 1999a,b), Kaugummikauen steigere merklich die kognitive Leistungsfähigkeit. Bei der Größe unserer Stichprobe werden durch die von uns vorgenommene Randomisierung denkbare Störfaktoren, die die interne Validität des Experiments beeinflussen können, kontrolliert. Effekte des Kaugummikauens auf unterschiedliche Facetten intellektueller Leistungsfähigkeit der Fünft- und Sechstklässler unserer Stichprobe sind nicht objektivierbar. Der Anteil der durch die Kaubedingung aufgeklärten Varianz der Leistungsvariablen ist verschwindend gering ($d < .13$). Die Behauptung, Kaugummikauen fördere die kognitive Leistungsfähigkeit, wird bezüglich der in unserem Experiment erhobenen Intelligenzfacetten (*reasoning, number, verbal comprehension, Informationsverarbeitungsgeschwindigkeit*) nicht gestützt.

## 3 Experiment II

### 3.1 Fragestellung

Im ersten Experiment waren die Auswirkungen von Kaugummikauen auf Intelligenztestvariablen – mit zu vernachlässigenden Effekten – analysiert worden. Nun könnte man argumentieren, die dort untersuchten Leistungsindikatoren seien eher *distale* Variablen («traits») und deshalb wenig geeignet, eventuell vorhandene Behandlungseffekte, aus einer Kaugummikauintervention resultierend, zu objektivieren. Deswegen zielte das zweite Experiment auf die Auswirkungen des Kaugummikauens auf *proximale* Leistungsvariablen ab: «Ist bei Kindern der fünften und sechsten Jahrgangsstufe ein kurzzeitiger Effekt des Kaugummikauens auf *Aufmerksamkeits-* bzw. *Konzentrations-* sowie *kurzfristige Gedächtnisleistungen* (numerisch, verbal, figural) nachweisbar?»

### 3.2 Methode

#### Stichprobe

Die Stichprobe bestand aus $n = 253$ Jungen sowie $n = 251$ Mädchen fünfter ($n = 250$) und sechster Klassenstufen ($n = 254$) dreier hessischer Schulen (zwei Gesamtschulen, eine Grundschule mit Förderstufe). Das durchschnittliche Alter betrug $M = 11.4$ Jahre ($S = 0.79$). Aufgrund von unvollständiger Bearbeitung schlossen wir insgesamt $n = 18$ Fälle von den weiteren Analysen aus. Es gab keinen Hinweis auf eine Interaktion «Ausfall × Versuchsbedingung»; gehäufte Ausfälle in einer der beiden Versuchsbedingungen wurden nicht beobachtet. Alle nachfolgenden Auswertungen beziehen sich auf mindestens $N = 486$ Datensätze.

#### Variablen

Zur Erhebung der kurzfristigen Aufmerksamkeits- und Konzentrationsleistung wurde der «Aufmerksamkeits- und Belastungstest» (d2) von Brickenkamp (2002) administriert. Des Weiteren gaben wir die Subtests «Zeichenlernen» (ZL1; ein aus einer Zeichnung herausgenommenes Zeichen soll gelernt werden, zwei Minuten Bearbeitungszeit), «Wörterfeld» (WF1; vorher vorgegebene Wörter sollen in einem Feld aus vielen Wörtern entdeckt werden, drei Minuten Bearbeitungszeit) sowie «Zahlenpaare» (ZP1; Lernen von zusammengehörigen Zahlenpaaren, zwei Minuten Bearbeitungszeit) (jeweils 15 Aufgaben) aus dem

Z. Pädagog. Psychol. 24 (1) © 2010 Verlag Hans Huber, Hogrefe AG, Bern

D.H. Rost et al.: Kaugummikauen und kognitive Leistung

Gruppenunterschiede in der Kon-
L» ($p < .01$; $\eta^2 = .03$) und im Tem-
; $\eta^2 = .01$), und zwar *zugunsten der*
ent F %: $p = .09$; $\eta^2 = .01$).

ng von Gruppenunterschieden be-
htniskomponenten führte bei statis-
schselwirkung ($p = .24$; $\eta^2 = .01$) zu
kanten Haupteffekten «Kaubedin-
.02) sowie (nicht weiter erläutert)
; $\eta^2 = .03$). Univariate Nachfolge-
glich einen statistisch signifikanten
*Nichtkauer* in «Zeichenlernen ZL»
hselwirkung: $p = .15$; $\eta^2 < .01$). Die
g des Gesamtwertes «allgemeines
Gedächtnis» erbrachte bezüglich des Haupteffekts «Kaubedingung» ($p = .11$; $\eta^2 = .01$) bzw. der Wechselwirkung ($p = .96$; $\eta^2 < .01$) keinen statistisch signifikanten Befund.

### 3.4 Fazit Experiment II

Sowohl hinsichtlich der drei erhobenen Facetten der Aufmerksamkeit als auch der Gedächtniskomponenten ergeben sich – wenn überhaupt – nur kleine Unterschiede *zugunsten der Nichtkauer* ($d < .32$). Aufgrund der Untersuchungsanlage (intern valides, weil randomisiertes Experiment mit hinreichender Versuchspersonenzahl) gibt es also keinen Anlass, eine positive Wirkung des Kaugummikauens auf Aufmerksamkeits- sowie Gedächtnisleistungen von Fünft- und Sechstklässlern zu vermuten. Anhand unserer Stichprobe lassen sich die von Wilkinson et al. (2002) bzw. Baker et al. (2004) sowie Allen et al. (2004) berichteten Befunde nicht replizieren.

## 4 Gesamtdiskussion

Die Förderung der kognitiven Leistungsfähigkeit sowie der Lernleistung war und ist eine in der schulischen Praxis und pädagogisch-psychologischen Wissenschaft forcierte Zielvorstellung. Verständlicherweise erfreuen sich Maßnahmen, die versprechen, mit möglichst geringem Aufwand dieses Ziel zu erreichen, einer enormen Beliebtheit. In diesem Zusammenhang tauchen immer wieder auch Berichte über (Forschungs-)Befunde auf, die mit besonders simplen Maßnahmen mühelos eine statistisch signifikante und praktisch bedeutsame Steigerung der kognitiven Leistungsfähigkeit und damit der Lernfähigkeit bewirkt, also gewissermaßen den «Nürnberger Trichter» (Harsdörffer, 1647) neu erfunden haben wollen.

Zu den Aufgaben der Pädagogischen Psychologie als (neben der ABO-Psychologie, Klinischen Psychologie, Rechtspsychologie, Verkehrspsychologie etc.) anwendungsorientierte Teildisziplin der Psychologie gehören nicht nur die Entwicklung und Verbesserung von Verstehensmodellen («Theorien»), sondern auch technologische Aufgaben wie z. B. eine Evaluation von Behauptungen über die Effektivität von Maßnahmen und Programmen, welche der Schulpraxis von Wissenschaftlern zur Optimierung der kognitiven Leistungsfähigkeit und der Lernleistung (vorschnell) offeriert werden, zu evaluieren. Zu solchen leichtfertig popularisierten Effekten, die – den üblichen empirischen Standards genügenden – Replikationen nicht standgehalten haben, zählt z. B. der «Pygmalion-Effekt», der eine massive Steigerung der Intelligenz von Kindern lediglich durch Erwartungseffekte behauptet. Die von Rosenthal und Jacobson (1968, deutsch 1971) als Beleg vorgelegte empirische Studie ist allerdings methodisch sehr fehlerhaft (zur Kritik bzw. zu fehlgeschlagenen Replikationen vgl. Dusek, 1985; Elashoff & Snow, 1971; Spitz, 1999b; zur methodologischen Kritik an Rosenthals Studien zu Lehrererwartungseffekten allgemein, vgl. Chow, 1990). Spitz (1999b) zog folgendes ernüchterndes Fazit: «Pygmalion mit seinen Spätfolgen kann von Interesse sein […], aber nur als Illustration dafür, wie Objektivität und Skepsis zu häufig durch Werbekampagnen und Interessen ersetzt worden sind. Ungeprüfte, voreilige Enthusiasmen für […] anscheinend mühelose psychologische und pädagogische Methoden zur […] Intelligenzförderung sind Chimären» (S. 229; Übersetzung durch den Autor). Es gibt Erwartungseffekte (Ludwig, 2006), aber sie bewirken keine Intelligenzsteigerung. Auch der ominöse «Mozart-Effekt» (Rauscher, Shaw & Ky, 1993), nach dem sich durch ein auf wenige Minuten begrenztes Hören einer ganz bestimmten Mozartsonate (KV 488) die Leistungen von Studenten in einem nachfolgenden Intelligenztest, speziell im abstrakträumlichen Schlussfolgern, statistisch signifikant und praktisch bedeutsam steigern ließen, hielt kritischen Replikationsversuchen mit vergleichbaren Versuchsanordnungen nicht stand, so dass Waterhouse (2006a,b) konstatierte, es gäbe überhaupt keinen «Mozart-Effekt». Mozart macht nicht schlau (vgl. Jäncke, 2008), der «Mozart-Effekt» ist nicht mehr als eine «wissenschaftliche Legende» (Bangerter & Heath, 2004; zur Kritik des «Mozart-Effekts» siehe z. B. auch Chabris, 1999; Crncec, Wilson & Prior, 2006; Schumacher, 2006; Spychiger, 2001; Steele, 2003; Steele, Bass & Crook, 1999).

In diese Reihe ist nach unseren Befunden auch der «Kaugummikaueffekt» einzuordnen. Auf der Grundlage von zwei Experimenten an Schülern der 5. und 6. Jahrgangsstufen lassen sich unter Berücksichtigung forschungsmethodisch strikter Kriterien (z. B. ausreichende Stichprobengröße mit jeweils pro Experimentalbedingung rund 250 Vpn; streng randomisierte experimentelle Versuchsanordnung; multiple Kriterienmaße) weder für Leistungen in ausgewählten *Intelligenz(sub)tests* (reasoning, number, verbal comprehension) bzw. in der *Informationsverarbeitungsgeschwindigkeit* noch für *Aufmerksamkeits- bzw. Konzentrations- und Gedächtnisleistungen* Unter-

Z. Pädagog. Psychol. 24 (1) © 2010 Verlag Hans Huber, Hogrefe AG, Bern

### Statistische Symbole

Die Ergebnisse werden deutlich leichter verständlich, wenn man die Symbole der wichtigsten statistischen Kennwerte kennt. Dies sind:
- Mittelwert: *M*
- Standardabweichung: *s*, *S* oder *SD*
- Korrelationskoeffizienten: *r* oder ρ
- Größen von (Teil-)Stichproben: *N* oder *n*
- Statistische Signifikanz: *p*
- Häufigkeit: *f*
- Stichprobenkennwerte: *t*, *F*, $\chi^2$, *z*
- Effektstärken: standardisierte Mittelwertdifferenz *d*, Varianzaufklärung $\eta^2$
- Reliabilitätskoeffizienten: interne Konsistenz α

### Diskussion

Im letzten (vierten) Hauptteil eines Fachartikels werden die Ergebnisse diskutiert und interpretiert. Eine solche Diskussion beginnt meist mit einer nochmaligen Kurzbeschreibung des zu untersuchenden Phänomens und der Forschungsfrage, gefolgt von einer verdichteten Beschreibung der Befunde. Ergebnisse zu diskutieren heißt, ihre Bedeutung und Schlussfolgerungen im Hinblick auf die formulierte Fragestellung festzuhalten, sie mit den Befunden anderer Untersuchungen zu vergleichen und daraus Implikationen für Theorie und Praxis abzuleiten.
Da in der Diskussion alle Stränge der Arbeit zusammengeführt werden, kann es eine gute Strategie zur ersten Orientierung über einen Fachartikel sein, nach der Lektüre der Zusammenfassung zunächst die Diskussion zu lesen.

### Nach der Lektüre

Am Ende der Lektüre eines Fachartikels lohnt es sich, noch einmal die gesamte Forschungsarbeit in den Blick zu nehmen – und über ihre methodische Qualität sowie ihre Bedeutung für die schulische Praxis nachzudenken.

**Abb. 27.11** (Fortsetzung)

**Zusammenfassung**

Viele für den Schulalltag relevante Fragestellungen lassen sich nur empirisch beantworten, d. h. mittels einer systematischen und methodisch kontrollierten Sammlung und Bewertung von Daten. Um als Lehrkraft von den vielfältigen und stets im Wandel befindlichen Erkenntnissen der Forschung im Bereich Schule profitieren zu können, sind grundlegende forschungsmethodische Kompetenzen erforderlich.

Empirische Forschung ist ein Prozess mit mehreren Schritten. Ausgehend von Praxisbeobachtungen oder theoretischen Überlegungen werden Hypothesen formuliert und überprüft. Dazu ist es nötig, auch Merkmale messbar zu machen, die sich nicht direkt beobachten lassen – dies wird möglich, indem Indikatoren des jeweiligen Konstrukts mit geeigneten Erhebungsverfahren erfasst werden. In der Psychologie stehen dazu eine Vielzahl an Verfahren zur Verfügung, wie systematische Verhaltensbeobachtung, Interviewverfahren, Fragebögen oder Tests. Die Qualität von Erhebungsverfahren lässt sich anhand von Gütekriterien beurteilen (insb. Objektivität, Reliabilität, Validität).

Für die Ableitung von Schlussfolgerungen stehen unterschiedliche Untersuchungsdesigns zur Verfügung. Experimentelle Forschungsdesigns zielen darauf ab, Ursache-Wirkungszusammenhänge aufzuklären. Bei Querschnittuntersuchungen geht es um die Analyse von Zusammenhängen zwischen Merkmalen, die alle zum gleichen Zeitpunkt erfasst werden. Dies ist unaufwendig, erlaubt jedoch keine kausalen Aussagen. Um Veränderungen und deren Bedingungen analysieren zu können, sind Längsschnittuntersuchungen nötig, bei denen die im Fokus stehenden Merkmale wiederholt gemessen werden. In Metaanalysen werden die Ergebnisse mehrerer, bereits vorliegender Studien zu einem Forschungsthema systematisch zusammengeführt.

Gewonnene Daten werden mit statistischen Methoden analysiert. Mit den Methoden der deskriptiven Statistik lassen sich die in einer Stichprobe erhobenen Daten übersichtlich und komprimiert darstellen. Lagemaße (z. B. Mittelwert) eignen sich zur Beschreibung der zentralen Tendenz der Daten. Streuungsmaße (z. B. Standardabweichung) geben Auskunft darüber, wie sich das gemessene Merkmal in der Stichprobe verteilt. Mit Hilfe der Inferenzstatistik lässt sich abschätzen, ob ein in der Stichprobe beobachteter Effekt zufällig auftritt oder die Bedingungen in der Grundgesamtheit reflektiert. Ist die Irrtumswahrscheinlichkeit, dass die Nullhypothese gilt, obwohl die Ergebnisse in der Stichprobe mit der Alternativhypothese in Einklang stehen (statistische Signifikanz), hinreichend gering (meist kleiner als 5 % oder 1 %), wird ein Befund als signifikant bezeichnet. Mit Hilfe von Effektstärkemaßen lässt sich die praktische Relevanz von Befunden abschätzen.

Um im Lehramtsstudium oder als Lehrkraft belastbare Informationen zu einem praktischen Phänomen zu erhalten, bietet sich die Lektüre von entsprechenden Fachartikeln an. Diese lassen sich mit Hilfe von Literaturdatenbanken auffinden. Die Kenntnis des typischen Aufbaus von Fachartikeln sowie ein strategisches Vorgehen beim Lesen ermöglicht ein leichteres Verständnis der Inhalte.

## Verständnisfragen

1. In welchem Verhältnis stehen Praxis, Theorie und Empirie zueinander und wo liegt der jeweilige Beitrag zur Weiterentwicklung von Wissen?
2. Welche Alltagstheorien rund um das Thema Lernen und Lehren kennen Sie? Greifen Sie eine heraus und überlegen Sie, worin sich diese von einer wissenschaftlichen Theorie unterscheidet.
3. Angenommen, Sie wollen das hypothetische Konstrukt „Prüfungsangst" messen. Wie genau würden Sie vorgehen und welche Indikatoren würden Sie zur Messung heranziehen?
4. Messinstrumente, wie sie in der psychologischen Forschung typischerweise eingesetzt werden, sind mehr oder weniger fehlerbehaftet. Überlegen Sie, wie solche Fehler zustande kommen und wie damit umzugehen ist.
5. Welche Erhebungsmethoden kennen Sie, um Erleben, Kognition und Verhalten von Schülerinnen und Schülern, Eltern und Lehrkräften zu erfassen?
6. Inwiefern hängt die Fragestellung einer empirischen Studie mit der Wahl des Forschungsdesigns zusammen?
7. Warum wird das Experiment oft als „starkes" Forschungsdesign bezeichnet?
8. Was ist unter interner Validität, was unter externer Validität zu verstehen?
9. In einer empirischen Studie lesen Sie, dass bei Lernenden die Häufigkeit, mit der Belohnungsstrategien zur Selbstmotivierung eingesetzt werden zu $r = -.21$ mit der Neigung korreliert, wichtige Lernaktivitäten aufzuschieben. Was drückt die Korrelation aus und als wie bedeutsam würden Sie den Befund einschätzen?
10. Reflektieren Sie die Grundidee der Inferenzstatistik. Was genau ist in diesem Zusammenhang mit dem Begriff der Signifikanz gemeint?
11. Wie würden Sie vorgehen, um belastbare Forschungsergebnisse zur Bedeutung von Humor im Unterricht zu finden?
12. Macht Kaugummikauen wirklich schlau? Lesen Sie den in ▶ Abschn. 27.6 kommentierten Artikel von Rost et al. (2010) im Original (◘ Abb. 27.11). In

welche Hauptabschnitte ist die Darstellung gegliedert und welche grundlegenden Informationen zur beschriebenen Studie finden sich in den jeweiligen Abschnitten?

## Literatur


Artelt, C. (1999). Lernstrategien und Lernerfolg: Eine handlungsnahe Studie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *31*, 86–96.

Baumert, J., Heyn, S., & Köller, O. (1992). *Das Kieler Lernstrategien-Inventar (KSI)*. Kiel: Institut für die Pädagogik der Naturwissenschaften.

Bieg, S., & Dresel, M. (2016). Entwicklung eines Fragebogens zur Erfassung des Humors von Lehrkräften aus Schülersicht (HUMLAS). *Diagnostica*, *62*, 3–15.

Bieg, S., Grassinger, R., & Dresel, M. (2017). Humor as a magic bullet? Associations of different teacher humor types with student emotions. *Learning and Individual Differences*, *56*, 24–33.

Boekaerts, M. (1999). Self-regulated learning: Where we are today? *International Journal of Educational Research*, *31*, 445–457.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.

Coe, R. (2002). It's the effect size, stupid: What "effect size" is and why it is important. Aufsatz präsentiert auf der 2002 Annual Conference of the British Educational Research Association, Exeter, England. http://www.leeds.ac.uk/educol/documents/00002182.htm (Erstellt: September).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, NJ: Erlbaum.

Deci, E. L., & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*, *39*, 223–238.

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, *3*, 231–264.

Dollase, R., & Koch, K.-C. (2010). Soziometrie. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl. S. 819–828). Weinheim: Beltz.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin: Springer.

Dresel, M., Schober, B., & Ziegler, A. (2005). Nothing more than dimensions? Evidence for a surplus meaning of specific attributions. *Journal of Educational Research*, *99*, 31–44.

Eid, M., Gollwitzer, M., & Schmitt, M. (2015). *Statistik und Forschungsmethoden* (4. Aufl.). Weinheim: Beltz.

Engelschalk, T., Steuer, G., & Dresel, M. (2015). Wie spezifisch regulieren Studierende ihre Motivation bei unterschiedlichen Anlässen? Ergebnisse einer Interviewstudie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *47*, 14–23.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2. Aufl.). Cambridge, MA: MIT Press.

Fahrenberg, J. (2010). Ambulantes Assessment. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (S. 201–212). Göttingen: Hogrefe.

Freud, A. (1936). *Das Ich und die Abwehrmechanismen*. Wien: Internationaler Psychoanalytischer Verlag.

Glock, S., & Kleen, H. (2017). Gender and student misbehavior: Evidence from implicit and explicit measures. *Teaching and Teacher Education*, *67*, 93–103.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Hattie, J. A. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Abingdon: Routledge.

Hadwin, A. F., Nesbit, J. D., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*, 107–124.

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (4. Aufl.). Seelze: Klett.

Hesse, I., & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3. Aufl.). Opladen: Budrich.

Hussy, W., Schreier, M., & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften*. Heidelberg: Springer.

Jansen, H., Mannhaupt, G., Marx, H., & Skowronek, H. (2002). *Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten* (2. Aufl.). BISC. Göttingen: Hogrefe.

Klauer, K. J. (2006). Forschungsmethoden in der Pädagogischen Psychologie. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (5. Aufl. S. 75–98). Weinheim: Beltz.

KMK (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Bericht der Arbeitsgruppe*. Bonn: KMK.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., & Weiß, M. (2003). *PISA 2000: Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.

Labuhn, A. A., Bögeholz, S., & Hasselhorn, M. (2008). Lernförderung durch Anregung der Selbstregulation im naturwissenschaftlichen Unterricht. *Zeitschrift für Pädagogische Psychologie*, *22*, 13–24.

Lehmann-Grube, S., Hartinger, A., Grassinger, R., Brandl-Bredenbeck, H.-P., Ohl, U., Riegger, M. & Dresel, M. (2017, März). *Struktur und Zusammenhangsmuster von Einstellungen zu Heterogenität: Ergebnisse einer Studie mit Lehramtsstudierenden*. Vortrag auf der 5. Jahrestagung der Gesellschaft für Empirische Bildungsforschung (GEBF) in Heidelberg.

Leopold, C., & Leutner, D. (2002). Der Einsatz von Lernstrategien in einer konkreten Lernsituation bei Schülern unterschiedlicher Jahrgangsstufen. *Zeitschrift für Pädagogik*, *45*(Beiheft), 240–258.

Maltby, J., Day, L., & Macaskill, A. (2011). *Differentielle Psychologie, Persönlichkeit und Intelligenz* (2. Aufl.). München: Pearson.

Mayring, P. (2010). Qualitative Inhaltsanalyse. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 601–613). Wiesbaden: VS.

Mehl, M. R. (2006). Textanalyse. In F. Petermann & M. Eid (Hrsg.), *Handbuch der psychologischen Diagnostik* (S. 196–202). Göttingen: Hogrefe.

Moosbrugger, H., & Kelava, A. (2011). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer.

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

Nett, U. E., Goetz, T., & Hall, N. C. (2011). Coping with boredom in school: An experience sampling perspective. *Contemporary Educational Psychology*, *36*, 49–59.

Petermann, F., & Eid, M. (Hrsg.). (2006). *Handbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.

Pfeiffer, C., Mößle, T., Kleimann, M., & Rehbein, F. (2008). Die PISA-Verlierer und ihr Medienkonsum. Eine Analyse auf der Basis verschiedener empirischer Untersuchungen. In. In U. Dittler & M. Hoyer (Hrsg.), *Aufwachsen in virtuellen Medienwelten. Chancen und Gefahren digitaler Medien aus medienpsychologischer und medienpädagogischer Perspektive* (S. 275–305). München: Kopaed.

Pintrich, P. R., & Garcia, T. (1994). Self-regulated learning in college students: Knowledge, strategies, and motivation. In P. R. Pintrich, D. R. Brown & C. E. Weinstein (Hrsg.), *Student motivation, cognition, and learning* (S. 113–133). Hillsdale, NJ: Erlbaum.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*, 387–400.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12.

Rauchfleisch, U. (2006). Projektive Tests. In F. Petermann & M. Eid (Hrsg.), *Handbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.

Rauchfleisch, U., Battegay, R., & Rosenzweig, S. (1979). *Handbuch zum Rosenzweig Picture-Frustration Test (PFT)*. Bern: Huber.

Rost, D. H. (2013). *Interpretation und Bewertung pädagogisch-psychologischer Studien* (3. Aufl.). Bad Heilbrunn: Klinkhardt.

Rost, D. H., Wirthwein, L., Frey, K., & Becker, E. (2010). Steigert Kaugummikauen das kognitive Leistungsvermögen? *Zeitschrift für Pädagogische Psychologie*, *24*, 39–49.

Schlagmüller, M., & Schneider, W. (2007). *Würzburger Lesestrategie-Wissenstest für die Klassen 7 bis 12 (WLST 7-12)*. Göttingen: Hogrefe.

Schmalt, H.-D., Sokolowski, K., & Langens, T. (2000). *Das Multi Motiv Gitter für Anschluss, Leistung und Macht (MMG)*. Frankfurt/M: Sweets Test Services.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology*, *31*, 64–96.

Schwamborn, A., Mayer, R. E., Thillmann, H., Leopold, C., & Leutner, D. (2010). Drawing as a generative activity and drawing as a prognostic activity. *Journal of Educational Psychology*, *102*, 872–879.

Seidel, T., & Prenzel, M. (2010). Beobachtungsverfahren: Vom Datenmaterial zur Datenanalyse. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (S. 139–152). Göttingen: Hogrefe.

Tobisch, A., & Dresel, M. (2017). Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education*, *20*(4), 731–752.

Tobisch, A. (2017). *Herkunftsassoziierte Lehrkrafturteile und -erwartungen: Soziale Kognitionen und Urteilsbildungsprozesse im Kontext einer ethnisch und sozial heterogenen Schülerschaft*. Dissertation. Augsburg: Universität.

Tulis, M., & Dresel, M. (2018). Emotionales Erleben und dessen Bedeutung für das Lernen aus Fehlern. In G. Hagenauer & T. Hascher (Hrsg.), *Emotionen und Emotionsregulation in der Schule und Hochschule*. Münster: Waxmann.

Von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). Judging the plausibility of arguments in scientific texts: A student-scientist comparison. *Thinking & Reasoning*, *22*, 221–249.

Weinstein, C. E., & Hume, L. M. (1998). *Study strategies for lifelong learning*. Washington, DC: American Psychological Association.

Wettstein, A. (2008). *BASYS. Beobachtungssystem zur Analyse aggressiven Verhaltens in schulischen Settings*. Bern, Schweiz: Huber.

Wirtz, M. A. (Hrsg.). (2017). *Dorsch – Lexikon der Psychologie* (18. Aufl.). Bern: Huber.