1 Forschungsmethoden in der empirischen Sportpsychologie

Karen Zentgraf & Axel Kohler

1.1 Forschungsmethoden als Mittel zur Erkenntnisgewinnung

Auf welche Art und Weise und mit welchem Ziel bearbeiten Forscher*innen in der Sportpsychologie wissenschaftliche Fragestellungen? In diesem Kapitel soll es um die methodischen Zugänge gehen, die helfen, diese Forschungsziele zu erreichen. Generell lassen sich in der Psychologie grob vier Zielbereiche definieren.

Zunächst kann es um die *Deskription* eines Phänomens gehen, also die Eingrenzung und Benennung eines Gegenstands, was er ist und was nicht. Wie lässt sich z. B. Versagen unter Druck beschreiben, durch was sind Leistungsabfälle charakterisiert, was sind überhaupt Drucksituationen in sportlichen Leistungssituationen (▶ Kap. 11)?

Ein weiteres Forschungsziel kann die *Vorhersage* des Eintretens eines Ereignisses oder eines Zustandes sein, d. h., wann wird das Ereignis/der Zustand zu beobachten sein und wann nicht? Welche Eigenschaften von Basketballtalenten sagen den späteren Erfolg vorher?

Um gute Vorhersagen erklären zu können, kann es helfen, die spezifischen Faktoren zu kennen, die das Auftreten des Ereignisses oder des Zustands bedingen – hiermit ist vornehmlich das Ziel der *Kontrolle der Kontext- oder Bedingungsvariablen* gemeint. Tritt Versagen unter Druck z. B. nur in Abhängigkeit eines bestimmten Persönlichkeitsmerkmals oder des Spielstandes ein? Welche anderen Faktoren sind denkbar, die zu Leistungsabfällen führen können? Wie generalisierbar ist die gedankliche Vorwegnahme (Antizipation) von Sportspielexpert*innen (▶ Kap. 2 und ▶ Kap. 6)?

Das Forschungsziel der *Erklärung* von Phänomenen soll Fragen beleuchten, warum ein Ereignis oder ein Zustand eintritt – welches Erklärungsmodell steckt hinter dem Phänomen? Ist eine erhöhte Selbstaufmerksamkeit Ursache für den Leistungsabfall oder werden antizipierte Misserfolge der Person handlungswirksam? Haben Sportspielexpert*innen aufmerksamkeitsbedingt eine bessere Sensitivitätsschwelle für die Detektion von spielrelevanten Hinweisreizen (▶ Kap. 3)?

Inzwischen gibt es eine Vielzahl an Methoden, die für sportwissenschaftliche Untersuchungen zur Verfügung stehen: von klassischen Verhaltensuntersuchungen bis hin zu hirnphysiologischen Messungen mit großem apparativen Aufwand.

Zuweilen kann der Eindruck entstehen, dass manche Methoden um der Methode willen angewandt werden und das Ziel, Antworten auf zentrale Forschungsfragen zu finden, in den Hintergrund rückt. Methoden dienen allerdings allein den Forschungszielen. Für bestimmte Forschungsziele eignen sich bestimmte methodische Zugänge besonders. Beobachtungsstudien, Einzelfallstudien, Berichte und Interviews sind Verfahren, die ohne eine untersucherbezogene Kontrolle des Gegenstandsbereichs arbeiten. Dies gilt auch für korrelative beschreibende Untersuchungen. Hier untersuchen Forschende, wie Variablen »natürlicherweise« zusammenhän-

gen, ohne auf kausale Wirkungszusammenhänge zu referieren. Wenn das Ausmaß körperlicher Aktivität mit einem geringeren Körpergewicht positiv korreliert, heißt dies nicht, dass körperliche Aktivität Körpergewichtsreduktionen verursacht. Aber aufgrund der Kenntnis dieses Zusammenhangs aus beschreibenden Untersuchungen ergeben sich ggf. weitere Fragen, die mit anderen und ergänzenden methodischen Ansätzen beantwortet werden können.

Allerdings sind unter bestimmten Bedingungen andere methodische Ansätze auch ethisch nicht vertretbar, korrelative Studien bieten dann die einzige Möglichkeit, sportpsychologische Forschung zu betreiben: Die Untersuchung der Frage, inwiefern schwere Sportverletzungserfahrungen wettkampfbezogene Selbstwirksamkeitserwartungen verändern, wird Sportpsycholog*innen aus plausiblen ethischen Gründen nicht auf die Idee bringen, Verletzungen in einem manipulativen Sinne herzustellen, sondern man wird Forschungsdesigns anwenden, die z. B. verschiedene Gruppen mit und ohne Sportverletzungserfahrungen vorsehen.

1.1.1 Experimentieren in der Sportpsychologie

Wilhelm Windelband formulierte bereits 1894 die – zumindest im deutschen Sprachraum – sehr einflussreiche Unterscheidung von nomothetischen und idiographischen Ansätzen in der Wissenschaft. Der idiographische Ansatz zielt auf die Beschreibung und Analyse einzigartiger Vorgänge und Ereignisse (z. B. historische Prozesse oder individuelle Biografien).

Die Forschungsziele nomothetischer Ansätze liegen hingegen im Besonderen in der Erklärung von Phänomenen und der Entwicklung von allgemein gültigen Gesetzmäßigkeiten. Die Idee der Verwendung experimenteller Methoden liegt darin begründet, durch strenge Kontrolle der möglichen beeinflussenden Variablen die wirksamen Faktoren zu isolieren. Dass die Auslegung, wie »streng« diese Bedingungen in der Forschungswirklichkeit zu kontrollieren sind, sehr heterogen ist, muss zur Kenntnis genommen werden.

Übergreifend fordern experimentelle Herangehensweisen die Forscher*innen auf, eine Variable systematisch zu verändern (unabhängige Variable, UV, z. B. verschiedene, von der Versuchsleitung vorgegebene sportliche Belastungen), möglichst alle anderen zu kontrollieren, z. B. konstant zu halten, und den Einfluss der Manipulation der UV auf eine weitere, die abhängige Variable (AV, z. B. die Treffergenauigkeit), zu messen. Dadurch scheint es möglich, auf theoretischer Basis begründete Erwartungen (Hypothesen) zu überprüfen, die Forscher*innen als verursachend für den zu beobachtenden Effekt annehmen.

Eine UV hat typischerweise Stufen, also verschiedene Ausprägungen, die manipuliert werden. Zwei Stufen sind minimal notwendig, um überhaupt Vergleiche vornehmen zu können (beim Beispiel der sportlichen Belastung könnte es eine hohe und niedrige Beanspruchung geben oder aber sogar feinere Abstufungen). Die Stufen können sich auf eine oder mehrere Versuchsgruppen beziehen: Entweder durchläuft ein*e Versuchsteilnehmer*in beide (oder alle) Stufen der UV (die Belastungsstufen würden bei der gleichen Person an verschiedenen Versuchstagen oder mit ausreichender Pause untersucht) oder eine Person wird zufällig einer Stufe der UV zugeordnet (dann gibt es unterschiedliche Gruppen für die Belastungsstufen). Im ersteren Fall spricht man von sog. »Within-Subject«- oder Intrapersonalen Designs mit Messwiederholung; im zweiten Fall von »Between-Subject-« oder Zwischengruppendesigns. Bei beiden Designs sind bzgl. der Kontrolle von Störvariablen und zur Vermeidung von Konfundierungen weitere Aspekte zu berücksichtigen, z. B. Reihenfolgeeffekte, Stichprobenfehler, nicht-zufällige Zuordnungen etc. Diese Verfahren sollen sicherstellen, dass das Ergebnis von der untersuchten Stichprobe auch überzeugend auf die Gesamtpopulation

übertragen werden kann. Aber auch hier können Fehler entstehen, wenn z. B. von der Untersuchung von Studienteilnehmer*innen aus einem Kulturkreis auf Personen aus einem anderen Kulturkreis generalisiert wird.

1.1.2 Experimentelle Validitäten, Forschungsdesigns und Versuchspläne in der Sportpsychologie

In Forschungsfeldern ergeben sich aus der Interaktion zwischen den formalen Ansprüchen an experimentelle Forschung (Kontrolle der Bedingungen, zufällige Zuordnung in Versuchsgruppen etc.) und den Gegebenheiten in der jeweiligen Disziplin (verfügbare Methoden und charakteristische Fragestellungen etc.) typische experimentelle Designs und Versuchspläne, die zur Untersuchung von wissenschaftlichen Fragestellungen Anwendung finden.

Dies kann man auch als Forderungen an die interne und externe Validität von Experimenten verstehen. Es geht in beiden Fällen entsprechend der Wortbedeutung um die Gültigkeit von Aussagen oder Schlussfolgerungen, die aus Studien gezogen werden. Die interne Validität bezieht sich auf die Frage, ob ein Versuch so gestaltet war, dass man auf eine Kausalbeziehung zwischen den UV und AV schließen kann. Wenn in einem Laborversuch zu mentalen Rotationsleistungen mit menschlichen Körpern ausschließlich die Sichtbarkeit des Reizes in verschiedenen Bedingungen manipuliert und ein Effekt auf die Antwortzeiten festgestellt wird, kann man sicher sein, dass die Manipulation einen direkten kausalen Einfluss hatte. Wenn in zwei Schulklassen zwei verschiedene Unterrichtsmethoden zum Beobachtungslernen einer Rolle rückwärts eingesetzt werden, gibt es viele Unterschiede zwischen den Klassen, die einen Einfluss auf die Ergebnisse haben könnten. Zuerst muss sichergestellt sein, dass die Klassen überhaupt vergleichbar sind. Ist die eine Klasse disziplinierter, erreichen sie bessere Ergebnisse, obwohl das nichts mit der Unterrichtsmethode zu tun hat. Selbst wenn Unterschiede zwischen den Klassen messbar gemacht werden, ist nicht auszuschließen, dass ein anderer Einflussfaktor übersehen wurde. Man geht deshalb bei solchen Feldstudien allgemein von einer eingeschränkten internen Validität im Vergleich zum Laborexperiment aus (auch wenn das bei gut geplanten Feldstudien nicht notwendigerweise der Fall ist).

Bei der externen Validität ist es eher die Frage, inwiefern sich Ergebnisse auf andere Personengruppen und Situationen übertragen lassen. Gerade im Sportbereich ist es oft notwendig, mit vorhandenen Gruppen (z. B. Spitzenathlet*innen, Schulklassen, Alzheimer-Patient*innen) zu arbeiten bzw. ist die Nützlichkeit von Laborexperimenten sehr begrenzt, wenn eine Anwendung auf den sportlichen Kontext gar nicht möglich oder fraglich ist.

Als weiteren Begriff in diesem Bereich wird oft noch die ökologische Validität als Kriterium angeführt. Sie bildet ab, in welchem Maß eine Untersuchungssituation einem natürlichen Kontext entspricht. »Natürlich« bezieht sich dabei nicht auf die belebte Natur, sondern auf Situationen, die auch im Alltag vorkommen können. Beim Menschen sind das in den meisten Teilen der Welt stark kulturell geprägte soziale Umgebungen.

Rein definitorisch kann man die ökologische von der externen Validität trennen. Wenn ein Experiment mit sehr vielen Personengruppen verschiedenen Alters, verschiedener Nationalitäten und in verschiedenen Laborsituationen durchgeführt wurde, sollte ein hohes Maß an externer Validität gegeben sein. Man könnte hier aber immer noch einwenden, dass das Experiment nicht in möglichen Alltagssituationen erprobt wurde (mangelnde ökologische Validität). Man kann also Szenarien beschreiben, in denen die beiden Aspekte nicht notwendig zusammenhängen.

Es können an dieser Stelle nicht alle Faktoren benannt und erläutert werden, welche die verschiedenen experimentellen Validitä-

ten beeinflussen. Bei den Beispielen und Erwägungen wurde aber bereits deutlich, dass die Validitäten teilweise im Widerspruch stehen können.

Wenn man die obigen Überlegungen auf konkrete Forschungsdesigns anwendet, lassen sich verschiedene Ansätze zur Durchführung von Studien nach der Qualität der Schlussfolgerungen und der Generalisierbarkeit unterscheiden. Die folgende Darstellung zu den Designfragen lehnt sich an einen klassischen Text von Campbell und Stanley (1963) an. Dort wird noch eine Vielzahl weiterer Designs besprochen und auch die jeweiligen Stärken und Schwächen im Detail dargelegt. An dieser Stelle sollen nur die wichtigsten Aspekte und die bekanntesten Designs betrachtet werden.

Aus der Ausprägung der internen und externen Validität lassen sich zwei Grundunterscheidungen von Studien ableiten, die oben zum Teil schon angesprochen wurden. In Bezug auf die externe bzw. ökologische Validität kann man zwischen Labor- und Feldstudien differenzieren. Die künstliche Laborsituation schränkt die externe Validität ein, die im Vergleich bei Studien in realistischen Situationen und Umgebungen (»im Feld«) höher ausgeprägt ist. Auch wenn es in Feldsituationen manchmal schwerer fällt, Störeinflüsse zu kontrollieren, ist die interne Validität bei sorgfältig geplanten und durchgeführten Feldstudien nicht notwendigerweise eingeschränkt. Bei der internen Validität kann man stattdessen zwischen Experimenten und Quasi-Experimenten unterscheiden. Für ein erstes Verständnis der Konzepte ist es ausreichend, von einer experimentellen Gruppe und einer Kontrollgruppe auszugehen. Bei »echten« Experimenten wird die Zuordnung der Versuchspersonen zu den Gruppen zufällig ermittelt. Das heißt, weder die Versuchsleitung noch die Versuchspersonen haben Einfluss darauf, wer welcher Gruppe angehört. Dadurch wird sichergestellt, dass es keine systematischen Unterschiede in der Gruppenzusammensetzung gibt. Könnten sich die Versuchspersonen eine Gruppe aussuchen, würden evtl. Persönlichkeitsmerkmale die Gruppenwahl beeinflussen (»Die Experimentalgruppe finde ich spannender« vs. »Egal, ich kann auch in die Kontrollgruppe«). Würde die Versuchsleitung die Einteilung bestimmen, könnte die Gruppenzusammensetzung – oft auch unbewusst – im Sinne der eigenen Erwartungen beeinflusst werden. Eine Vielzahl solcher Versuchsleitungseffekte war selbst Gegenstand experimenteller Untersuchungen und konnten überzeugend als relevant dargelegt werden.

In manchen Fällen ist es jedoch nicht möglich, die Zuordnung zu kontrollieren oder randomisiert vorzunehmen. Werden Geschlechterunterschiede untersucht, müssen die Unterschiede als gegeben hingenommen werden. Wir können Menschen nicht zufällig in eine Geschlechterkategorie einordnen. In sportwissenschaftlichen Untersuchungen besteht häufig Forschungsinteresse an Expertinnen und Experten, die bestimmte Bewegungen oder die Ausführung einer Sportart über lange Jahre perfektioniert haben (▶ Kap. 6). Auch in diesem Fall müssen wir mit den bereits vorhandenen Gruppen zurechtkommen. Erfolgt die Zuordnung in die jeweiligen Gruppen nicht randomisiert, ist die interne Validität eingeschränkt. Man spricht dann von »quasiexperimentellen Designs«, da die Durchführung einer solchen Studie trotzdem nach hohen experimentellen Standards erfolgen kann, die entscheidende Anforderung der Randomisierung aber von vornherein nicht erfüllt ist.

Unter den experimentellen Designs ist zunächst ein Grunddesign zu betrachten, das die wichtigsten Aspekte eines Zwischengruppen-Experiments abdeckt: das Prätest-Posttest-Kontrollgruppen-Design. Auch wenn Gruppen zufällig eingeteilt werden können, kann es gerade wegen des Zufalls auch vor der experimentellen Manipulation bestehende Unterschiede zwischen den Gruppen in den abhängigen Variablen geben. Es ist deshalb oft ratsam, eine erste Referenztestung (Prätest) zu Beginn durchzuführen. Dann folgt für die Experimentalgruppe die Intervention, wäh-

rend die Kontrollgruppe entweder keine Intervention erhält oder idealerweise eine Kontrollbedingung durchläuft, die sonst vergleichbar ist mit der Experimentalbedingung, aber sich im entscheidenden Faktor unterscheidet. Wenn die Hypothese ist, dass ein Krafttraining das physische Selbstkonzept stärkt, dann könnte ein Ausdauertraining als Kontrollbedingung wenigstens ausschließen, dass der Effekt allein auf der körperlichen Betätigung beruht. Wenn die Kontrollgruppe keine Intervention erhält, bleibt die Zahl der Alternativerklärungen unbefriedigend hoch, sodass es weiterer Experimente bedarf, um einen spezifischen Mechanismus zu isolieren.

Nach der experimentellen Manipulation (unabhängige Variable) soll der Effekt auf die abhängige Variable getestet werden (Posttest). Der Vergleich von Posttest und Prätest zeigt, ob beim Experiment ein Effekt erzeugt wurde, der vorher noch nicht oder in einem geringeren Maß vorhanden war. Im Prinzip kann ein solches Design auch ohne Prätest mit den entsprechenden Einschränkungen bei der Interpretation durchgeführt werden. Da nicht ausgeschlossen werden kann, dass ein Prätest die Ergebnisse einer experimentellen Intervention oder Manipulation und/oder die Effekte im Posttest verändert, wurde von Solomon (1949) vorgeschlagen, in einem umfassenden Design die Interaktionen zwischen allen Einflussfaktoren direkt zu testen. Wenn im Beispiel oben bereits vorher ein Test zum physischen Selbstkonzept durchgeführt wurde, könnten die Versuchspersonen anders an die Trainingsintervention herangehen und sie würden sich für den Posttest vielleicht schon überlegen, was die erwarteten Effekte sind und ihr Antwortverhalten (oft auch unbewusst) darauf einstellen. Beim Solomon-Vier-Gruppen-Design gibt es mindestens zwei Gruppen mit und zwei Gruppen ohne Prätest. Damit wird der Einfluss der Vortestung überprüft. Bei jedem Paar gibt es eine Kontrollgruppe für die Abschätzung des Interventionseffekts. Einziger Nachteil bei dem Design ist der recht hohe Aufwand. Ist ein Einfluss der Vortestung unwahrscheinlich, ist zu überlegen, ob die Ressourcen für die zusätzlich erforderlichen Gruppen gut investiert sind.

In vielen Forschungsbereichen ist das Ziel, mehr als einen Einflussfaktor zu untersuchen. Zum einen steigert es die Effizienz unserer Forschungsaktivität, weil mehrere Hypothesen in einem Experiment getestet werden können. Zum anderen können nur bei gleichzeitiger Betrachtung mehrerer Faktoren in einem Datensatz sogenannte »Interaktionseffekte« überprüft werden. Bei Interaktionen geht es um die Abhängigkeit oder Unabhängigkeit zwischen verschiedenen Einflussgrößen. Was ist, wenn der Effekt des Krafttrainings auf das Selbstkonzept davon abhängt, mit wem ich mich beim Training vergleiche? Experimentatoren könnten den Versuchspersonen bestimmte Trainingspartner*innen zuweisen. Bei leistungsgleichen oder etwas schwächeren Vergleichspersonen zeigt sich vielleicht der erwartete, positive Effekt auf das Selbstkonzept. Wird einer Versuchsperson jedoch eine leistungsstärkere Person zugewiesen, zeigt sich evtl. ein negativer Effekt. Wichtig für das Bestehen einer Interaktion ist, dass bei der Kontrolltrainingsbedingung (z. B. Ausdauertraining anstatt Krafttraining) dieser Vergleichsfaktor keine Rolle spielt. Um den Effekt des einen Faktors (Vergleichsgruppe) auf das Selbstkonzept vorherzusagen, muss also auch die Ausprägung des anderen Faktors bekannt sein (Art des Trainings). In diesem Fall ist eine Interaktion vorhanden und die Faktoren sind nicht voneinander unabhängig. Das Design liefert also Zusatzinformationen, die über den getrennten Einfluss der einzelnen Manipulationen (die sogenannten »Haupteffekte« für »Vergleichsgruppe« und »Art des Trainings«) hinausgeht.

Im Vergleich zu den experimentellen Designs können bei quasi-experimentellen Untersuchungen bestimmte Faktoren nicht vollständig kontrolliert werden. Vielleicht sollen vorhandene Gruppen verglichen werden, es besteht Interesse an speziellen Gruppen (z. B. Hochleistungssportler*innen, Alzheimer-Patient*innen) oder es müssen gegebene Orga-

nisationsformen hingenommen werden (z. B. Schulklassen). In all diesen Fällen ist bei der Interpretation der Ergebnisse besondere Vorsicht geboten. Es kann nicht ausgeschlossen werden, dass die bereits vorher bestehenden Unterschiede zwischen den Gruppen auch die Effekte der experimentellen Manipulation erklären. Es können durchaus Designs verwendet werden, die grundsätzlich den experimentellen Formen entsprechen. Wird die Vergleichbarkeit der Gruppen auf möglichst vielen Dimensionen sichergestellt, kann die Interpretierbarkeit der des echten Experiments sehr nahekommen.

Abgesehen von den experimentellen und quasi-experimentellen Designs, bei denen verschiedene Bedingungen aktiv realisiert werden, wird bei vielen Studien nur der Zusammenhang zwischen verschiedenen Maßen betrachtet. Man spricht hier von Korrelationsstudien. Kann man anhand physiologischer Messungen den Erfolg bei einem Wettkampf vorhersagen? Werden Menschen mit bestimmten Persönlichkeitseigenschaften mit höherer Wahrscheinlichkeit Leistungssportler*innen? Die Einschränkung der internen Validität ist in diesem Fall besonders stark. Streng genommen sollten gar keine Schlussfolgerungen über kausale Mechanismen gezogen werden.

In vielen Fällen scheint allerdings eine kausale Interpretation plausibel, z. B. dass der Zusammenhang zwischen höherer Lebenserwartung und körperlicher Aktivität darauf zurückgeht, dass sportliche Betätigung die Lebenserwartung erhöht. Es gibt aber eine Vielzahl von Alternativerklärungen, die den gefundenen Zusammenhang erklären können und bei denen der kausale Mechanismus ein anderer ist: Personen, die sich sportlich betätigen, ernähren sich möglicherweise auch anders, was die eigentliche Ursache für den Effekt auf die Lebenszeit sein könnte. Eventuell gibt es Erbanlagen, die sich sowohl auf die Sportlichkeit als auch auf die Lebensdauer auswirken. Kausale Belege in diesem Kontext stammen aktuell aus einer Vielzahl sorgfältig geplanter experimenteller Studien, in denen verschiedene Alternativerklärungen ausgeschlossen werden konnten. Auch Experimente an Tieren unterstützen den Zusammenhang, der – mit Einschränkungen – auf den Menschen übertragbar ist. Im Hinblick auf die Forschungsziele ist demnach zu konstatieren: Korrelative Daten sind zunächst beschreibend und nicht erklärend.

1.2 Maße und Messmethoden in der Sportpsychologie

Entsprechend der Ausrichtung dieses Buches auf quantitative, experimentelle Ansätze soll an dieser Stelle insbesondere auf Methoden eingegangen werden, die in empirischen, natur- und verhaltenswissenschaftlich orientierten Studien zum Einsatz kommen.

1.2.1 Verhaltensmaße

Die solide Basis für die experimentelle Arbeit in der empirischen Sportwissenschaft ist die sorgfältige Erhebung von Verhaltensdaten. In vielen Fällen werden bei einer bestimmten Aufgabenstellung auch bestimmte Reaktionen erwartet. Grundvariablen für die Auswertung sind dabei die Reaktionszeiten und die Korrektheit der Antworten. Reaktionszeitparadigmen spielen auch in der Sportpsychologie eine wichtige Rolle, im Rahmen vieler Fragestellungen werden Einfach- oder Wahlreaktionen abgefragt. Weitere Paradigmen sind »Go/no-go«- und »Two-alternative-forced-choice«-Aufgaben; bei letzteren muss

über die räumliche oder zeitliche Ordnung zweier Reize entschieden werden. Primingstudien sind in der Sportpsychologie weit verbreitet: Hierbei geht es um die Wirkung der Bahnungsreize (»Primes«) auf Zielreize (»Targets«). Zahlreiche weitere Standardparadigmen wie Stroop, Eriksen-Flanker, Posner-Cuing, visuelle Suche etc. werden auch in sportpsychologischen Studien verwendet (▶ Kasten 1.1).

Kasten 1.1: Einige in der Sportpsychologie häufig verwendete Paradigmen

- *Wahlreaktionen*: Der zeitliche Ablauf eines Durchgangs wird bestimmt durch eine Reizdarbietung, die Verarbeitungszeit und die Antwort der Versuchspersonen (Vp). Für die Antwort stehen der Vp zwei Optionen zur Verfügung, aber es wird pro Durchgang nur *ein* Reiz präsentiert. Die experimentellen Bedingungen werden meist durch die unterschiedlichen Reize implementiert. Als AV werden sowohl die Dauer der Reaktions-/Antwortzeit (d. h. vom Beginn der Reizdarbietung bis – meist – zum Drücken einer Taste) als auch die Qualität der Antwort betrachtet. Beispiel: In einem Wahrnehmungsexperiment reagieren Handballtorhüter (Helm, Reiser & Munzert, 2016) auf visuelle Reize mit dem rechten oder linken Arm. Zu der Reaktionszeit (Dauer zwischen Reizpräsentation und Start der motorischen Antwort) wurde zudem auch die Bewegungszeit (Dauer der Armstreckung) erfasst.
- »*Two-alternative forced choice« (TAFC)*: Hier werden der Vp *zwei* Reize innerhalb eines Durchgangs präsentiert und sie muss sich für eine Option entscheiden, basierend auf einem instruierten Kriterium (z. B. »Welcher Reiz ist heller?« »In welchem Reiz bewegen sich Punkte nach oben oder unten?« etc.). Auch hier interessieren die Antwortdauern und -qualitäten. Beispiel: Kennel, Hohmann und Raab (2014) verwenden eine akustische TAFC-Aufgabe, um die Rolle motorischer Expertise bei der Diskriminierung von bewegungsbezogenen Geräuschen beim Hürdenlaufen zu untersuchen.
- *Gleich-Verschieden*: Dargeboten werden entweder gleichzeitig oder kurz nacheinander zwei Reize, über deren Gleich- oder Verschiedenheit die Vp entscheiden muss. Beispiel: Shepard und Metzler (1971) präsentierten zwei Würfelfiguren aus unterschiedlichen Perspektiven und manipulierten systematisch den Drehwinkel, der die beiden in Überdeckung bringen konnte. Die Forscher zeigten, dass die Zeit für die Entscheidung über Gleichheit oder Unterschiedlichkeit proportional zur Disparität der Figuren anstieg. In der mentalen Rotationsforschung innerhalb der Sportpsychologie werden heute viele Paradigmen verwendet, die sich vom klassischen Gleich-Verschieden-Paradigma unterscheiden (Heppe, Kohler, Fleddermann et al., 2016; Jansen, Lehmann & Van Doren, 2012; Steggemann, Engbert & Weigelt, 2011) und die Fragestellungen untersuchen, wie körperliche und sportliche Aktivität die mentalen Rotationsleistungen verändert.
- *Stroop*: Hier werden Reize präsentiert, bei denen geschriebene Farbwörter (»ROT«, »BLAU« etc.) mit Farben versehen werden: entweder kongruent, d. h. das Wort »ROT« wird in roter Schrift dargeboten, oder inkongruent, d. h. das Wort »ROT« in blauer Farbe. Auch neutrale Bedingungen sind möglich (z. B. das Wort »ROT« in schwarzer Farbe oder einfach nur ein farbiger Reiz, z. B. ein Kreis). Variiert werden auch die Aufgaben: also entweder das Wort vorlesen (unabhängig von der Farbe) oder die Farbe benennen (unabhängig vom Wort). Beispiel: Das Paradigma wird seit einigen Jahren im Rahmen einer größeren Test-Batterie verwendet, um die kognitiven Leistungseffekte von im Sport erlittenen Gehirnerschütterungen zu untersuchen (Echemendia, Putukian, Mackin et al., 2001; McCrea, Guskiewicz & Marshall, 2003).

- *Eriksen-Flanker*: Die Vp hat die Aufgabe, einen zentral präsentierten Reiz zu benennen (z. B. »Zeigt der Pfeil nach rechts oder links?«). Dabei sind die benachbart (flankierend) präsentierten Reize nicht zu beachten. In manchen Studien werden Ziel- und Flankierreiz auch zeitlich versetzt präsentiert (mit einer sog. Stimulus-Onset-Asynchronie, SOA). Beispiel: Varianten des Paradigmas wurden verwendet, um z. B. zu untersuchen, ob sich körperlich Aktive oder Personen, die einer sportlichen Interventionsgruppe zugeordnet sind, in ihren Leistungen und neuralen Substraten unterscheiden (z. B. Chaddock, Hillman, Pontifex et al., 2012).
- *Posner-Cueing*: Auch hier werden unterschiedliche Typen von Reizen eingesetzt. Zentral wird ein Hinweisreiz (Cue) dargeboten, der mit hoher Wahrscheinlichkeit angibt, ob nachfolgend rechts oder links ein Zielreiz erscheint. Die Aufgabe der Vp ist es, auf den Zielreiz rechts oder links mit einem Tastendruck zu antworten. Unterschieden werden valide und invalide Durchgänge, d. h. einmal gibt der Cue tatsächlich an, auf welcher Seite der Reiz kommen wird und in anderen Durchgängen nicht (invalide Durchgänge). Die Reaktionszeiten bei Durchgängen mit validem Hinweisreiz sind in der Regel niedriger als bei invaliden. In der Sportpsychologie werden Varianten des Paradigmas eingesetzt, um z. B. Zusammenhänge zwischen dem motorischen und dem kognitiven Leistungszustand von Kindern, Jugendlichen oder Erwachsenen im mittleren oder höheren Alter zu untersuchen (z. B. Wang, Liang, Tseng et al., 2015; Cereatti, Casella, Manganelli et al., 2009).
- *Priming*: Dazu existieren zahlreiche Varianten. Die methodische Grundlogik ist, dass ein vorhergehender Bahnungsreiz (»prime«) den nachfolgend präsentierten Zielreiz (»target«) in seiner Verarbeitung beeinflusst. Bahnungsreize können unterschiedliche Modalitäten ansprechen (visuell: Bild; akustisch: Ton; olfaktorisch: Geruch etc.) und sollen (meist implizit) einen Gedächtnisinhalt aufrufen, der die Verarbeitung des nachfolgenden Zielreizes, der bewusst verarbeitet wird, beeinflusst. In einer Studie von Stone, Harrison und Mottley (2012) wurden am College studierende Athlet*innen unterschiedlicher ethnischer Herkunft vor der Bearbeitung von Aufgaben entweder als »Scholar Athletes«, »Athletes« oder »Research Participants« gebahnt. Entsprechend der Erwartungen eines Stereotyps »Threats« zeigten die afroamerikanischen Sportler*innen besonders schlechte Leistungen, wenn sie bzgl. ihrer athletischen Identität gebahnt wurden.

Mit entsprechenden Versuchsanordnungen können mit Verhaltensmaßen dann auch Aussagen über den Zeitverlauf mentaler Prozesse getroffen werden.

Die Verwendung von Verhaltensmaßen in der Forschung hat eine lange Tradition. Es war einer der ersten Ansätze, um »geistige« Prozesse sichtbar zu machen. Wichtige erste Schritte in diesem Bereich waren die Arbeiten der »Psychophysiker« des 19. Jahrhunderts. Auch auf persönlicher Ebene gab es eine enge Verknüpfung mit der Entwicklung der Psychologie als wissenschaftlicher Disziplin in Deutschland. Wilhelm Wundt (1832–1920), der 1879 das erste Psychologische Institut in Leipzig gründete, war in intensivem wissenschaftlichen Austausch mit den ebenfalls in Leipzig arbeitenden Gründungsvätern der Psychophysik, Gustav Theodor Fechner (1801–1887) und Ernst Heinrich Weber (1795–1878). Fechner verallgemeinerte die Erkenntnis von Weber, dass bei der Entdeckung eines sensorischen Unterschieds – z. B. beim Vergleich der Helligkeit von zwei Lichtquellen oder der Lautstärke von zwei Tönen – die Unterschiedsschwelle vom Intensitätsniveau der Reize abhängt. Bei sehr leisen Tönen kann schon ein geringer Unterschied entdeckt werden, bei lauten Tönen bedarf es hingegen eines deutlich größeren Unterschieds zwischen dem Reizpaar. Anders

ausgedrückt ist die für eine Unterscheidung notwendige Reizdifferenz als prozentualer Anteil der Grundintensität bestimmt. Die Verallgemeinerung dieses Zusammenhangs wird auch als Weber-Fechner-Gesetz bezeichnet: Es definiert einen funktionalen Zusammenhang zwischen der Reizintensität und der Empfindung. Die Art des Zusammenhangs ist logarithmisch, was sich einfach mathematisch aus der von Weber beschriebenen Eigenschaft der Unterschiedsschwelle ergibt.

In der Mitte des 20. Jahrhunderts stellte sich heraus, dass das Weber-Fechner-Gesetz nicht auf alle Schwellenphänomene in sensorischen Systemen umfassend zutraf. Stanley S. Stevens entwickelte neue Messmethoden zur Erfassung von Empfindungen; d. h., Versuchspersonen mussten nicht Unterschiede zwischen Größen bestimmen, sondern gaben direkte Schätzungen der absoluten Wahrnehmungsintensität ab. Außerdem berücksichtigte er bei der mathematischen Herleitung des psychophysischen Gesetzes, dass auch auf der Seite der Empfindung der wahrgenommene Unterschied von der absoluten Intensität der Empfindung abhängt (wie bei der Reizschwelle). Die Form des funktionalen Zusammenhangs ist dann nicht mehr logarithmisch, sondern entspricht einer Potenzfunktion. In vielen Fällen haben diese beiden Funktionen einen ähnlichen Verlauf, doch es zeigte sich, dass die Potenzfunktion die empirischen Daten deutlich besser abbilden konnte und sich bereits beschriebene Abweichungen vom Weber-Fechner-Gesetz dadurch erklären ließen.

Parallel zu Stevens' Weiterentwicklung der Psychophysik entstand ein neuer Ansatz zur mathematischen Behandlung von Schwellenmessungen, der auch heute noch von zentraler Bedeutung ist. Die Signalentdeckungstheorie entwickelte sich ursprünglich als Antwort auf Fragen im Umgang mit der Radartechnik. Das erklärt die manchmal eigentümlich technisch klingende Terminologie, die in diesem Kontext Anwendung findet. Der Grundgedanke der Signalentdeckungstheorie ist, dass bei Wahrnehmungen und insbesondere bei darauf basierenden Entscheidungen das Signal fast immer mit Rauschen vermischt ist. Das Rauschen beruht auf variablen Eigenschaften der technischen Systeme, die für eine Messung eingesetzt werden, kann aber auch in der beobachtenden Person selbst begründet sein, wenn z. B. die Aufmerksamkeit fluktuiert oder auch der Zustand der sensorischen Systeme im Gehirn variabel ist. Das heißt, in den meisten Situationen werden nicht konstante, isolierte Signale präsentiert, sondern es ist ein Signal zu entdecken, das mit Rauschanteilen versetzt ist.

Ein zweiter wichtiger Aspekt der Signalentdeckungstheorie ist, dass bei Beobachter*innen nicht nur die Sensibilität für Signale entscheidend ist, sondern auch das angewandte Kriterium. Angenommen, zwei angehende Schiedsrichterinnen im Basketball nehmen an einer Schulung teil und sollen im Rahmen der Beobachtung von Spielszenen Fouls entdecken. Sie sind darüber informiert, dass bei einem Teil der Szenen Fouls begangen wurden. Eine Schiedsrichterin möchte auf keinen Fall ein Foul verpassen und entscheidet fast immer auf »Foul«. Dadurch berichtet sie alle vorhandenen Fouls korrekt, erzeugt aber auch sehr viele falsche Entscheidungen. Das heißt, dass zusätzlich zur korrekten Entdeckung von Signalen (Foul) auch die Anzahl der Falschmeldungen berücksichtigt werden muss. Die andere Schiedsrichterin ist deutlich kritischer: Sie übersieht zwar zwei Fouls, erzeugt aber sonst keine Fehlentscheidung. Welche Schiedsrichterin hat nun die bessere Leistung gezeigt? Es ist klar, dass die erste Schiedsrichterin mehr Fouls entdeckt, allerdings ist ihr Kriterium so liberal, dass es viele Pfiffe für Spieler gibt, die gar kein Foul begangen haben – mit entsprechenden Folgen wie vorzeitigem Spielausschluss durch Erreichen der Foulhöchstgrenze. Die zweite Schiedsrichterin entdeckt nicht alle Fouls, aber es wird auch kein Spieler mit einem Foul belastet, das er nicht begangen hat.

Die Signalentdeckungstheorie hilft, diesen Aspekt des Antwortkriteriums von der eigentlichen sensorischen Sensibilität zu unterscheiden. Die Sensibilität gibt an, wie gut Personen

oder auch Apparaturen das Signal vom Rauschen trennen können. Das Antwortkriterium spiegelt davon unabhängig wider, ob die Entscheidungen eher konservativ (Signale werden manchmal verpasst, aber es gibt auch kaum »falschen Alarm«) oder liberal sind (mehr Signale werden entdeckt, aber auch häufiger Rauschen als Signal interpretiert). Innerhalb einer Person kann dieses Kriterium unabhängig von der Sensibilität auch verschoben werden, indem man z. B. Anreize setzt oder bestimmte Fehler höher »bestraft«. Im Fall der Foulansage könnte man dieses Kriterium der Basketball-Schiedsrichterinnen dadurch beeinflussen, dass direkt Feedback gegeben wird (siehe z. B. Schweizer & Plessner, 2013).

Heutzutage wird der Begriff »Psychophysik« oftmals allgemein für jegliche Art von Verhaltensmessung eingesetzt. Das ist mit Bezug auf die Originalkonzeption und angesichts der historischen Entwicklung dieses Feldes eigentlich nicht angemessen, aber in den meisten Fällen unproblematisch. Für Schwellenmessungen werden inzwischen komplexe mathematische Modelle angewandt, mit denen Schwellen adaptiv und effizient bestimmt werden können. Unter anderem kommen dabei Bayes'sche Statistiken zum Einsatz, welche die Beziehung zwischen zugrundeliegenden mathematischen Modellen und empirischen Daten optimal formalisieren.

1.2.2 Tests

Im Folgenden wird unter dem Begriff des Tests ein »wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung« (Bortz & Döring, 2002, S. 189) verstanden. Im Rahmen sportpsychologischer Forschung werden Tests zur Messung von motorischen und kognitiven Leistungen oder von Persönlichkeitsmerkmalen bei verschiedenen Populationen (Kinder, Jugendliche, Erwachsene) und Zielgruppen angewendet. Zur Abgrenzung sei angemerkt, dass bei Leistungstests ein objektiver Beurteilungsmaßstab existiert: Wird beim Zahlenverbindungstest die 17 nicht mit der 18, sondern mit der 19 verbunden, ist die Aufgabe falsch gelöst. Bei Persönlichkeitstests ist dies nicht der Fall: Einstellungen oder Motive sind nicht falsch oder richtig, sondern stärker und weniger stark ausgeprägt. Allerdings sollte differenziert werden zwischen Persönlichkeitstests und Persönlichkeitsfragebogen: Die Anwendung von Tests findet stärker im Rahmen der Diagnostik latenter Merkmale und des Vergleichs von individuellen Daten mit Referenz- oder Normwerten statt; dies ist bei Fragebögen meist nicht der Fall.

Auch motorische Tests können als Leistungstests kategorisiert werden. Sie werden häufig im Rahmen von Entwicklungsuntersuchungen eingesetzt: Der motorische Entwicklungsstand soll erfasst, ggf. Förderbedarf erkannt bzw. ein prozessbegleitendes Beobachtungsverfahren (z. B. Zimmer & Volkamer, 1987, MOT 4-6) zur Verfügung gestellt werden. Beim Körperkoordinationstest für Kinder (KTK; Kiphard & Schilling, 2007) oder beim Movement-ABC (M-ABC; Henderson, Sugden & Barnett, 2007) werden Komponenten motorischer Funktionen mit standardisierten Aufgaben analysiert. Dabei wird – bei aufgabenangemessener Bearbeitung – der Ausprägungsgrad (z. B. die Anzahl der gefangenen Bälle) dokumentiert, aber nur, wenn die Aufgabe auch bearbeitet wurde (z. B. wird beim M-ABC im Protokollbogen ein »V« für Verweigerung oder ein »B« für Beeinträchtigung markiert, wenn ein definierter Durchlauf gar nicht beendet wurde).

Im Leistungssport werden z. B. im Rahmen der Talententwicklung oder der Talentselektion motorische Leistungstests eingesetzt, allerdings häufig nicht im Sinne eines differenzierten Motoriktests mit mehreren Aufgaben, sondern es werden einzelne Testaufgaben (20 Meter Sprint zur Messung der