BSB Bayerische StaatsBibliothek
MDZ Münchener DigitalisierungsZentrum Digitale Bibliothek

Allgemeine Psychologie und deduktivistische Methodologie mit ... 22 Tabellen

Göttingen 2004

2010.26559
urn:nbn:de:bvb:12-bsb00050138-3

Die PDF-Datei kann elektronisch durchsucht werden.

Edgar Erdfelder / Joachim Funke (Hg.)

Allgemeine Psychologie und deduktivistische Methodologie

Mit 11 Abbildungen und 22 Tabellen

Vandenhoeck & Ruprecht

2010.
26559

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

ISBN 3-525-46210-7

© 2004 Vandenhoeck & Ruprecht, Göttingen / www.v-r.de
Alle Rechte vorbehalten. Das Werk und seine Teile sind urheberrechtlich geschützt. Jede Verwertung in anderen als den gesetzlich zugelassenen Fällen bedarf der vorherigen schriftlichen Einwilligung des Verlages. Hinweis zu § 52a UrhG: Weder das Werk noch seine Teile dürfen ohne vorherige schriftliche Einwilligung des Verlages öffentlich zugänglich gemacht werden. Dies gilt auch bei einer entsprechenden Nutzung für Lehr- und Unterrichtszwecke.
Printed in Germany.
Satz: Satzspiegel, Nörten-Hardenberg
Druck und Bindung: Hubert & Co., Göttingen

Gedruckt auf alterungsbeständigem Papier.

G10

Bayerische
Staatsbibliothek
München

Inhalt

Vorwort 7

Teil I: Methodologie 13

Volker Gadenne: Die deduktivistische Methodologie: Grundlagen, Probleme und Weiterentwicklung 15
Arndt Bröder: Interne und externe Validität aus deduktivistischer Sicht: Versuch einer Systematisierung und Bestandsaufnahme 37
Rainer Westermann: Strukturalistische Theorienkonzeption und deduktive Forschungsmethodologie 56
Willi Hager: Psychologische und statistische Hypothesen: Ableitung von Vorhersagen und Entscheidungen 72
Edgar Erdfelder: Angewandte deduktivistische Methodologie in der empirischen Psychologie 98

Teil II: Methoden 125

Rolf Steyer: Was wollen und was können wir durch empirische Kausalforschung erfahren? 127
Edgar Erdfelder, Axel Buchner, Franz Faul und *Martin Brandt:* GPOWER: Teststärkeanalysen leicht gemacht 148
Klaus Willmes: Permutations- und Randomisierungstests 167
Matthias Siemer: Teststärkekonzept und Teststärkebestimmung bei Randomisierungstests 191

Teil III: Anwendungen 205

Cristina Massen: Vorteile deduktiver Forschungsstrategien am Beispiel eines Paradigmas aus der Sakkadenforschung 207
Silke Hamm: Prognose der Kurzzeitgedächtnisspanne durch verschiedene Invarianzhypothesen 225

Bianca Vaterrodt-Plünnecke: Die Prozess-Dissoziations-Prozedur mit der lexikalischen Entscheidungsaufgabe: Bestandsaufnahme und Validierung eines multinomialen Modells 245
Karl Christoph Klauer: Deduktivistische Ansätze beim deduktiven Schlussfolgern 265
Joachim Funke: Komplexes Problemlösen: Möglichkeiten deduktivistischen Vorgehens 281
Edeltrud Marx: Methoden der Erhebung von Versprecherdaten in der Sprachproduktionsforschung: Probleme und Lösungsansätze 301

Teil IV: Über Jürgen Bredenkamp 319

Kurzbiografie von Jürgen Bredenkamp 321
Veröffentlichungen von Jürgen Bredenkamp 322
Liste der Promovenden von Jürgen Bredenkamp 327

Die Autorinnen und Autoren 331

Vorwort

Methodologie – die Lehre von den wissenschaftlichen Methoden – ist in der psychologischen Forschungspraxis allgegenwärtig und doch zumeist unsichtbar. Allgegenwärtig ist sie, weil jeder konkrete Forschungsprozess unvermeidlich mit einer Vielzahl methodischer Entscheidungen verknüpft ist. Unabhängig davon, ob sich Forschende der Entscheidungsoptionen in allen Phasen des Forschungsprozesses bewusst sind, ob sie ihre Entscheidungen sorgfältig und unter Abwägung der damit verbundenen Vor- und Nachteile treffen oder aber mehr oder minder unreflektiert methodischen Konventionen folgen: Man hat sich mit der Realisierung bestimmter Methoden und der Nichtrealisierung denkbarer alternativer Methoden automatisch für eine Methodologie entschieden, sei diese über verschiedene Entscheidungen hinweg nun konsistent realisiert oder eher hybrider Natur. Die Festlegung auf eine Methodologie lässt sich also in der Wissenschaft nicht vermeiden; zur Disposition steht lediglich, ob diese Festlegung explizit und unter rationaler Erwägung der damit verbundenen Konsequenzen oder aber implizit und unreflektiert erfolgt.

Unsichtbar ist Methodologie in der Regel deshalb, weil Begründungen für die Wahl bestimmter Methoden sowie Diskussionen der Konsequenzen dieser Wahl in psychologischen Publikationen typischerweise unerwähnt bleiben. Folgt man dem Willen der Herausgeber vieler Fachzeitschriften, so soll das vermutlich auch so sein. Nicht selten werden in Publikationsrichtlinien und Editorials wissenschaftliche Veröffentlichungen bis hin zur Festlegung auf bestimmte Konventionen der Datenanalyse (z. B. Melton, 1965) und der Ergebnisdarstellung (z. B. Hunt, 1990) reglementiert und normiert, sodass wenig Freiraum für eigene Entscheidungen oder gar methodische Erörterungen bleibt. Dies hat zwar immer wieder zu Debatten über die tradierten Konventionen und denkbare Alternativen geführt, zuletzt in der Zeitschrift »Psychological Science« (z. B. Abelson, 1997; Hunter, 1997; Shrout, 1997; s. auch Loftus, 1996). Geändert hat diese bislang letzte Signifikanztestdebatte an der Publikationspraxis allerdings ebenso wenig wie ihre Vorgängerdebatten.

Mit der vorliegenden Edition soll der skizzierte Rahmen in zweierlei Hinsicht durchbrochen werden, sodass die Bedeutung der Methodologie für die Forschungspraxis sichtbar wird. Zum einen soll die Frage einer vernünftigen Methodologie aus verschiedenen Perspektiven ausführlich diskutiert und ohne Rücksichtnahme auf übliche methodische Konventionen beantwortet werden. Zum anderen soll der Nachweis erbracht werden, dass *nichts so praktisch ist wie eine gute Methodologie* – um es in Abwandlung eines bekannten Zitats von Kurt Lewin (1982, S. 217) auszudrücken. Gemeint ist damit, dass eine explizit formulierte Methodologie, deren Prämissen akzeptabel sind, ganz praktische Entscheidungshilfen bei allfälligen methodischen Entscheidungen in der Forschungspraxis bietet und somit unmittelbar anwendungsrelevant ist.

Die deduktivistische Methodologie psychologischer Forschung, die wir in den Mittelpunkt dieser Edition stellen, geht zurück auf Arbeiten von Jürgen Bredenkamp (1969, 1972, 1980) und Volker Gadenne (1976, 1984). Sie basiert auf dem kritischen Rationalismus Karl Poppers (1934/1982), welcher der induktiven Methodologie der logischen Empiristen eine strikt deduktive Methode der Theorienprüfung entgegengestellt hat: Nicht die induktive Herleitung von theoretischen Aussagen aus Beobachtungsaussagen, sondern umgekehrt die deduktive Herleitung von Beobachtungsaussagen aus universellen theoretischen Aussagen bildet demnach das einzig verlässliche Fundament der empirischen Wissenschaften. Aus einer zu prüfenden Theorie werden vor allem solche singulären Beobachtungsaussagen deduziert, die zu Konkurrenztheorien im Widerspruch stehen, zumindest aber aus ihnen nicht deduzierbar sind. Über diese Prognosen wird dann anhand von Beobachtungen entschieden. »Fällt die Entscheidung positiv aus, werden die singulären Folgerungen anerkannt, *verifiziert*, so hat das System die Prüfung vorläufig bestanden; wir haben keinen Anlaß, es zu verwerfen. Fällt eine Entscheidung negativ aus, werden Folgerungen *falsifiziert*, so trifft ihre Falsifikation auch das System, aus dem sie deduziert wurden. Die positive Entscheidung kann das System immer nur vorläufig stützen; es kann durch spätere negative Entscheidungen immer wieder umgestoßen werden. Solange ein System eingehenden und strengen deduktiven Nachprüfungen standhält . . ., sagen wir, daß es sich *bewährt*« (Popper, 1934/1982, S. 8).

Die Übertragung der Popperschen Falsifikationstheorie auf die psychologische Forschung wirft, vor allem wegen des Problems der Kontrolle von Störeinflüssen und der damit zusammenhängenden inferenzstatistischen Analyse der Ergebnisse, Schwierigkeiten auf. Gadenne (1976) hat gezeigt, dass die in der Psychologie über viele Jahre dominierende Methodologie von Campbell und Stanley (1963), welche die Forderung nach interner und

externer Validität in den Mittelpunkt stellt, mit Induktionsproblemen behaftet ist und deshalb als Ausgangspunkt für eine deduktive Methodologie ausscheidet. Bredenkamp (1969, 1972) und Gadenne (1976) gelang es jedoch, wichtige Grundzüge einer Methodologie auszuarbeiten, die ohne induktive Elemente auskommt. Später wurden diese Ideen von ihnen selbst und anderen – darunter auch Autoren dieser Edition – zu einer deduktivistischen Methodologie weiterentwickelt.

Das vorliegende Buch gliedert sich in drei Teile: I. Methodologie, II. Methoden und III. Anwendungen. In Teil I wird der gegenwärtige Stand der deduktivistischen Methodologie dargestellt und aus verschiedenen Perspektiven diskutiert. *Volker Gadenne* (Linz) gibt einen Überblick über Grundlagen, Probleme und Weiterentwicklungen der deduktivistischen Methodologie, *Arndt Bröder* (Bonn) diskutiert das aus der Campbell-Stanley-Methodologie stammende Konzept der externen Validität im Lichte einer deduktivistischen Methodologie und *Rainer Westermann* (Greifswald) betrachtet die deduktivistische Methodologie aus der Perspektive einer strukturalistischen Theorienkonzeption. Unterschiedliche Vorstellungen darüber, woran man die Adäquatheit der Ableitung statistischer Hypothesen aus psychologischen Hypothesen festmachen kann, bilden den Gegenstand des Beitrags von *Willi Hager* (Göttingen). *Edgar Erdfelder* (Mannheim) behandelt die konkrete Anwendung der deduktivistischen Methodologie in der Forschungspraxis und schließt damit den methodologischen Teil ab.

Teil II der Edition behandelt Methoden, die im Rahmen einer deduktivistischen Methodologie von zentraler Bedeutung sind: Methoden der Prüfung von Kausalhypothesen, Randomisierungstests sowie Methoden der Teststärkekontrolle. *Rolf Steyer* (Jena) zeigt, wie Hypothesen über kausale Effekte von unabhängigen auf abhängige Variable unterschiedlich strengen Überprüfungen zugeführt werden können, die über die Prüfung reiner Zusammenhangshypothesen weit hinausgehen. Verschiedene Techniken der Teststärkeanalyse und ihre Umsetzung in der Forschungspraxis behandeln *Edgar Erdfelder, Axel Buchner, Franz Faul* und *Martin Brandt* (Mannheim, Düsseldorf und Kiel). Teststärkeanalysen spielen im Rahmen einer deduktivistischen Methodologie eine sehr wichtige Rolle, weil erst sie eine strenge Prüfung vieler psychologischer Hypothesen mittels statistischer Hypothesentests ermöglichen (s. bereits Bredenkamp, 1969, 1972). Ähnliches gilt für Randomisierungstests, die – wie *Klaus Willmes* (Aachen) zeigt – ideal auf Hypothesen und Stichprobenverfahren der experimentellen Psychologie zugeschnitten sind und zur Lösung einer Vielzahl unterschiedlicher Forschungsprobleme eingesetzt werden können. Ein lange Zeit nicht befriedi-

gend gelöstes Problem – die notwendige Teststärkekontrolle bei Randomisierungstests – behandelt *Matthias Siemer* (Greifswald).

Gegenstand von Teil III der Edition sind Anwendungen der deduktivistischen Methodologie in verschiedenen Disziplinen der Allgemeinen Psychologie. Unter »Allgemeiner Psychologie« wird hierbei mit Bredenkamp (1969, 1972) das Teilgebiet der Psychologie verstanden, das sich mit der Prüfung von psychologischen Allsätzen über offene Mengen von Individuen beschäftigt. Inhaltlich stammen die Beispiele aus den Bereichen Aufmerksamkeit, Gedächtnis, Denken und Problemlösen sowie Sprache und veranschaulichen damit die breite Anwendbarkeit des deduktivistischen Ansatzes. *Cristina Massen* (Bonn) erläutert deduktivistisches Vorgehen beispielhaft in der Blickbewegungsforschung. Sie demonstriert die gegenüber herkömmlichen »Effektforschungen« bestehenden Vorteile anhand von Experimenten über den Zusammenhang von kontrollierter Aufmerksamkeit, Arbeitsgedächtnis und fluider Intelligenz. *Silke Hamm* (Bonn) befasst sich mit einem Vergleich verschiedener Invarianzhypothesen bezüglich der Prognosegüte für die Kurzzeitgedächtnisspanne. *Bianca Vaterrodt-Plünnecke* (Bonn) greift die deduktivistische Vorgehensweise anhand der Prozess-Dissoziations-Prozedur auf. Die Validierung eines multinomialen Modells für lexikalische Entscheidungsaufgaben und dessen Anwendungen stehen im Mittelpunkt ihres Beitrags. *Christoph Klauer* (Bonn) befasst sich mit deduktivistischen Ansätzen beim deduktiven Schlussfolgern. Anhand von Beispielen des propositionalen und syllogistischen Schließens wird verdeutlicht, welche Schwierigkeiten in diesem Bereich bei der Inferenz von Theorie auf die Daten auftreten und mit welchen Methoden man diese Schwierigkeiten überwinden kann. *Joachim Funke* (Heidelberg) behandelt Problemlösungsforschung auf deduktivistischer Basis. Die Vorteile einer solchen Vorgehensweise im Vergleich zu »Schrotschuss«-Forschungsstrategien werden durch Beispiele aus dem Bereich komplexer Probleme illustriert. *Edeltrud Marx* (Köln) integriert schließlich Ergebnisse aus Sprachwahrnehmungsexperimenten, Sprechfehlersammlungen sowie Sprechfehlerinduktionsexperimenten und diskutiert Möglichkeiten der Reliabilitäts- und Validitätssteigerung von Versprecherdaten, sodass eine strengere Hypothesenprüfung in diesem Bereich möglich wird.

Diese einleitende Übersicht wäre unvollständig, würden sie eine wichtige Gemeinsamkeit aller in diesem Buch versammelten Beiträge unerwähnt lassen: Sie alle stammen von Personen, die in unterschiedlicher Weise von den Gedanken und Werken einer anderen Person beeinflusst wurden – *Jürgen Bredenkamp*. Seine Vorstellungen, Anregungen und Ideen, die seit den sechziger Jahren auf verschiedene Bereiche der Allgemeinen Psychologie und der

Psychologischen Methodenlehre Einfluss genommen haben, spiegeln sich in den vorliegenden Beiträgen mehr oder weniger explizit wider. Für viele der Autorinnen und Autoren dieses Bandes ist Jürgen Bredenkamp auch ihr akademischer Lehrer und Mentor, eine Persönlichkeit, die tragfähige wissenschaftliche Prinzipien vermittelt und veranschaulicht, wesentliche Denkanstöße gegeben und ihre Umsetzung tatkräftig unterstützt hat. Mit diesen Beiträgen wollen wir ihm anlässlich seines 65. Geburtstags und seiner Emeritierung unseren Dank zum Ausdruck bringen – dies auch im Namen derjenigen Schülerinnen, Schüler, Mitarbeiterinnen und Mitarbeiter, die in diesem Band nicht als Autorinnen und Autoren vertreten sind. Jürgen Bredenkamps Lebenslauf, ein Verzeichnis seiner Publikationen und eine Liste seiner akademischen Schülerinnen und Schüler schließen diesen Band ab.

Im Zuge der Herstellung dieses Bandes sind wir vielen Personen zu Dank verpflichtet. Neben den Autorinnen und Autoren, die nicht nur ihre Beiträge fristgerecht verfasst, sondern auch im internen Review-Verfahren gegenseitig konstruktive Kritik geäußert haben, danken wir auch den externen Reviewern aus Bonn (Dipl.-Psych. Stefanie Schiffer), Heidelberg (Dipl.-Psych. Carsten Schulz, Dipl.-Psych. Miriam Spering, Dipl.-Psych. Safir Yousfi) und Mannheim (Dipl.-Psych. Lutz Cüpper, cand. psych. Roan de Bock, Dr. Jochen Musch, cand. psych. Stephan Rückert). In unseren Sekretariaten haben Frau Gustava Heß und Frau Brigitte Krieg hilfreich am Zustandekommen dieses Bandes mitgewirkt, ebenfalls die Studentinnen Anna Frommherz, Silja Taxis, Monika Undorf und Sandra Zillig. Schließlich ist Herrn Dr. Harry Lundberg (Göttingen) zu danken, der den Kontakt zu unserem Verlag hergestellt hat. Sein Einsatz für unser Projekt hat geholfen, sowohl dem Band die repräsentative äußere Form zu geben, die wir uns gewünscht haben, als auch die Produktion zeitnah sicherzustellen.

Edgar Erdfelder und Joachim Funke
Mannheim und Heidelberg

Literatur

Abelson, R.P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15.

Bredenkamp, J. (1969). Über die Anwendung von Signifikanztests bei theorie-testenden Experimenten. *Psychologische Beiträge, 11,* 275–285.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Campbell, D.T., Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Hg.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Gadenne, V. (1976). *Die Gültigkeit psychologischer Untersuchungen.* Stuttgart: Kohlhammer.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung.* Tübingen: Mohr.

Hunt, E. (1990). The traditional editorial statement. *Journal of Experimental Psychology: General, 119,* 3–4.

Hunter, J.E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7.

Lewin, K. (1982). *Feldtheorie* (= Kurt-Lewin-Werkausgabe, Band 4). Bern: Huber.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

Melton, A.W. (1965). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Popper, K.R. (1934/1982). *Logik der Forschung* (7. Auflage). Tübingen: Mohr.

Shrout, P.E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1–2.

Teil I:
Methodologie

Volker Gadenne

Die deduktivistische Methodologie: Grundlagen, Probleme und Weiterentwicklung

Zusammenfassung

In seinen Arbeiten zur psychologischen Methodenlehre ging Jürgen Bredenkamp von der deduktivistischen Methodologie Poppers aus. Letztere ist nun eingehend kritisiert worden und wird heute von vielen als unhaltbar erachtet. Einer der Hauptkritikpunkte lautet, dass Popper kein überzeugendes Kriterium dafür angegeben hat, wann eine Theorie definitiv als falsifiziert zu beurteilen ist. Ein anderer Einwand besagt, dass die deduktivistische Methodologie ein verstecktes induktives Prinzip enthält und ohne ein solches auch nicht auskommen kann. Die Auseinandersetzung mit diesen und weiteren Einwänden reicht bis in die heutige Zeit. Im Beitrag wird zu zeigen versucht, dass einige Kritikpunkte am Deduktivismus berechtigt sind, dass es aber eine neuere, weiterentwickelte Version dieser Methodologie gibt, die sich mit überzeugenden Gründen verteidigen lässt.

Einleitung

Karl Poppers deduktivistische Methodologie, die in seiner Auseinandersetzung mit dem logischen Positivismus, insbesondere mit dem Wiener Kreis, entstand, spielte für die Entwicklung der modernen Wissenschaftstheorie eine zentrale Rolle. In den sechziger Jahren war sie Gegenstand wissenschaftstheoretischer Debatten, an denen unter anderem Kuhn, Lakatos und Feyerabend mitwirkten. Um diese Zeit bestand auch in den Sozialwissenschaften ein großes Interesse an der Wissenschaftstheorie. 1961 kam es zu der Kontroverse zwischen dem kritischen Rationalismus und der kritischen Theorie der Frankfurter Schule, die später unter der Bezeichnung »Positivismusstreit« veröffentlicht wurde. In der Psychologie nahm zunächst Holzkamp (1964) auf Popper Bezug, jedoch eher am Rande, da er sich vor allem am Konstruktivismus Dinglers orientierte.

Die detaillierte Auseinandersetzung mit dem Deduktivismus und seine Anwendung auf die Psychologie begann Ende der sechziger Jahre. Jürgen Bredenkamp kommt das Verdienst zu, hierbei die Pionierarbeit geleistet zu

haben. Auf der Grundlage des Deduktivismus hat er die psychologische Methodenlehre, vor allem den Signifikanztest und die Versuchsplanung, analysiert, Probleme aufgezeigt und zahlreiche konstruktive Vorschläge für das methodische Vorgehen und die Interpretation empirischer Resultate gemacht (Bredenkamp, 1969, 1972, 1980; Erdfelder u. Bredenkamp, 1994). Poppers Wissenschaftstheorie fordert dazu auf, die Methoden und Vorgehensweisen in den empirischen Wissenschaften auf ihre Eignung zum Erreichen des gesetzten Erkenntnisziels hin zu reflektieren. Dementsprechend begann Bredenkamp beispielsweise seine Untersuchung des Signifikanztests nicht mit Lehrsätzen und Beweisen der mathematischen Statistik, sondern mit der Überlegung, was für Arten von Aussagen es sind, für die sich die Psychologie eigentlich interessiert. Wie hängen diese inhaltlich-psychologischen Aussagen mit den statistischen Hypothesen zusammen, die üblicherweise getestet werden? Ist der Signifikanztest geeignet, zur Widerlegung oder Bewährung der psychologischen Hypothesen beizutragen? Wie könnte man ihn modifizieren, damit er diesem Ziel besser dient? Bredenkamps Ergebnisse haben dazu beigetragen, die empirisch-psychologische Forschung durchsichtiger zu machen und zu verbessern, und sie haben viele andere dazu angeregt, sich ähnlichen Problemstellungen zu widmen, teils ebenfalls auf der Grundlage der deduktivistischen Methodologie (dies gilt für meine Arbeiten von 1976 und 1984), teils von anderen wissenschaftstheoretischen Konzeptionen ausgehend (z. B. Westermann, 1987; Hager, 1992).

Poppers kritizistische Lehre ist nun bekanntlich selbst einer eingehenden Kritik unterzogen worden und viele halten diese Kritik für vernichtend. Popper habe niemals ein überzeugendes Kriterium dafür angegeben, wann eine Theorie definitiv als falsifiziert zu beurteilen sei, lautet ein Einwand. Ein anderer besagt: Entgegen Poppers Meinung enthält seine Konzeption der Bewährung von Theorien ein verstecktes induktives Prinzip, und ohne ein solches kann sie auch gar nicht auskommen. Die Auseinandersetzung mit diesen und weiteren Einwänden reicht bis in die heutige Zeit (Böhm et al., 2002). Was ist das Ergebnis? Kann die deduktivistische Methodologie noch verteidigt werden? Wenn ja, in welcher Form? Um diese Fragen wird es im Folgenden gehen, wobei zunächst die Problematik der Falsifikation und danach die der Bewährung und Induktion behandelt wird. Ich will zu zeigen versuchen, dass einige Kritikpunkte berechtigt sind, dass es aber eine neuere, weiterentwickelte Version der deduktivistischen Methodologie gibt, die sich mit überzeugenden Gründen verteidigen lässt.

Die Grundzüge der deduktivistischen Methodologie

Eine prägnante Zusammenfassung seiner Methodologie gab Popper gleich auf den ersten Seiten seiner *Logik der Forschung* (im Folgenden kurz *Logik*): »Aus dem System (gemeint sind Theorien als Aussagensysteme, V. G.) werden (unter Verwendung bereits anerkannter Sätze) empirisch möglichst leicht nachprüfbare bzw. anwendbare singuläre Folgerungen (›Prognosen‹) deduziert und aus diesen insbesondere jene ausgewählt, die aus bekannten Systemen nicht ableitbar sind bzw. mit ihnen in Widerspruch stehen. Über diese – und andere – Folgerungen wird nun im Zusammenhang mit der praktischen Anwendung, den Experimenten usw., entschieden. Fällt die Entscheidung positiv aus, werden die singulären Folgerungen anerkannt, *verifiziert*, so hat das System die Prüfung vorläufig bestanden; wir haben keinen Anlass, es zu verwerfen. Fällt eine Entscheidung negativ aus, werden Folgerungen *falsifiziert*, so trifft die Falsifikation auch das System, aus dem sie deduziert wurden.

Die positive Entscheidung kann das System immer nur vorläufig stützen; es kann durch spätere negative Entscheidungen immer wieder umgestoßen werden. Solange ein System eingehenden und strengen deduktiven Nachprüfungen standhält und durch die fortschreitende Entwicklung der Wissenschaft nicht überholt wird, sagen wir, dass es sich *bewährt*.

Induktionslogische Elemente treten in dem hier skizzierten Verfahren nicht auf; niemals schließen wir von der Geltung der singulären Sätze auf die der Theorien. Auch durch ihre verifizierten Folgerungen können Theorien niemals als ›wahr‹ oder auch nur als ›wahrscheinlich‹ erwiesen werden« (Popper, 1935/1994, S. 8).

Wegen der zentralen Rolle der Falsifikationsidee wurde diese Methodologie später meist *Falsifikationismus* genannt. Die andere Bezeichnung *Deduktivismus* hebt das deduktive Prüfverfahren und die Ablehnung der Induktion hervor.

Die deduktivistische Methodologie steht in unmittelbarem Zusammenhang mit dem von Popper aufgeworfenen *Abgrenzungsproblem*: Wodurch zeichnet sich *empirische Wissenschaft* aus? Was unterscheidet sie einerseits von Mathematik und Logik, andererseits von Metaphysik sowie von Pseudowissenschaft? Es kann nicht die *Verifizierbarkeit* von Theorien mithilfe der induktiven Methode sein, wie die logischen Positivisten glaubten, denn induktive Schlüsse lassen sich nicht rational rechtfertigen. Noch so viele Beobachtungen weißer Schwäne können die Aussage »Alle Schwäne sind weiß« nicht zwingend als wahr erweisen. Und ein entsprechender induktiver Schluss lässt sich auch nicht dadurch rechtfertigen, dass man auf die

erfolgreiche Verwendung der Induktion in der Vergangenheit verweist. Denn dieses Argument hätte selbst die Form eines induktiven Schlusses, würde also dasjenige schon voraussetzen, für das argumentiert werden soll. In Bezug auf diese Problematik übernahm Popper die Kritik David Humes an der Induktion.

Als Abgrenzungskriterium schlug Popper stattdessen vor: Es ist die *Widerlegbarkeit*, die *Falsifizierbarkeit* von Hypothesen und Theorien anhand von Beobachtungen, die empirische Wissenschaft auszeichnet. Eine empirisch-wissenschaftliche Theorie muss an der Erfahrung scheitern können.

Er betonte in diesem Zusammenhang, dass es nicht nur darauf ankommt, Hypothesen eine für ihre Falsifizierbarkeit erforderliche logische Form zu geben (z. B. keine Tautologien, keine allgemeinen Es-gibt-Sätze), sondern auch darauf, dass man sie *aktiv zu falsifizieren* versucht und darauf verzichtet, sie angesichts widerlegender Befunde dogmatisch zu verteidigen. Um dies zu gewährleisten, entwarf er ein System *methodologischer Regeln*. Das deduktive Prüfverfahren bildet das Grundgerüst dieser Methodologie. Es lässt dem Vorgehen aber einen sehr weiten Spielraum und daher wird es ergänzt durch einzelne Regeln, die nähere Empfehlungen zu speziellen Fragen geben. Eine dieser Regeln verbietet etwa Ad-hoc-Maßnahmen zur Verteidigung von Theorien: Man sollte nicht eine Theorie verteidigen, indem man einen widersprechenden empirischen Befund durch ad hoc eingeführte Annahmen als irrelevant erklärt, etwa durch die Annahme, es hätte im Experiment »Störungen« unbekannter Art gegeben, die nachträglich nicht nachweisbar seien. (Einen systematischen Überblick über die methodologischen Regeln gibt Jarvie, 2001.)

Ergänzt wird diese Methodologie durch das erkenntnistheoretische Prinzip des *Fallibilismus*, dem in Poppers Lehre eine zentrale Bedeutung zukommt: Alle Annahmen gelten als grundsätzlich *fehlbar* und daher *revidierbar*. Es gibt in der Erkenntnis (und auch beim übrigen Problemlösen) keine Gewissheit. Dies gilt für wissenschaftliche Theorien ebenso wie für Beobachtungsaussagen.

Wenn Beobachtungsaussagen ebenso fehlbar sind wie allgemeine Hypothesen, welchen Zweck hat es dann, sie zum Test von Hypothesen heranzuziehen? Die Antwort lautet, dass man sich über Beobachtungsaussagen, obwohl sie fehlbar sind, dennoch leichter einig wird als über allgemeine Hypothesen. Wissenschaftler können sich gut darüber einig werden, dass das Display eines Messinstruments die Zahl 15 anzeigt oder dass eine Testperson bei einer bestimmten Frage die Antwort »ja« angekreuzt hat. Sie streiten sich aber möglicherweise darüber, welche Theorie die beste Erklärung für die vorliegenden Daten darstellt. Es ist sinnvoll, gemäß der metho-

dologischen Regel zu verfahren, die schwieriger zu beurteilenden Aussagen anhand der leichter zu beurteilenden zu prüfen. Beobachtungsaussagen gelten als akzeptabel, solange keine Hinweise vorliegen, dass sie auf einer Täuschung beruhen. Wenn es aber solche Hinweise gibt, dann dürfen auch Beobachtungsaussagen in Zweifel gezogen und anhand neuer Beobachtungen überprüft werden. Poppers Methodologie nimmt an, dass es auch ohne eine absolut gesicherte Beobachtungsbasis möglich ist, aus Beobachtungen zu lernen beziehungsweise sie effektiv zur Korrektur von Hypothesen heranzuziehen.

Wann genau soll eine Theorie als falsifiziert gelten? Gibt es dazu eine methodologische Regel? Popper legte sich nicht exakt fest, drückte aber in *Logik* einen grundsätzlichen Optimismus dahingehend aus, dass Falsifikationsentscheidungen gut begründbar sind: »Wir betrachten also im allgemeinen eine (methodisch entsprechend gesicherte) intersubjektiv nachprüfbare Falsifikation als endgültig . . .« (S. 214). Diese Aussage ist etwas vage, und der Grund dafür ist verständlich: Eine sehr strenge Festlegung könnte einen dogmatischen Falsifikationismus ergeben, der dem Problem der Fehlbarkeit von Beobachtungsaussagen nicht Rechnung trägt. Bei einer zu liberalen Fassung der Methodologie läuft man dagegen Gefahr, dass diese zu einer leeren Angelegenheit wird, indem sie alles zulässt.

Zum richtigen Verständnis von Poppers Denken muss man nun wissen, dass es in ihm eine Entwicklung gibt, die unter anderem auch eine gemäßigtere Auffassung der Falsifikation mit sich brachte. In den vierziger Jahren begann Popper, sozialphilosophische Theorien zu behandeln, und in diesem Zusammenhang entwickelte er die Auffassung, dass man nicht nur empirische, sondern auch *philosophische* (z. B. erkenntnistheoretische, metaphysische, sozialphilosophische) Aussagen rational diskutieren und bewerten kann (vgl. dazu Wettersten, 1992). Dies eröffnete ihm die Möglichkeit, den erkenntnistheoretischen *Realismus* und die *Korrespondenztheorie der Wahrheit* zu verteidigen, gegenüber denen er in *Logik* noch eine Neutralität bewahrt hatte. Damit vollzog er den Übergang vom frühen Falsifikationismus, der noch durch die vom Positivismus beeinflusste Distanz zur Metaphysik geprägt war, zum *kritischen Rationalismus.* Letzterer umfasst nach Albert (1982, S. 9 ff.) den *Fallibilismus*, den *Realismus* und den *methodischen Rationalismus.* Der Fallibilismus wurde schon erläutert. *Realismus* ist die Position, dass es eine objektive Welt gibt, die unabhängig von unserem Wahrnehmen und Denken existiert, und dass wir dazu in der Lage sind, zumindest über Teile dieser Realität ein (fehlbares) Wissen zu erlangen. Der *methodische Rationalismus* ist die Ausdehnung der Idee der kritischen Prüfung, des kritischen Argumentierens auf alle Problemlösungen,

sei es in der Wissenschaft, der Moral, dem Recht oder der Politik. Er verlangt, dass man die jeweils vorgeschlagenen Problemlösungen der Kritik unterzieht und vergleicht, welche dabei am besten abschneiden. Die deduktivistische Methodologie kann als die Ausgestaltung des methodischen Rationalismus für das Gebiet der empirischen Wissenschaften verstanden werden.

Diese Darstellung könnte den Eindruck erzeugen, als sei der kritische Rationalismus eine sehr einheitliche Position, aber so ist es nicht. Die heutigen Vertreter des kritischen Rationalismus, zu denen etwa Agassi (1975), Albert (1968), Andersson (1988), Jarvie (2001), Miller (1994) und Musgrave (1999) gehören, akzeptieren zwar alle den Fallibilismus und den Realismus und hier gibt es keine wesentlichen Unterschiede. Aber wie wir noch sehen werden, haben sie zum Teil verschiedene Vorstellungen darüber, wie der methodische Rationalismus umzusetzen ist.

Im Rahmen der geschilderten Entwicklung änderte Popper auch seine Auffassung über die Möglichkeit und Strenge von Falsifikationsentscheidungen. Grob gesprochen wurde der Falsifikationismus gemäßigter. Im Postskript zu *Logik*, das Popper Anfang der fünfziger Jahre verfasste, betonte er, dass es endgültige Falsifikationen *nicht* gebe (1983, S. xxii ff.). Damit widersprach er wörtlich seiner früheren Aussage; er hat dies jedoch als »Klarstellung« ausgegeben. Es gibt weitere Äußerungen aus seinen späteren Schriften, aus denen hervorgeht, dass er Falsifikationsentscheidungen als schwierig und keineswegs sehr eindeutig ansah. Er sagte, »dass wir nicht sämtliche Immunisierungen ausschließen dürfen, nicht einmal alle jene, die ad hoc Hilfshypothesen einführen«; und er gestand zu, »dass ein gewisses Maß an Dogmatismus sogar in der Wissenschaft fruchtbar ist« (1979, S. 53).

Für die folgende Analyse ist es wichtig, die einzelnen Bestandteile des kritischen Rationalismus auseinander zu halten. Das *deduktive Prüfverfahren*, für sich allein genommen, ist niemals infrage gestellt worden. Es war im Übrigen lange vor Popper bekannt, dass es sinnvoll ist, Hypothesen anhand von Aussagen zu prüfen, die deduktiv aus ihnen folgen. In einem weiten Sinne macht auch jede statistische Testtheorie von dieser Idee Gebrauch. Weiterhin steht das deduktive Prüfverfahren allein nicht in Widerspruch zu induktiven Prinzipien. J. S. Mill beispielsweise vertrat eine bestimmte Form der Induktion, wies aber auch der deduktiven Prüfung eine wichtige Rolle zu. Bei Popper kommen zum deduktiven Verfahren die *Ablehnung der Induktion* hinzu sowie die *Regeln der Falsifikation*, wodurch aus diesem Verfahren der *Deduktivismus* wird. Wenn von Letzterem gesprochen wird, ist leider oft nicht klar, ob dabei die frühere, strengere oder die spätere, gemäßigtere Version gemeint ist. Halten wir den Unterschied noch-

mals fest. Die frühere Auffassung hält endgültige Falsifikationen für möglich, die zwar nicht garantiert irrtumsfrei sind (dies wäre dogmatischer Falsifikationismus und würde dem Fallibilismus zuwider laufen), die aber »im Allgemeinen« nicht angezweifelt werden müssen. Poppers spätere Auffassung sieht Falsifikationsentscheidungen nicht mehr als (praktisch) endgültig an. Und von den genannten heutigen Vertretern des kritischen Rationalismus hält keiner mehr die Suche nach einem Falsifikationskriterium für ein wichtiges Thema. Damit wird die deduktivistische Methodologie in ihren Regeln allerdings unbestimmter, wie wir sehen werden.

Die Kritik an der Falsifikationsidee

Ein Teil der Einwände gegen den Deduktivismus nimmt Bezug auf die Wissenschaftsgeschichte (Kuhn, 1967; Feyerabend, 1976, Kap. 5). Es wird darauf verwiesen, dass bedeutende Wissenschaftler an ihren Theorien festgehalten haben, obwohl diese eindeutig mit anerkannten Fakten in Widerspruch standen. Aus späterer Sicht stellte sich dieses Ignorieren der Fakten als vorteilhaft für den Erkenntnisfortschritt heraus. Nach Poppers Lehre hätten diese Wissenschaftler ihre Theorien jedoch angesichts der widersprechenden Fakten längst aufgeben müssen.

Andere Einwände sind methodologischer Natur und knüpfen vor allem an eine Einsicht an, die schon Pierre Duhem hatte und die später als Duhem-Quine-These bekannt geworden ist (Duhem, 1908, S. 243 ff.; Gadenne, 1998; Harding, 1976). Duhem zeigte, dass bei der Prüfung wissenschaftlicher Theorien die Prüfaussage P so gut wie immer aus mehreren Aussagen einer Theorie T, bestimmten Anfangsbedingungen A und Hilfsannahmen H (etwa über den Aufbau der Untersuchung und über das richtige Funktionieren der Messinstrumente) deduziert wird. Manchmal enthalten die Hilfsannahmen sogar ganze Theorien, etwa die Optik als Grundlage des Teleskops. Sollte nun die Untersuchung das Ergebnis $A \wedge \neg P$ erbringen, so widerspricht dies nicht T allein, geschweige denn einer einzelnen Aussage von T, sondern dem ganzen Aussagensystem $T \wedge H$. Welche Aussage soll man dann als falsifiziert betrachten und verwerfen? Dies ist das Duhemsche Problem.

Mit Blick auf die Psychologie und wohl auch viele andere empirische Wissenschaften kann man das Duhemsche Problem noch um zwei Punkte erweitern. Erstens sind die zu prüfenden Aussagen oft *Ceteris-paribus-Hypothesen*. In diesem Fall gehört zu den vorausgesetzten Hilfsannahmen je-

weils eine Ceteris-paribus-Annahme, die behauptet, dass in der gegebenen Untersuchungssituation außer den kontrollierten Variablen keine weiteren eine Rolle spielen (Gadenne, 1984, 1998). Zweitens wird die zu prüfende Theorie eventuell statistische Aussagen enthalten oder auf dem Umweg über statistische Hypothesen geprüft werden (Bredenkamp, 1972, S. 10 ff.) und dann kommen statistische Entscheidungsregeln zur Anwendung. Die damit verbundenen Entscheidungsrisiken bilden, zusätzlich zu den eventuell falschen Hilfsannahmen, eine weitere Quelle möglicher Irrtümer.

Es gibt nun angesichts des Duhemschen Problems durchaus eine systematische Vorgehensweise, die sich der deduktiven Prüfmethode bedient. Sie garantiert zwar keine kurzfristige Lösung, verspricht aber dennoch einen Fortschritt auf längere Sicht: Man ändert versuchsweise eine bestimmte Annahme aus T oder H und untersucht, ob nun zutreffende Prüfvorhersagen resultieren. Wir ändern beispielsweise die in Zweifel gezogene Theorie T_1 und machen daraus T_2. Wir kombinieren beide mit denselben Hilfsannahmen und Anfangsbedingungen derart, dass aus $T_1 \wedge H \wedge A$ die Prüfvorhersage P_1 folgt, während aus $T_2 \wedge H \wedge A$ die von P_1 abweichende Vorhersage P_2 ableitbar ist. Wenn in diesem Fall P_1 verworfen und P_2 akzeptiert wird, so haben wir einen weiteren Hinweis darauf gewonnen, dass T_1 falsch ist.

Zwingend falsifizieren kann dieses Vorgehen eine Theorie nicht. Es liegt hier zwar nahe, das Ergebnis P_2 als Bewährung von T_2 anzusehen. Es ist aber sehr wohl möglich, dass trotz dieses Resultats T_1 wahr ist und irrtümlich als falsifiziert erklärt wird. Zum Beispiel könnte H falsch sein und zusammen mit T_1 die falsche Vorhersage P_1 implizieren, während T_2, sei es wahr oder falsch, zusammen mit H die wahre Aussage P_2 impliziert (Musgrave, 1999, S. 196). Allerdings kann es einen Fortschritt durchaus auch ohne ein Kriterium für eine endgültige Falsifikation geben. Davon war Duhem übrigens ganz überzeugt. Er vertraute auf den »gesunden Menschenverstand«. Was passiert, wenn Forschergruppen sich darüber streiten, ob eine Theorie zugunsten einer anderen als falsifiziert erklärt werden soll? »Eines Tages erklärt sich der gesunde Menschenverstand so deutlich zugunsten einer der beiden Parteien, dass die andere auf den Kampf verzichtet, wenn auch die reine Logik seine Fortsetzung nicht verbieten würde« (Duhem, 1908, S. 292). Wissenschaftler fahren nicht endlos darin fort, eine Beobachtungsaussage oder die mit ihr verbundenen Hilfshypothesen ständig neu zu debattieren. Irgendwann einigt man sich. Wann dies geschieht, kann nicht durch eine methodologische Regel vorgeschrieben werden, aber dies ist auch nicht notwendig.

Lakatos (1974) hat mit seiner *Methodologie wissenschaftlicher Forschungsprogramme* ebenfalls eine Antwort auf die von Duhem aufgeworfe-

ne Frage gegeben. Ein Forschungsprogramm nach Lakatos enthält jeweils einen *harten Kern* und einen *Schutzgürtel*. Der harte Kern entspricht dem, was wir oben die Theorie T nennen, und der Schutzgürtel entspricht den Hilfsannahmen H. Wenn nun das System T ∧ H empirisch geprüft wird und sich Daten einstellen, die ihm widersprechen, so verlangt Lakatos' Methodologie, dass H geändert werden muss. T ist gegen Änderung tabu. Das Gesamtsystem T ∧ H kann falsifiziert und durch ein verbessertes T ∧ H' ersetzt werden, aber die Korrektur muss an H vorgenommen werden. Erst, wenn es einmal nachhaltig misslingt, den angesammelten empirischen Befunden durch Anpassungen am Schutzgürtel Rechnung zu tragen, wenn das Forschungsprogramm *degeneriert*, dann stellt sich die Option, T und damit das gesamte Programm aufzugeben.

Mit dem Fallibilismus steht dieser Lösungsversuch jedoch nicht in Einklang, denn danach kann es nicht gerechtfertigt werden, T, den harten Kern, zeitweilig für tabu gegen Kritik zu erklären. Warum sollte man dies auch tun? Es ist denkbar, dass ein Wissenschaftler frühzeitig, das heißt solange das Forschungsprogramm zutreffende neue Prognosen erzeugt, eine geniale Idee zur Änderung von T hat, etwa zur Änderung eines Gesetzes von T. Die Wissenschaftsgeschichte belegt übrigens nicht, dass Wissenschaftler bei »großen« Theorien stets so verfahren, wie Lakatos' Methodologie es nahe legt. So gab es wiederholt Versuche, Newtons Theorie durch Änderung der zentralen Gesetze und nicht des Schutzgürtels an die empirischen Fakten anzupassen (Musgrave, 1999, S. 212 ff.).

Doch wenn man einmal von dieser Abweichung der Methodologie wissenschaftlicher Forschungsprogramme vom Deduktivismus absieht: Gibt sie denn eine Antwort auf das von Duhem aufgezeigte Problem? Sie gibt eine Antwort auf der einen Ebene, die aber zur Folge hat, dass das Problem auf einer anderen Ebene wieder auftaucht (Feyerabend, 1976, S. 256 ff.). Die Falsifikation ist gegen H und nicht gegen T zu richten, dies wird auf der ersten Ebene vorgeschrieben. Die nächste Frage ist: Wie viel Kredit soll T gegeben werden, bis das Programm als degeneriert erklärt und T aufgegeben werden soll? Dafür gibt es in Lakatos' Methodologie keine genaue Angabe. Er lässt es zu, dass sich Programme, die als degeneriert erscheinen, wieder erholen.

Diese Überlegungen zeigen, dass es nicht möglich ist, von einer Theorie zwingend nachzuweisen, dass sie falsch ist. Dies bedeutet nicht, dass das deduktive Prüfverfahren überhaupt keinen Wert hätte. Aber was folgt nun speziell für Poppers Lehre? Der in *Logik* enthaltene Falsifikationismus wird von den Einwänden getroffen. Popper sprach dort von »endgültigen« Falsifikationen. Seine Aussage war etwas unbestimmt, aber wie auch immer

man sie auslegt, sie ist im Licht der Wissenschaftsgeschichte und des Duhem-Problems sicherlich zu optimistisch gewesen. Falsifikationen können nicht endgültig sein, auch nicht »im Allgemeinen«. Der spätere Deduktivismus sagt nun explizit, dass sie nicht zwingend oder endgültig sein können, er wird also durch die Einwände nicht berührt. Doch ist er mit einem anderen Problem konfrontiert. Kuhn (1974, S. 16) hat es folgendermaßen auf den Punkt gebracht: »Was ist eine Falsifikation, wenn nicht eine zwingende Widerlegung? ... Solange die Aufklärung dieser Frage in der Schwebe gelassen wird, bin ich mir dessen gar nicht sicher, ob wir Sir Karl überhaupt eine Logik der Forschung zu verdanken haben.« Feyerabend (1981, S. 328 f.) hat denselben Gedanken noch polemischer ausgedrückt: »Entweder ist der kritische Rationalismus ... eine Philosophie, die die Forschung in eine bestimmte Richtung zu lenken vermag, weil ihre Regeln oder Maßstäbe in den Forschungsbetrieb eingreifen; oder er ist leeres Geschwätz, das sich mit jedem Vorgehen vereinbaren läßt.«

Was bleibt von der Falsifikationsidee?

Betrachtet man einzelne Prinzipien des Deduktivismus getrennt, so fällt sogleich auf, dass die angeführte Kritik den *Fallibilismus* nicht trifft, sondern sogar unterstützt. Was man aus der Diskussion der Duhem-Quine-These vor allem lernen kann, deckt sich zu einem großen Teil mit dem, was der Fallibilismus für die wissenschaftliche Forschung nahe legt: Hypothesen und Theorien können nicht bewiesen werden, Beobachtungsaussagen und Hilfsannahmen aber auch nicht, und daher sind auch Falsifikationsurteile fehlbar.

Aber kann der Fallibilismus die Forschung »lenken«? Allein hat er keine sehr spezifischen Konsequenzen, wenn er auch manche Urteile und Vorgehensweisen auszuschließen vermag wie: T stellt den endgültigen Stand unserer Wissenschaft dar und braucht nicht weiter überprüft werden. Der Fallibilismus bewahrt vor Dogmatismus, er regt an zu einem empfehlenswerten Maß an Skepsis und zu pluralistischem Denken (Albert 1968, S. 49). Auf jeden Fall ist der Fallibilismus allein aber noch keine Methodologie.

Auch das *deduktive Prüfverfahren* selbst gibt der Forschung eine grobe Orientierung, denn es verlangt, dass 1) aus Theorien Beobachtungsaussagen deduziert, 2) Beobachtungen angestellt und 3) Widersprüche zwischen Theorie, Hilfsannahmen und Beobachtungsergebnissen beseitigt werden. Dies sind freilich sehr geringe Anforderungen. Sie gehen zwar über ein

»anything goes« hinaus, da sie aber von anderen Methodologien der empirischen Wissenschaften auch erhoben werden, können sie nicht das Spezifische des Deduktivismus ausmachen.

Nun besteht dieser aber auch nicht aus dem Fallibilismus und der Methode der deduktiven Prüfung allein, sondern enthält eine Reihe weiterer methodologischer Regeln. Jarvie (2001) zählt 15 solcher Regeln auf. Wir können sie hier nicht alle einer detaillierten Betrachtung unterziehen. Ich möchte aber eine bestimmte Regel diskutieren, die mir als die wichtigste erscheint. In *Logik* (S. 16) wird gefordert, ein zu überprüfendes Annahmensystem *in jeder Weise einer Falsifikation auszusetzen.* An anderer Stelle fordert Popper »ernsthafte« oder »strenge« Tests von Theorien. Was soll dies heißen? Es sind im groben Rahmen der deduktiven Prüfung verschiedene Strategien des Vorgehens denkbar und der Falsifikationismus schlägt vor, bestimmte von ihnen zu bevorzugen. Man soll solche Folgerungen testen, »die aus bekannten Systemen nicht ableitbar sind bzw. mit ihnen in Widerspruch stehen« (1935/1994, S. 8). Beispielsweise fand eine strenge Prüfung der allgemeinen Relativitätstheorie statt, als am 29. Mai 1919 zwei britische Expeditionen während einer Sonnenfinsternis die Sterne in der Nähe der Sonne fotografierten, mit dem Resultat, dass das Licht dieser Sterne durch die Gravitation der Sonne um einen Winkel abgelenkt wurde, der mit Einsteins Theorie vereinbar ist, mit Newtons jedoch nicht. Es handelte sich um eine kühne, für Einsteins Theorie riskante Vorhersage, denn nach der Theorie Newtons, die damals als ungeheuer gut gesichert galt, hätte die aus Einsteins Theorie abgeleitete Vorhersage nicht eintreten dürfen.

Popper versuchte, die Strenge eines Prüfversuchs und in Verbindung damit, den Grad der Bewährung einer Theorie formal zu definieren (1935/1994, Neuer Anhang, Abschnitt *IX; 1997, S. 565 ff.). Seine Vorschläge haben sich aber als nicht haltbar herausgestellt (Vetter, 1967; Kutschera, 1972). Eher lässt sich ein qualitativer, eventuell auch ein komparativer Begriff der strengen Prüfung und der Bewährung einführen. Kutschera (1972, S. 407 ff.) machte einen entsprechenden Vorschlag. Er definierte den Begriff der Bewährung folgendermaßen: Eine Theorie T bewährt sich an einem empirischen Befund P bezüglich des (mit T konsistenten) Hintergrundwissens HW, wenn 1) P aus T $\wedge$ HW deduzierbar ist; 2) P aus HW allein nicht deduzierbar ist. (Kutschera formuliert drei Bedingungen; ich habe seine dritte Bedingung, dass nämlich T mit H konsistent ist, bereits in die Voraussetzungen der Definition aufgenommen.)

Dass P aus T $\wedge$ HW deduzierbar ist, entspricht der Forderung, dass bei einer deduktiven Prüfung P aus T $\wedge$ H $\wedge$ A abgeleitet wird, denn HW umfasst unter anderem die Randbedingungen A und die Hilfsannahmen H.

Die beiden Bedingungen Kutscheras bringen zum Ausdruck, dass P eine *nichttriviale* Folgerung aus T ist, eine Vorhersage, deren Eintreffen nicht schon unabhängig von T erwartet werden kann. Legen demnach die Bedingungen (1) und (2) fest, was ein strenger Test ist (der dann, wenn P anschließend als zutreffend befunden wird, in eine Bewährung von T einmündet)? Am Ende zweifelt Kutschera doch an, dass seine Definition die Idee eines »ernsthaften« oder »strengen« Widerlegungsversuchs ausdrücken könne, weil Popper selbst gesagt habe, dass diese Idee nicht formal definierbar sei. Tatsächlich sagte Popper einmal, dass man den Gedanken eines »ernstgemeinten und gut ausgedachten Widerlegungsversuches« nicht vollständig formalisieren könne (1935/1994, S. 354).

Ich denke aber, dass sich diese Idee doch ausreichend klar fassen lässt. Die von Kutschera angegebenen Bedingungen sind dazu aber nicht ausreichend. In ihnen fehlen zwei wichtige Aspekte strenger Prüfungen: dass P nicht aus *konkurrierenden*, mit T logisch unvereinbaren Theorien deduzierbar sein darf und dass P, wenn möglich, sogar einer konkurrierenden Theorie *widersprechen* sollte. Die Gesamtheit dieser Forderungen kann man *nicht* allein mithilfe des Begriffs des Hintergrundwissens HW ausdrücken. Die Forderung, dass P nicht aus HW ableitbar ist, kann diese weiteren Forderungen nicht schon enthalten, da die mit T konkurrierenden Theorien nicht Teil von HW sein können. HW muss nämlich, nach Voraussetzung, mit T konsistent sein, sonst kann T nicht mit HW zur Ableitung von P kombiniert werden. Wenn aber HW eine mit T konkurrierende (und T widersprechende) Theorie T' enthält, dann ist T $\wedge$ HW nicht mehr konsistent.

Ich habe vor längerer Zeit (1976, S. 54 ff.) einen Vorschlag zur Lösung dieses Problems gemacht. Heute scheint mir eine andere Lösung besser, nämlich die zu bestimmende Idee in zwei Komponenten aufzuteilen (Gadenne, 2002). Dies ist nicht nur aus logischen Gründen, sondern auch deshalb von Vorteil, weil daraus *zwei* methodologische Regeln resultieren, die unterschiedlichen Zwecken dienen. Es geht *erstens* darum, dass P nicht aus von T verschiedenen Annahmen ableitbar sein sollte, die zu HW gehören. Nur dann handelt es sich um einen *ernsthaften Test*. Denn wenn P auch ohne T vorhergesagt werden kann, und wenn P darüber hinaus aufgrund des vorhandenen Wissens sogar erwartet werden kann, dann wird T hier *keinem echten Risiko* ausgesetzt. Dass so etwas zu vermeiden ist, dies gilt in den empirischen Wissenschaften als selbstverständlich. Kontrolltechniken wie die Elimination, Konstanthaltung und Randomisierung dienen diesem Zweck. Eine Untersuchung, die sich um eine derartige Kontrolle von »Störfaktoren« überhaupt nicht kümmert, wird nicht als wissenschaftlich anerkannt.

Zweitens kann man *kritische Tests* durchführen, man kann Theorien *gegeneinander* testen, sie empirisch miteinander *konfrontieren*. »Kritisch« soll hier zum Ausdruck bringen, dass eine Folgerung von T gesucht wird, die aus der Sicht einer anderen Theorie oder des etablierten Wissens falsch ist. Aus T wird eine Folgerung P gezogen, die einer konkurrierenden Theorie T' widerspricht, eventuell einer älteren, etablierten Theorie. Ein kritischer Test ist der Idee des »experimentum crucis« verwandt. Es muss sich jedoch nicht um eine experimentelle Testsituation handeln. Außerdem ist das »experimentum crucis« von dem Gedanken getragen, eine der beteiligten Theorien zu beweisen, während bei einem kritischen Testen davon ausgegangen wird, dass beide Theorien falsch sein könnten.

Im Unterschied zur Durchführung eines korrekten Versuchsplans, der den üblichen Standards genügt und damit einen ernsthaften Test darstellt, handelt es sich bei einem kritischen Test nicht um eine Routineangelegenheit, sondern um etwas, das Fantasie erfordert (Poppers »gut ausgedachter Widerlegungsversuch«). In den empirischen Wissenschaften wird ein kritisches Testen nicht ausdrücklich gefordert, im Unterschied zur Durchführung ernsthafter Tests.

Die empirische Konfrontation von Theorien bringt einen großen Gewinn: Das empirische Resultat spricht entweder gegen T, und dies ist ein Gewinn, denn es kann es nur von Vorteil sein, die Fehler innerhalb von neuen Annahmensystemen herauszufinden. Oder das Ergebnis spricht für T und damit gegen T', und in diesem Fall haben wir herausgefunden, dass in einer anderen, eventuell gut etablierten Theorie ein Irrtum enthalten ist. Unter der Voraussetzung des Fallibilismus, dass wir stets mit Irrtümern rechnen müssen, und mit der Zielsetzung, diese aufzufinden und zu beseitigen, kann man ein solches Vorgehen nur positiv beurteilen. Wer überhaupt empirische Forschung betreibt und aus der Empirie lernen will, hat Grund, so vorzugehen.

Wie erscheinen diese Regeln aus der Perspektive der dargestellten Kritik am Deduktivismus? Sie sind von den dort aufgeführten Einwänden nicht betroffen. Diese Regeln fordern nicht dazu auf, einen konträren empirischen Befund sofort zum Anlass zu nehmen, T aufzugeben. Sie zielen wohl darauf ab, konträre Tatsachen zu finden, sei es gegen T oder gegen ein altes Annahmensystem. Ob und wann T zu verwerfen ist, lässt sich aber nicht methodologisch fixieren.

Es zeigt sich, dass der Deduktivismus auch ohne Kriterium einer gesicherten Falsifikation methodologische Regeln enthält, die der empirischen Forschung Empfehlungen geben können. Die deduktive Methode gibt einen weiten Rahmen vor. Sie muss stets in der von Duhem beschriebenen

Problemsituation operieren, die viele Wege zulässt. Die beiden Regeln, ernsthafte Tests und, wenn möglich, kritische Tests durchzuführen, geben genauere Hinweise. Sie lenken die Fantasie der Forscher beim Ausdenken und Planen von empirischen Untersuchungen in eine bestimmte Richtung, sie sind also heuristischen Prinzipien ähnlich. Albert (1982, S. 52) fasst Methodologie als *rationale Heuristik* auf. Ein Kriterium für endgültige Falsifikationen gibt es nicht. Aber auch ohne ein solches Kriterium können heuristische Prinzipien nützlich sein.

Ich habe oben angedeutet, dass es im Deduktivismus einige weitere methodologische Regeln gibt. Die Forderung nach einem kritischen Testen ist nicht die einzige, die sinnvoll erhoben werden kann, andere Gesichtspunkte müssen hinzukommen, insbesondere, wenn man es mit Ceteris-paribus-Aussagen zu tun hat (vgl. dazu Gadenne, 1984, Kap. 4 bis 6). Wenn die zu prüfende Hypothese statistischer Natur ist, oder aber durch Ableitung einer statistischen Hypothese getestet wird, ist die Problemsituation noch komplexer. Hier kann man durch Entscheidungsregeln die Risiken unterschiedlicher Arten von Fehlern steuern und man sollte diese so wählen, dass falsche Hypothesen nach Möglichkeit scheitern und wahre Hypothesen nach Möglichkeit den Test bestehen. Westermann (1987), Hager (1987) sowie Erdfelder und Bredenkamp (1994) folgern daraus, dass Hypothesentests nicht nur streng, sondern auch *fair* sein müssen. Auf diese und viele andere Fragen, die sich bei der Anwendung methodologischer Prinzipien auf die Psychologie ergeben, kann ich hier leider nicht näher eingehen.

Bewährung, Induktion und rationales Fürwahrhalten

Auch im Zusammenhang mit der Idee der *Bewährung* sind gegen den Deduktivismus wichtige Einwände erhoben worden (z. B. Salmon, 1968; Lakatos, 1968; Schurz, 2002). Das Problem, um das es geht, lässt sich folgendermaßen entwickeln: Nehmen wir an, T ist zunächst noch ungeprüft, hat dann einen ersten Test bestanden und somit eine erste Bewährung erlangt, und zu einem noch späteren Zeitpunkt gilt T schließlich als hochgradig bewährt. Haben wir nun am Ende mehr Grund zu der Annahme, dass T wahr ist, als zu Anfang? Und dürfen wir uns stärker darauf verlassen, dass T zutreffende Vorhersagen macht? Wenn der Deduktivismus diese Fragen verneint, so muss man anzweifeln, dass die empirischen Tests irgendeinen Zuwachs an Erkenntnis erbracht haben, und die hohe Bewährung von T wäre dann auch ohne jeden praktischen Nutzen. Antwortet der Deduktivismus

aber mit Ja, so erlaubt er ein Urteil, das aus der deduktiven Logik allein nicht folgt und offenbar eine Art induktiven Schließens darstellt, denn zutreffende Vorhersagen in vergangenen Tests sagen deduktiv-logisch nichts darüber aus, ob zukünftige Vorhersagen mit T zutreffen.

Der Deduktivismus scheint in diesem Punkt in Schwierigkeiten zu geraten. Insbesondere in der Praxis stellt sich die Frage, ob man sich auf eine bestimmte, bewährte Theorie verlassen soll, oder, falls mehrere Theorien zur Auswahl stehen, an welcher man sich beim Handeln orientieren soll. Im praktischen Leben verlassen sich Menschen auf Hypothesen, die sich in der Vergangenheit bestätigt haben. Wenn wir nicht aus dem Fenster eines Hochhauses springen, um auf schnellstem Weg nach unten zu gelangen, so hat dies unter anderem den Grund, dass wir die Hypothese für wahr halten, dass solche Stürze lebensgefährlich sind. Wie kann der Deduktivismus dieser Praxis Rechnung tragen, ohne Prinzipien der rationalen Rechtfertigung und Auswahl einzuführen?

Auch was die Methodologie und den Erkenntnisfortschritt angeht, gibt es ein Problem: Die Annahme, dass das Befolgen bestimmter methodologischer Regeln (wie kritisches Testen) den Fortschritt fördert, dass es zu dem Ziel beiträgt, im Lauf der Zeit unter den nicht falsifizierten Theorien möglichst viele wahre Theorien zu haben, lässt sich nicht aus der deduktiven Logik ableiten. Welchen Status hat diese Annahme? Ist sie nicht einem induktiven Prinzip ähnlich?

Mit diesen Fragen konfrontiert, äußerte sich Popper folgendermaßen: Er erklärte es als rational gerechtfertigt, unter konkurrierenden Theorien die bestbewährte bis auf weiteres *für die der Wahrheit nächste* zu halten (1973, S. 97). Außerdem formulierte er ein Prinzip der *pragmatischen Bevorzugung*: Vorziehen als Grundlage für unsere Handlungen sollten wir jeweils die bestbewährte Theorie (1973, S. 34). Er sagte weiterhin, Bewährung könne als »Anzeichen« für die »Annäherung an die Wahrheit« dienen (1973, S. 97). Und er machte auch einmal das Zugeständnis, dass im Deduktivismus ein »Hauch« von Induktion enthalten sei (1974, S. 1192).

An anderer Stelle widersprach er dem jedoch und betonte, Bewährung sage »nicht das geringste über die zukünftigen Leistungen oder die ›Verlässlichkeit‹ einer Theorie« (1973, S. 30). Und zu glauben, dass eine Theorie wahr sei, dazu gäbe es »ganz einfach *keinen* Grund« (1983, S. 67). Dies ist einigermaßen verwirrend. Betrachten wir, wie seine Nachfolger mit dieser Problematik umgehen.

Der kritische Rationalismus ohne Rechtfertigungsprinzip

Agassi äußerte als Erster den Gedanken, dass die Idee der Bewährung einen Rest von »Rechtfertigungsdenken« darstellt, den es zu überwinden gelte. Auf jeden Fall sei Bewährung in der Grundlagenforschung verzichtbar, allenfalls in der Technologie könne sie eine Rolle spielen (Agassi, 1985). Bartley (1987) befürwortete einen »pankritischen« Rationalismus, der die in der Tradition für grundlegend gehaltene Frage »Wie kann man diese Aussage begründen?« schlichtweg für nicht beantwortbar erklärt und durch die Frage ersetzt: »Wie kann man diese Aussage am besten kritisieren?« Rationalität sei mit *Offenheit für Kritik* und kritischem Vorgehen gleichzusetzen.

Am radikalsten wird dieser Lösungsansatz von David Miller vertreten (1994, 2002). Nach dieser Auffassung gibt der kritische Rationalismus jeglichen Versuch auf, Aussagen zu begründen oder rechtfertigen zu wollen, sei es durch hinreichende oder nicht hinreichende Gründe. Es gibt *keine guten Gründe* für irgendeine Annahme. Auch singuläre (Beobachtungs-)Aussagen können nicht gerechtfertigt werden. Die Klassifikation von Aussagen in wahre und falsche wird nicht dadurch erreicht, dass man bestimmte Aussagen als gestützt oder irgendwie gerechtfertigt erweist. Sie wird erreicht, indem man bestimmte Aussagen ohne Rechtfertigung als falsch klassifiziert. Hypothesen und Theorien werden in den Bereich der empirischen Wissenschaften aufgenommen, wenn sie mit potenziellen Beobachtungsaussagen in Widerspruch geraten können. Wenn eine Hypothese geprüft wird und die Prüfung besteht, geschieht mit ihr nichts. Der positive Ausgang des Tests verändert an ihrem Status nichts. Sie verbleibt im Bereich der Wissenschaft, ohne so etwas wie Stützung oder Glaubwürdigkeit gewonnen zu haben. Wenn sie den Test nicht besteht, wird sie aus dem Bereich der Wissenschaft entfernt. Es ist hierbei wichtig, dass Verfahren zur Kritik und Verwerfung von Hypothesen bei jeder möglichen Gelegenheit ins Spiel gebracht werden: Die offene Aufnahme in den Bereich der Wissenschaft (ohne Forderung irgendeiner Art von Stützung) verlangt rigorose Eliminierungsverfahren. Es gibt durchaus »Wissen« oder »Erkenntnis« in Gestalt von Theorien, nur hat und behält dieses Wissen grundsätzlich den Status von *Rateversuchen*.

Ist diese Lösung des Induktionsproblems zufrieden stellend? Ist sie von einem Skeptizismus zu unterscheiden? Sind hier nicht alle Theorien *gleich ungerechtfertigt*? Es gibt zwar angeblich den Unterschied zwischen falsifizierten und nicht falsifizierten Theorien, der die Möglichkeit eröffnet, nicht falsifizierte Theorien den falsifizierten vorzuziehen. Aber auch dieses »Vorziehen« scheint beliebig zu sein, da keine Möglichkeit vorgesehen ist, falsifizierende Beobachtungsaussagen rational zu rechtfertigen.

Und was sagt diese Methodologie zu den Problemen des praktischen Handelns? Man verlässt sich in einem mehr oder weniger hohen Grad auf die Theorie, von der man sein Handeln leiten lässt. Wer eine Brücke baut, einen Ofen beheizt, Essen zubereitet, verlässt sich bewusst oder automatisch darauf, dass gewisse Gesetzmäßigkeiten weiter bestehen. Kann man dies bestreiten?

Miller (1994, S. 38 ff.) geht diesen Einwand so an, dass er zunächst das seiner Meinung nach naive Bild menschlicher Praxis, von dem hier ausgegangen wird, bestreitet: Die Anwendung von Theorien besteht nicht darin, dass eine bestimmte Theorie als die glaubwürdigste, am besten gerechtfertigte ausgewählt wird, um dann aus ihr Folgerungen für das Handeln abzuleiten. Wissenschaftliche Theorien geben überhaupt keine Handlungsanweisungen, sondern haben die logische Form von Verboten, sie sagen uns, welche Wege nicht funktionieren. Praxis geht in Wirklichkeit so vor sich: Es werden Vorschläge zur Lösung eines anstehenden Problems gemacht und kritisch diskutiert. Hierbei können wissenschaftliche Theorien in kritischen Argumenten vorkommen. Manche der Vorschläge werden am Ende »überleben«, der Kritik besser standhalten als andere. Es ist rational, denjenigen Vorschlägen zu folgen, die der Kritik am besten standhalten. Bei diesem Verfahren werden keinerlei Theorien als »verlässlich« vorausgesetzt, und das Problem einer Rechtfertigung von Theorien stellt sich nicht. Miller hält alle Einwände, wonach auch eine deduktivistische Position auf Induktion nicht völlig verzichten könnte, für ein grundlegendes Missverständnis. Dass so lange Zeit nach Hume manche immer noch diesem Missverständnis erliegen, hält er für einen »Skandal der Philosophie« (2002, S. 81).

Ich denke, dass es teilweise richtig ist, was Miller über Praxis sagt. Man findet zu einem praktischen Problem im Allgemeinen nicht gleich *die* passende Theorie und deduziert daraus die Lösung. Das Rechtfertigungsproblem taucht aber auch in der Millerschen Schilderung von Praxis wieder auf. Angenommen, jemand schlägt die Errichtung eines Gebäudes vor, und der Vorschlag wird kritisch diskutiert. Gegen das Projekt wird der Einwand erhoben, dass das Gebäude in der geplanten Form nicht stabil genug sei, und es wird ein Gegenvorschlag gemacht, der allerdings höhere Kosten verursacht. Größere Stabilität erfordert höhere Kosten, daher müssen die Kritiker gut argumentieren, und sie versuchen dies mithilfe von Berechnungen auf der Grundlage der Gesetze der Mechanik. Nehmen wir an, dass die Partei sich durchsetzt, die mithilfe der Gesetze der Mechanik den ersten Vorschlag kritisiert hat. Wir haben hier die Anwendung einer Theorie zur Kritik eines praktischen Vorschlags, so wie Miller es anspricht. Und dennoch ist das Rechtfertigungsproblem unschwer zu erkennen. Wenn die Beteilig-

ten nämlich für den zweiten Vorschlag stimmen, so tun sie dies deshalb, weil sie glauben, dass die Gesetze der Mechanik gelten und auch in Zukunft noch gelten werden. Sie verlassen sich auf bewährte wissenschaftliche Theorie.

Der kritische Rationalismus mit Rechtfertigungsprinzip

Alan Musgrave (1993, 1999) vertritt, gegen Bartley und Miller, folgende Auffassung: Wir können Theorien nicht begründen, wir können aber begründen oder rechtfertigen, dass wir an einige Theorien glauben oder sie für wahr halten, nämlich jeweils die zu einer bestimmten Frage am besten bewährte Theorie. Nennen wir dies das kritisch-rationale *Rechtfertigungsprinzip* (kurz *R*) für Hypothesen beziehungsweise Theorien: Es ist vernünftig, (zum Zeitpunkt t) diejenige Theorie für wahr zu halten (zu glauben), die (zum Zeitpunkt t) am besten bewährt ist.

Wenn es auf einem bestimmten Gebiet nur eine vorgeschlagene Theorie gibt, die ernsthafte Tests bestanden hat, dann ist sie die bestbewährte Theorie und man darf sie vorläufig für wahr halten und zur Anwendung empfehlen. Dieses Urteil ist natürlich revidierbar, da sich Gegenbefunde einstellen können. Gibt es mehrere konkurrierende Theorien, so sollten diese kritischen Tests unterzogen werden, bis sich deutliche Unterschiede in der Bewährung zwischen ihnen ergeben. Hat T_2 einige kritische Tests bestanden, an denen T_1 gescheitert ist, so hat T_2 den höheren Bewährungsgrad. Wenn es zwei konkurrierende Theorien gibt, die beide eine Reihe von kritischen Tests bestanden haben, so sollte man ein Urteil aufschieben, bis es durch kritische Tests gelungen ist, einige falsifizierende Befunde zu einer der beiden zu gewinnen.

Musgrave meint, dass dies die zentrale Idee des kritischen Rationalismus ist und zugleich zusammenfasst, was Popper dazu gesagt hat. Letzteres ist jedoch nicht zutreffend. Die Aussage Poppers, die *R* am nächsten kommt, lautet, dass es gerechtfertigt ist, *die bestbewährte Theorie* für die *der Wahrheit nächste* zu halten. Aus heutiger Sicht erscheint diese Lösung allerdings eher problematischer als diejenige, die Musgrave vorschlägt (und Popper zuschreibt). Es gibt nämlich leider bisher keine überzeugende Theorie der Wahrheitsnähe (vgl. zu diesem Problem Kuipers, 1987). Der Begriff der Wahrheit ist weniger problematisch als der der Wahrheitsnähe, und dies ist ein Grund, eher *R* als Poppers Rechtfertigungsprinzip zu akzeptieren.

Warum meinte Popper, dass es in diesem Zusammenhang besser sei, von

Wahrheitsnähe als von Wahrheit zu sprechen? Er glaubte, dass sich auf lange Sicht jede Theorie als falsch erweisen würde: Keine Theorie stimmt vollkommen mit der Wirklichkeit überein. Popper (1983, S. 25 f.) verwies darauf, dass selbst Einstein seine spezielle und seine allgemeine Relativitätstheorie nur als Annäherungen betrachtet habe. Wenn nun von keiner Theorie erwartet werden kann, dass sie im vollen Sinne des Worts wahr ist, kann Fortschritt nur darin bestehen, dass Theorien der Wahrheit näher kommen.

In der Tat müsste man es als unvernünftig ansehen, selbst die bestbewährte Theorie vorläufig für wahr zu halten, wenn es Grund gäbe, jede Theorie für falsch zu halten. Aber wie könnte es hierfür a priori einen Grund geben? Der Fallibilismus impliziert einen solchen Grund nicht. Er besagt, dass sich jede Theorie als falsch herausstellen *kann*, aber er besagt *nicht*, dass sie es wird. Die Wissenschaft hat eine Vielzahl an Theorien entwickelt, die bis heute noch nicht falsifiziert worden sind und die es möglicherweise niemals werden. Jedenfalls spricht die bloße Möglichkeit, dass eine Theorie sich als falsch herausstellen könnte, nicht dagegen, eine gut bewährte Theorie vorläufig für wahr zu halten.

Ist *R* ein *induktives* Prinzip? Man mag *R* »induktiv« nennen, weil es Urteile über die Glaubwürdigkeit und Vorzugswürdigkeit von Theorien erlaubt, die nicht aus akzeptierten Beobachtungsaussagen deduktiv ableitbar sind. Es spielt aber genau genommen keine Rolle, welchen Namen man der Sache gibt. Wichtig ist vielmehr, dass es sich *nicht* um Induktion handelt, wie sie in der Erkenntnistheorie überwiegend verstanden und kritisiert worden ist. *R* gehört *nicht* zu den induktiven Verfahren zur Entdeckung von Gesetzen (an die J. S. Mill glaubte). Es ist *kein* induktives Verfahren zum Wahrheitsnachweis von Gesetzen (das es nach Hume auch nicht geben kann). Und es ist *kein* Kalkül einer induktiven Wahrscheinlichkeit (wie Carnaps induktive Logik).

R unterscheidet sich aber auch von Theorien deduktiver Bestätigung, die allein mithilfe deduktiv-logischer Beziehungen zu definieren versuchen, unter welchen Bedingungen eine Hypothese relativ zu bestimmten Beobachtungssätzen als bestätigt gelten soll. *R* enthält nämlich ein *pragmatisches* Element, das nicht auf logische Beziehungen zwischen Sätzen zurückführbar ist: Die Bewährung resultiert aus *aktiven Versuchen*, die entsprechende Hypothese oder Theorie zu falsifizieren. Wenn etwas einer Theorie überhaupt Glaubwürdigkeit verleihen kann, dann ist es am ehesten der Umstand, dass sie solchen Widerlegungsversuchen standgehalten hat.

Von der Rechtfertigung der Induktion, die Westermann und Gerjets (1994) sowie Westermann (2000) vornehmen, unterscheidet sich *R* insofern, als diese Autoren für induktives Schließen nicht notwendigerweise

verlangen, dass zunächst eine allgemeine Hypothese aufgestellt und dann getestet wurde. Dagegen rechtfertigt *R* die Akzeptanz einer Hypothese nur für den Fall, dass sie einen Widerlegungsversuch überstanden hat (ernsthafter Test), und falls es konkurrierende Hypothesen gibt, dass sie sich in kritischen Tests als die beste erwiesen hat.

Kann es für *R* noch eine weitere Begründung geben? Muss *R* eventuell ohne Rechtfertigung akzeptiert werden? Worrall (1989) scheint dieser Auffassung zu sein. Musgrave (1999, S. 330) vertritt, dass *R* zwar gerechtfertigt werden kann, jedoch nur auf zirkuläre Weise: *R* ist akzeptabel, weil *R* der Kritik besser standhält als konkurrierende (induktive) Prinzipien, die bisher vorgeschlagen worden sind. Dieser Rechtfertigungsversuch macht selbst von einem Prinzip Gebrauch, das *R* sehr ähnlich ist: Es ist vernünftig, unter konkurrierenden Vorschlägen .empirischen Annahmen, methodologischen Prinzipien, moralischen Normen usw.) jeweils denjenigen vorläufig zu akzeptieren, der bislang der Kritik am besten standgehalten hat. Dies ist nichts anderes als der oben erwähnte *methodische Rationalismus*, und *R* ist dessen Konkretisierung für die Bewertung empirisch-wissenschaftlicher Hypothesen. Musgrave hält auf dieser Ebene der Analyse eine zirkuläre Rechtfertigung für besser als eine dogmatische Akzeptanz.

Kann man *R* eventuell *pragmatisch* rechtfertigen? Eine Rechtfertigung, die auf Nützlichkeit abzielt, ist natürlich ebenfalls zirkulär. Um die zukünftige Nützlichkeit von *R* zu belegen, würde man wiederum ein Prinzip wie *R* benötigen. Aber eine andere pragmatische Rechtfertigung ist vielleicht überzeugender, eine, die zunächst darauf verweist, dass etwas, das praktisch unvermeidlich ist, nicht irrational sein kann: Es ist in der Wissenschaft wie allgemein im Leben unvermeidlich, Handlungsentscheidungen zu treffen, und im Zusammenhang mit solchen Handlungsentscheidungen haben Menschen stets bestimmte Gesetzeshypothesen darüber, was zu was führt. Die Möglichkeit, solche Gesetzeshypothesen nicht zu haben, scheint uns nicht zur Wahl zu stehen. Daher kann es nicht als irrational erklärt werden, sich überhaupt an irgendeiner Hypothese zu orientieren. Nun kommt der zweite Teil der Argumentation: Wenn es richtig ist, dass Menschen immer die eine oder andere Hypothese als die glaubwürdigste akzeptieren, dann erscheint es rational, die Glaubwürdigkeit davon abhängig zu machen, wie intensiv eine Theorie auf Irrtümer hin getestet worden ist. Dass eine Theorie falsche Folgerungen enthält, findet man schließlich im Großen und Ganzen eher heraus, wenn man nach ihnen aktiv sucht, als wenn man Theorien gar nicht testet oder sie nur mit den Beobachtungen konfrontiert, die sich zufällig ergeben. (Dies kann man zu bestreiten versuchen, wird dabei aber argumentativ einen schweren Stand haben.) Wenn wir ohnehin

nicht vermeiden können, stets irgendetwas zu glauben, dann doch am besten das, was wir aktiv auf falsche Folgerungen hin überprüft und als haltbar befunden haben. Was sonst sollten wir vernünftigerweise tun?

Literatur

Agassi, J. (1975). *Science in flux.* Dordrecht: Reidel.

Agassi, J. (1985). *Technology: Philosophical and social aspects.* Dordrecht: Reidel.

Albert, H. (1968). *Traktat über kritische Vernunft.* Tübingen: Mohr-Siebeck.

Albert, H. (1982). *Die Wissenschaft und die Fehlbarkeit der Vernunft.* Tübingen: Mohr-Siebeck.

Andersson, G. (1988). *Kritik und Wissenschaftsgeschichte.* Tübingen: Mohr-Siebeck.

Bartley III, W.W. (1987). *Flucht ins Engagement.* Tübingen: Mohr-Siebeck.

Böhm, J.M., Holweg, H., Hoock, C. (Hg.) (2002). *Karl Poppers kritischer Rationalismus heute.* Tübingen: Mohr-Siebeck.

Bredenkamp, J. (1969). *Über die Anwendung von Signifikanztests bei theorietestenden Experimenten.* Psychologische Beiträge, 11, 275–285.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Duhem, P. (1908). *Ziel und Struktur der physikalischen Theorien.* Originalausgabe Leipzig. Nachdruck Hamburg: Meiner 1978.

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In Th. Herrmann u. W. Tack (Hg.), *Enzyklopädie der Psychologie. Methodologische Grundlagen der Psychologie* (S. 604–648). Göttingen: Hogrefe.

Feyerabend, P. (1976). *Wider den Methodenzwang.* Frankfurt a. M.: Suhrkamp.

Feyerabend, P. (1981). Rückblick. In H.P. Duerr (Hg.), *Versuchungen: Aufsätze zur Philosophie Paul Feyerabends* (S. 320–372). Frankfurt a. M.: Suhrkamp.

Gadenne, V. (1976). *Die Gültigkeit psychologischer Untersuchungen.* Stuttgart: Kohlhammer.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung.* Tübingen: Mohr-Siebeck.

Gadenne, V. (1998). Spielarten des Duhem-Quine-Problems. *Logos, 5,* 117–148.

Gadenne, V. (2002). Hat der kritische Rationalismus noch etwas zu lehren? In J.M. Böhm, H. Holweg, C. Hoock (Hg.), *Karl Poppers kritischer Rationalismus heute* (S. 58–78). Tübingen: Mohr-Siebeck.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen der Psychologie. In G. Lüer (Hg.), *Allgemeine Experimentelle Psychologie* (S. 43–264). Stuttgart: Fischer.

Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen.* Göttingen: Hogrefe.

Harding, S.G. (Hg.) (1976). *Can theories be refuted? Essays on the Duhem-Quine Thesis.* Dordrecht: Reidel.

Holzkamp, K. (1964). *Theorie und Experiment in der Psychologie.* Berlin: de Gruyter.

Jarvie, I.C. (2001). *The republic of science.* Amsterdam: Rodopi.

Kuhn, T.S. (1967). *Die Struktur wissenschaftlicher Revolutionen.* Frankfurt a. M.: Suhrkamp.

Kuhn, T.S. (1974). Logik der Forschung oder Psychologie der wissenschaftlichen Arbeit? In I. Lakatos u. A. Musgrave (Hg.), *Kritik und Erkenntnisfortschritt* (S. 1–24). Braunschweig: Vieweg.

Kuipers, T.A.F. (Hg.) (1987). *What is closer-to-the-truth?* Amsterdam: Rodopi.

Kutschera, F. v. (1972). *Wissenschaftstheorie.* Bd. 2. München: Fink.

Lakatos, I. (1968). Changes in the problem of inductive logic. In I. Lakatos (Hg.), *The problem of inductive logic* (S. 315–417). Amsterdam: North Holland.

Lakatos, I. (1974). Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme. In I. Lakatos, A. Musgrave (Hg.), *Kritik und Erkenntnisfortschritt* (S. 89–189). Braunschweig: Vieweg.

Miller, D. (1994). *Critical rationalism: A restatement and defence.* Chicago: Open Court.

Miller, D. (2002). Induction: A problem solved. In J.M. Böhm, H. Holweg u. C. Hoock (Hg.), *Karl Poppers kritischer Rationalismus heute* (S. 81–106). Tübingen: Mohr-Siebeck.

Musgrave, A. (1993). Popper on induction. *Philosophy of the Social Sciences, 23,* 516–527.

Musgrave, A. (1999). *Essays on realism and rationalism.* Amsterdam: Rodopi.

Popper, K. (1935/1994). *Logik der Forschung.* 10. Auflage. Tübingen: Mohr-Siebeck.

Popper, K. (1973). *Objektive Erkenntnis.* Hamburg: Hoffmann und Campe.

Popper, K. (1974). Replies to my critics. In P.A. Schilpp (Hg.), *The philosophy of Karl Popper* (S. 961–1197). La Salle, IL: Open Court.

Popper, K. (1979). *Ausgangspunkte.* Hamburg: Hoffmann und Campe.

Popper, K. (1983). *The postskript to the logic of scientific discovery. Vol. 1: Realism and the aim of science.* Totowa, NJ: Rowman and Littlefield.

Popper, K. (1994). *Vermutungen und Widerlegungen. Teilband I: Vermutungen.* Tübingen: Mohr-Siebeck.

Popper, K. (1997). *Vermutungen und Widerlegungen. Teilband II: Widerlegungen.* Tübingen: Mohr-Siebeck.

Salmon, W.C. (1968). The justification of inductive rules of inference. In I. Lakatos (Hg.), *The problem of inductive logic* (S. 24–43). Amsterdam: North Holland.

Schurz, G. (2002). Karl Popper, Deduktion, Induktion und Abduktion. In J.M. Böhm, H. Holweg u. C. Hoock (Hg.), *Karl Poppers kritischer Rationalismus heute* (S. 126–143). Tübingen: Mohr-Siebeck.

Vetter, H. (1967). *Wahrscheinlichkeit und logischer Spielraum.* Tübingen: Mohr-Siebeck.

Westermann, R. (1987). *Strukturalistische Theorienkonzeption und empirische Forschung in der Psychologie.* Berlin: Springer-Verlag.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik.* Göttingen: Hogrefe.

Westermann, R., Gerjets, P. (1994). Induktion. In Th. Herrmann u. W. Tack (Hg.), *Enzyklopädie der Psychologie. Methodologische Grundlagen der Psychologie* (S. 428–471). Göttingen: Hogrefe.

Wettersten, J.R. (1992). *The roots of critical rationalism.* Amsterdam: Rodopi.

Worrall, J. (1989). Why both Popper and Watkins fail to solve the problem of induction. In F. d'Agostino u. I.C. Jarvie (Hg.), *Freedom and rationality: Essays in honor of John Watkins* (S. 257–296). Dordrecht: Kluwer.

Arndt Bröder

Interne und externe Validität aus deduktivistischer Sicht: Versuch einer Systematisierung und Bestandsaufnahme

Zusammenfassung

Anhand zweier rezenter Debatten zwischen Verfechtern »naturalistischer« Forschung und solchen, die die rigorosen Kontrollmöglichkeiten des »Labors« bevorzugen, wird die mangelnde Begriffsklärung bezüglich der so genannten *externen Validität* aufgezeigt. Die Begriffe der internen und externen Validität werden aufgrund der sie laut Gadenne (1976) bedrohenden Störfaktoren klassifiziert. Die aus einer deduktivistischen Perspektive hergeleiteten Gütekriterien der Populationsvalidität, Variablenvalidität und ökologischen Validität (Bredenkamp, 1980) werden dazu in Beziehung gesetzt. Es wird gezeigt, dass der externen Validität auch in der deduktivistischen Methodologie eine überragende Bedeutung zukommt, wenngleich die vorgeschlagenen Maßnahmen zu ihrer Erhöhung von den üblicherweise vorgeschlagenen abweichen. Einem großen Teil psychologischer Laborforschung kann tatsächlich mangelnde externe Validität vorgeworfen werden, aber aus anderen als den üblicherweise vorgetragenen Gründen.

Einleitung

Seit ihren Anfängen sieht sich die experimentell orientierte Psychologie in regelmäßigen Abständen mit dem Vorwurf der Lebensferne konfrontiert und man fordert, die psychologischen Labors zu verlassen und »realitätsnähere« Forschung zu betreiben. Zu einseitig habe die Psychologie die Sicherung der »internen Validität« von Untersuchungen betont und dabei die »externe Validität« vernachlässigt. Die beiden Validitätskriterien haben Einzug in fast alle Lehrbücher der Versuchsplanung gehalten und stehen laut Campbell und Stanley (1963) – den Begründern dieser Begrifflichkeit – in einem Spannungsverhältnis zueinander, wonach jede dieser Validitätsarten jeweils nur auf Kosten der anderen zu erhöhen sei. Prinz und Müsseler (2002) kommentieren, es sei »wohlfeil, das Vorgehen der Experimentellen Psychologie für seine Künstlichkeit zu kritisieren und über die große Distanz zwischen Labor und Leben die Nase zu rümpfen« (S. 7).

Die Forderung nach mehr externer Validität erscheint jedoch plausibel und attraktiv (daher wohlfeil), und die Debatten zwischen Verfechtern einer naturalistischen Forschung und Befürwortern der Laborforschung werden mit großer Leidenschaft geführt. Karl Popper schreibt: »Kritische Vernunft ist besser als Leidenschaft, besonders in Dingen, die mit Logik zu tun haben« (Popper, 1984, S. 12). Es soll daher ein Versuch gemacht werden, die Argumente nüchtern zu analysieren und zu systematisieren, wobei dies aus einer deduktivistischen Forschungsperspektive geschieht, die Hypothesen testend vorgeht und auf induktive Verallgemeinerungen einzelner Resultate verzichtet (Bredenkamp, 1980; Gadenne, in diesem Band). Popper (1984, S. 12) schreibt jedoch weiter: »Ich gebe aber gerne zu, daß ohne ein gewisses Maß von Leidenschaft nie irgend etwas zustande kommt«, denn sie verweist ja darauf, dass hier ein Problem vorliegt, das für wichtig gehalten wird.

In ihrer ursprünglichen Definition der externen Validität schreiben Campbell und Stanley (1963, S. 175): »*Externe Validität* stellt die Frage der *Generalisierbarkeit:* Auf welche Populationen, Umstände, Behandlungsvariablen und Messvariablen kann ein Effekt generalisiert werden?« (Übers. vom Verf.). Antworten auf diese Frage sind nach Popper (1966; 1984) prinzipiell logisch nicht zu rechtfertigen und in einer nach der oben genannten Charakterisierung des Begriffs »deduktivistisch« zu nennenden Methodologie offenbar auch gar nicht angestrebt. Insofern ist externe Validität für deduktivistisch orientierte Forscher auf den ersten Blick anscheinend nicht wichtig. Dennoch finden wir bei Bredenkamp (1980) folgende, zunächst überraschende Aussage: »Externe Validität, deduktivistisch interpretiert, ist genauso wichtig wie interne Validität. Verfahren zur Ermöglichung der externen Validität sind genauso wie die zur Herstellung der internen Validität auf die Vermeidung von Scheinfalsifikationen und -bestätigungen gerichtet« (S. 13). Da die Theorie der internen und externen Validität von Gadenne (1976) bezüglich *beider* Validitätskriterien als »induktivistisch« gekennzeichnet wurde, verzichtet Bredenkamp (1979, 1980) auf die weitere Verwendung dieser Begriffe und beschreibt eine Reihe von Validitätskriterien, die zu unterschiedlichen Anteilen der Sicherung interner und externer Validität (deduktivistisch interpretiert) dienen sollen. Da die Begriffe »interne und externe Validität« ein fester Bestandteil der Methodenliteratur sind (Bredenkamp, 2001; Brewer, 2000; Smith, 2000) und deduktivistisch motivierte Bezeichnungen verschiedener Validitätskriterien sich offenbar nicht durchsetzen (s. aber Westermann, 2000), erscheint mir der Versuch sinnvoll, diese Begriffe zueinander in Beziehung zu setzen, denn »confusion is the likely result of lack of clarity

about separate validity types and about their interrelations« (Vissers et al., 2001, S. 131), was vermutlich wesentlich dazu beiträgt, dass beide Validitätsarten immer wieder gegeneinander ausgespielt werden (z. B. Banaji u. Crowder, 1991; Mook, 1983).

Zunächst werden exemplarisch kurz zwei Debatten um die »externe Validität« psychologischer Forschung sowie die immer wieder vorgetragenen Argumente referiert, um einerseits die Aktualität und andererseits die mangelnde theoretische Analyse in diesem Bereich zu dokumentieren. Dann werden die Begriffe der internen und externen Validität von Untersuchungen aufgrund einer Klassifikation von Störbedingungen, wie sie Gadenne (1976) vorgenommen hat, aus deduktivistischer Perspektive reinterpretiert.[1] Die von Bredenkamp (1979, 1980) vorgeschlagenen Validitätskriterien werden daraufhin diesen Definitionen – so weit möglich – zugeordnet und Maßnahmen zur Erhöhung der Validitätsarten kurz angesprochen, wobei eine ausführliche Behandlung in diesem Rahmen nicht möglich ist. Besonders bezüglich Kriterien und Maßnahmen der *internen Validität* scheint es zwischen deduktivistisch und induktivistisch motivierten Forschern keinen Dissens zu geben, wenn man von unterschiedlichen Interpretationen der erhaltenen Ergebnisse absieht.

Bestandsaufnahme: Debatten über externe Validität

Im Folgenden sollen kurz zwei Debatten über die Künstlichkeit von Laborforschung und die Forderung nach mehr externer Validität geschildert werden. Es handelt sich dabei um die Anfang der neunziger Jahre ausgetragene Kontroverse zur Alltagsgedächtnisforschung (»Everyday memory«) und eine rezentere Diskussion zum so genannten »Naturalistic Decision Making (NDM)«. Es sei vorher angemerkt, dass die in solchen Kontroversen anzutreffende begriffliche Dichotomie »Labor« versus »Feld« hier lediglich der einfacheren Darstellung wegen übernommen wird, wenngleich es weder eine scharfe Grenze noch sinnvolle allgemeine Abgrenzungskriterien zwischen beiden Settings gibt (s. Vissers et al., 2001). Was kennzeichnet ein psychologisches »Labor«? Ist ein Gedächtnisexperiment mit künstlichen Stimuli, bei dem die Daten mittels eines Laptops bei den Versuchspersonen

1 Von der »Validität einer Untersuchung« zu sprechen ist ein vereinfachender Sprachgebrauch, denn Validität ist ein Kennzeichen der aus der Untersuchung gezogenen Schlussfolgerungen, nicht der Untersuchung selbst.

zu Hause erhoben werden, ein Labor- oder ein Feldexperiment? Ist eine gefilmte Gruppendiskussion zur Analyse von Interaktionsprozessen allein deshalb als »Laborstudie« zu bezeichnen, weil sie in den Räumen eines Psychologischen Instituts stattfindet? Die begriffliche Unklarheit stellt ein zusätzliches Problem in der Bewertung der Argumente dar.

Ende der siebziger Jahre zeigten sich eine Reihe von Forschern im Bereich der Gedächtnisforschung unzufrieden mit der zu beobachtenden Ferne der experimentellen Forschung zu alltäglichen Gedächtnisphänomenen und -leistungen und begannen, Gedächtnisleistungen in »realistischeren« Settings und bezüglich »realistischerer« Stimuli zu untersuchen. Die durch dieses Kriterium locker verbundene Forschungsrichtung wird durch die Begriffe *autobiographical memory, ecological memory, practical memory* oder *everyday memory* umschrieben (Gruneberg et al., 1988). Die heftige Debatte begann in diesem Fall mit einer vernichtenden Kritik dieser Richtung durch Banaji und Crowder (1989), die ihr vorwarfen, dass trotz »passionate rhetoric claiming that important questions about memory could be tackled only if researchers looked to the ›real world‹ for hypothesis validation« (S. 1185) weder neue Methoden noch neue Prinzipien oder gar Theorien aus dieser Richtung erwachsen seien. Banaji und Crowder erklärten daher den »Bankrott« dieser Richtung und verteidigten die Laborforschung, die in ihren Augen zu generalisierbareren Ergebnissen führe als die alltagsnahe Forschung. Selbstverständlich rief dieses Urteil heftige Gegenwehr hervor. Bahrick (1991), Conway (1991) und Loftus (1991) argumentierten, dass manche Gedächtnisphänomene ohne das Verlassen des Labors unentdeckt geblieben wären, zum Beispiel die von Bahrick et al. (1975) untersuchten Vergessenskurven über mehrere Jahrzehnte. Gruneberg et al. (1991) verwiesen vor allem auf die Zielsetzung des »psychological engineering«, die durch eine naturalistische Forschung eher erfüllt werden könne. Versucht man zu extrahieren, was in dieser Debatte mit dem Begriff der externen oder ökologischen Validität gemeint ist, so wird darunter entweder *irgendein* lebensnaher Aspekt einer Studie verstanden (Banaji u. Crowder, 1989), generell das Verlassen des »Labors« (Bahrick, 1991; Ceci u. Bronfenbrenner, 1991; Conway, 1991; Klatzky, 1991; Loftus, 1991), die praktische Relevanz (Gruneberg et al., 1991) oder einfach die »Ähnlichkeit« zum realen Leben (Aanstos, 1991). Bezogen auf die von Gruneberg et al. (1991) angemahnte Nützlichkeit der Forschung verwiesen Banaji und Crowder (1989) auf die »essential difference between science and technology« (S. 1191). Die Unterscheidung dieser Zielsetzungen von Forschung ist zwar zentral, in gewissem Sinne erweisen Banaji und Crowder der Forschung jedoch einen Bärendienst, da ihre unkritische Gleichsetzung aller alltagsnahen Methoden und

»externer« bzw. »ökologischer« Validität mit technologischer Forschung suggerieren, externe Validität sei für nicht-technologische Forschung irrelevant. In einem etwas versöhnlicheren Abschlusskommentar bemerken Banaji und Crowder (1991), der einzige Streitpunkt zwischen interner und externer Validität »arises when one dimension must be sacrificed in favour of the other« (S. 78).

Hat man diese im *American Psychologist* abgedruckte Debatte gelesen, so bleibt das ungute Gefühl, dass viele lange bekannte Argumente ausgetauscht wurden, ohne dass es zu einer Verständigung im Sinne der Modifikation eigener Standpunkte über angemessene Methodologie gekommen wäre, die über den versöhnlichen Appell hinausginge, beide Ansätze als »komplementär« zu begreifen (Klatzky, 1991). Folgende Ergebnisse der Debatte kann man konstatieren: (1.) Der Begriff der externen Validität bleibt undefiniert, verschiedene Bedeutungen werden munter durcheinander verwendet. (2.) Alle Beiträge suggerieren unhinterfragt, dass nach beliebigen Aspekten »naturalistische« Forschung eine höhere externe Validität aufweise als »künstliche«. Das gilt auch für Banaji und Crowder (1989), die jedoch den Wert externer Validität generell infrage stellen. (3.) Die Auffassung, interne und externe Validität stünden in einem Spannungsverhältnis, wird zementiert. (4.) Es wird nicht zwischen einem »Entdeckungszusammenhang« und einem »Begründungszusammenhang« von Forschung unterschieden, was für eine Diskussion über die Notwendigkeit der Kontrolle von Störvariablen von großem Vorteil wäre (Ausnahme: Morton, 1991).

Etwa zehn Jahre nach der soeben berichteten Kontroverse entbrannte eine ähnliche, aber deutlich gemäßigtere Debatte um die Forschung zum so genannten »Naturalistic Decision Making (NDM)«, die von Lipshitz et al. (2001) definiert wird als »attempt to understand how people make decisions in real-world contexts that are meaningful and familiar to them« (S. 332). Wenngleich in dieser programmatischen Übersicht zur Naturalistic Decision Making einige begriffliche Verwirrung bzgl. deskriptiver, normativer und präskriptiver Modelle herrscht und Naturalistic Decision Making verschiedentlich als »Modell«, als »Methode«, als »Richtung« oder »Paradigma« bezeichnet wird, so wird in ihrem Bericht dennoch deutlich, dass sie erstens die *Hypothesen generierende* Natur des Ansatzes betonen und zweitens den Fokus dieser Richtung in der *Anwendung* sehen. Zum ersten Punkt betonen sie, dass die im Feld gewonnen Hypothesen anschließend idealerweise durch streng kontrollierte Laborexperimente zu prüfen sind. Gegen dieses Vorgehen ist aus deduktivistischer Perspektive in frühen Stadien der Erforschung eines Gegenstandsbereichs ohne ausgereifte Theoriebildung nichts einzuwenden. Bei der zweiten Zielsetzung, die als »tech-

nologische Forschung« bezeichnet werden kann, ist allerdings aus deduktivistischer Sicht anzumerken, dass eine – wie auch immer definierte – »Realitätsnähe« der Untersuchungsbedingungen den Erfolg nicht garantiert. Vielmehr ist aus dieser Sicht auch technologische Forschung durch die Anwendung streng getesteter und bewährter *Theorien* gekennzeichnet, aus denen Anwendungsprinzipien (so genannte technologische Prognosen) hergeleitet und auf ihre Nützlichkeit überprüft werden können (s. Bredenkamp, 1980; Gadenne, 1976). Cooksey (2001) kritisiert in der genannten Debatte, dass der »empty term ›real world‹« (S. 361) undefiniert bleibe, und Jungermann (2001) bezweifelt, dass die in typischen NDM-Studien untersuchten Gegenstandsbereiche tatsächlich repräsentativ für menschliches Entscheiden seien. Beiden Kritikpunkten ist gemeinsam, dass sie es als hilfreicher ansehen, die als relevant vermuteten Variablen zu benennen, die das Entscheiden beeinflussen anstatt einfach auf »Realitätsnähe« zu rekurrieren. Roelofsma (2001) fasst die Argumente in einer der unten zu gebenden Definition von Variablenvalidität sehr nahe kommenden Kennzeichnung zusammen: »In order to study the real world, the laboratory need not be at all similar to the real world. When NDM tries to understand the underlying decision processes . . ., the major concern of the researcher becomes to provide the same decision processes to occur« (S. 378).

Im Unterschied zur Kontroverse über das Alltagsgedächtnis wird hier nicht unkritisch die Labor-Feld-Dichotomie benutzt, sondern kritisch in Frage gestellt. In ihrer klaren Zielsetzung von Naturalistic Decision Making als explorative Forschung wird ihr Stellenwert sehr viel deutlicher, und der scheinbare Gegensatz zur Laborforschung verschwindet. Allerdings bleibt festzuhalten, dass auch hier (1.) der Begriff der externen Validität undefiniert bleibt und (2.) Naturalistic Decision Making offenbar davon ausgeht, dass »Realitätsnähe« ein notwendiges Kennzeichen technologischer Forschung sei. Zu ganz ähnlichen Ergebnissen kommen Vissers et al. (2001) in ihrem Versuch der Begriffsklärung.

Interne und externe Validität aus deduktivistischer Perspektive

Wissenschaft verfolgt unterschiedliche Zielsetzungen: einerseits die Sammlung und Systematisierung von Tatsachenwissen sowie andererseits die Aufstellung und Prüfung von Theorien zur *Erklärung* von Sachverhalten (Brewer, 2000; Huber, 1995). In beiden Fällen unterscheidet die Strenge der Me-

thode die Wissenschaft von Alltagsbeobachtungen. Im ersten Fall (z. B. Kartierung von Feuchtbiotopen, Vermessung des Ozonlochs, Meinungsumfragen etc.) zielt die Methodenstrenge auf eine möglichst objektive Übersetzung der vorgefundenen Sachverhalte in eine Beobachtungssprache. Im zweiten Fall erfüllen die empirischen Untersuchungen die darüber hinaus gehende Funktion, Prüfinstanzen für allgemeine Aussagen zu sein. Die hier zu diskutierenden Kriterien der internen und externen Validität beziehen sich auf Untersuchungen mit der zweiten genannten Zielsetzung.

Die zu prüfenden Hypothesen in der Psychologie lassen sich zum großen Teil als universelle Allsätze der Form »Wenn *A*, dann *B*« ausdrücken oder in der quantitativen Variante »Je mehr *A*, desto mehr *B*«. »A« und »B« nehmen hier auf theoretische, nicht direkt beobachtbare Variable Bezug, über deren Zusammenhang (meist kausal verstanden) die Hypothese etwas aussagt. »Universeller Allsatz« bedeutet hier, dass der behauptete Zusammenhang für eine nach bestimmten Kriterien definierte, aber im Prinzip offene Population gelten soll, etwa für »alle Kinder auf der sensomotorischen Entwicklungsstufe«. Diese Sätze sind deshalb von besonderem Interesse, da nur diese Form sie potenziell als allgemeine Gesetzesaussagen qualifiziert, die in wissenschaftlichen Erklärungen oder Prognosen als Argumente zugelassen sind (s. Gadenne, 1994; Westmeyer, 1973).

In einer als Prüfversuch einer Hypothese angelegten Untersuchung sind sowohl mit der Hypothese kompatible als auch inkompatible Ergebnisse möglich, die als Feststellung *bewährender* oder *falsifizierender* Basissätze bezeichnet werden. Unter der *Strenge* einer Untersuchung *U* versteht man die Wahrscheinlichkeit, dass eine falsche Hypothese in *U* zu einem Falsifikationsurteil führt, unter ihrer *Fairness* dagegen die Wahrscheinlichkeit, dass eine wahre Hypothese sich in *U* bewährt (s. Erdfelder u. Bredenkamp, 1994; Westermann, 2000). Beide Wahrscheinlichkeiten sind nicht quantifizierbar, machen aber deutlich, dass ein Bewährungsurteil bezüglich der Hypothese umso informativer ist, je strenger die Untersuchung war, während ein Falsifikationsurteil umso informativer ist, je »wohlwollender« oder fairer die Prüfung war. Die Validität einer Untersuchung ist demnach umso höher, je geringer beide Fehlentscheidungswahrscheinlichkeiten sind, wobei der Verringerung einer Fehlentscheidungswahrscheinlichkeit ohne Erhöhung der anderen formale Grenzen gesetzt sind (s. dazu Erdfelder u. Bredenkamp, 1994).[2]

2 Von einer psychologischen Untersuchung kann laut Gadenne (1976) niemals behauptet werden, dass alle validitätsbedrohenden Störfaktoren ausgeschlossen wurden, daher gibt es keine »validen« versus »invaliden« Untersuchungen. Angemessener scheint ein komparativer Validitätsbegriff, wie er von Erdfelder und Bredenkamp (1994) formalisiert wurde.

Alle Maßnahmen zur Erhöhung der Validität von Untersuchungen müssen darauf abzielen, die Gefahr von Fehlurteilen bezüglich der Hypothese zu vermeiden. Das betrifft sowohl die interne als auch die externe Validität. Im Folgenden sollen diese Begriffe – Gadenne (1976) folgend – bezüglich der unterschiedlichen wirksamen Störfaktoren voneinander abgegrenzt werden. Danach werden die von Bredenkamp (1980) erläuterten Validitätsbegriffe dazu in Beziehung gesetzt. Diese Betrachtung wird zeigen, dass es kein Spannungsverhältnis zwischen interner und externer Validität gibt.

Interne Validität

Soll in einer Untersuchung U die kausale Rolle der Variablen A für die Variable B geprüft werden, so werden konkrete Realisierungen der theoretischen Variablen A_U und B_U beobachtet, die über Hilfshypothesen (Messmodelle, empirisch begründete Zusammenhänge, Plausibilitätsüberlegungen etc.) mit den theoretischen Größen zusammenhängen. Ist B_U diejenige Realisierung von B, die laut Hypothese mit A_U verbunden ist und konstatiert man die Beobachtung »$A_U \wedge B_U$«, so kommt dies einem Bewährungsurteil gleich, während die Beobachtung »$A_U \wedge \neg B_U$« als Falsifikationsurteil gewertet wird. In der Psychologie ist die Feststellung solcher Beobachtungen am Einzelfall oft durch Messungenauigkeiten, Reaktivität der Messung und Ähnliches nicht möglich. Die Prüfung erfolgt dann ersatzweise über statistische Hypothesen. Dieses Vorgehen ist dann gerechtfertigt, wenn die statistische Hypothese (bei Gültigkeit der Hilfshypothesen in U) eine Implikation der zu Grunde liegenden psychologischen Hypothese ist (Bredenkamp, 1986). Für eine tiefere Abhandlung der damit verbundenen komplexen Entscheidungsregeln, konventionalistischen Elemente, Schlussfolgerungen bezüglich der zugrunde liegenden Theorie und statistischen Spezialprobleme verweise ich auf die entsprechende Literatur (Erdfelder u. Bredenkamp, 1994; Lakatos, 1974; Westermann, 2000). Unbezweifelbar hängt die Validität der Untersuchung einerseits von der Gültigkeit der Hilfshypothesen ab. Eine Störung der internen Validität liegt darüber hinaus vor, wenn *unabhängig von A wirksame Faktoren* das Auftreten von B und/oder B_U bewirkt oder verhindert haben, wodurch die Prüfung der Hypothese entweder nicht streng (Bewährungsurteil) oder nicht fair (Falsifikationsurteil) war. Diese möglichen Faktoren nennt Gadenne (1976) »Störbedingungen 1. Art«. Die Existenz derartiger Störbedingungen außerhalb von U ist für den Wahrheitswert der zu prüfenden Hypothese selbst im Prinzip irrelevant, da eine Hypothese über die kausale Relevanz von A für

B andere mögliche kausale Faktoren (meist) nicht ausschließt. Für *U* als Prüfinstanz der Hypothese reduzieren sie aber deren Informationsgehalt bezüglich der Hypothese.

»Ein ganz wesentliches Ziel der Maßnahmen der Versuchsplanung besteht darin, die gemeinsame systematische Variation von möglichen Störfaktoren mit der hypothesenrelevanten UV zu verhindern und die statistische Assoziation zwischen den potentiellen Störfaktoren und der UV auf den Wert Null zu bringen« (Hager, 1987, S. 92). Über die dazu notwendigen Maßnahmen (z. B. Konstanthaltung, Parallelisierung, Ausbalancieren und vor allem Randomisierung) herrscht zwischen Deduktivisten und Induktivisten kein Dissens, wenngleich das Ergebnis einer nach diesen Kriterien intern validen Untersuchung von jenen etwas bescheidener interpretiert werden dürfte.

Externe Validität

Anders als in der Frage nach interner Validität besteht in den Ansichten zur externen Validität und ihrer Erreichung offenbar kein Konsens zwischen den Forschern, nicht einmal die *Definition* ist einheitlich. In Campbell und Stanleys (1963) Fassung ist es die Frage nach der Generalisierbarkeit von Ergebnissen und daher auch nach ihrer eigenen Auffassung mit dem Induktionsproblem behaftet (S. 175). Aus deduktivistischer Perspektive kann die induktive Verallgemeinerung einzelner Resultate kein Ziel sein, aber was ist dann externe Validität?

Gadenne (1976) sieht die externe Validität dann bedroht, wenn so genannte Störbedingungen 2. Art wirksam sind: »Ein Ereignis F' ist eine störende Bedingung 2. Art, wenn A nur zusammen mit F' hinreichend für das Auftreten von B ist« (S. 16). Diese Definition ist allerdings unvollständig, da sie nur Scheinbestätigungen umfasst. Eine störende Bedingung *F'* liegt ebenfalls vor, wenn das Auftreten von *B* in Gegenwart von *A* in der Kombination von *A* und *F' verhindert* wird. Störvariable 2. Art sind also solche, die mit der unabhängigen Variablen *interagieren*, was schon Campbell und Stanley (1963) so formulierten. Anders als bei Störbedingungen 1. Art ist die Existenz von Störbedingungen 2. Art (sowohl in *U* als auch außerhalb von *U*) von großer theoretischer Bedeutung, da sie immer eine Einschränkung des Geltungsbereichs der zu prüfenden Hypothese implizieren. Streng genommen ist ein durch diese Störbedingungen hervorgerufenes Falsifikationsurteil keine scheinbare, sondern eine echte falsifizierende Prüfinstanz, da die Allgemeingültigkeit der Implikation »Wenn *A*, dann *B*«

eingeschränkt wird, denn offenbar ist die Hypothese nicht gültig, wenn *A und F'* vorliegen. Wegen ihres gegenüber Störbedingungen 1. Art anderen theoretischen Status werden diese Störfaktoren auch »Moderatorvariable« genannt (Westermann, 2000).

Wenn generell dem Laborexperiment mangelnde externe Validität vorgeworfen wird, dann wird nach dieser Kennzeichnung entweder angenommen, dass wesentliche Variable in der Laborsituation *nicht* vorhanden sind, sodass es zu Scheinfalsifikationen kommt, oder es wird vermutet, dass *B* durch *A* nur unter den spezifischen Laborbedingungen bewirkt wird, wodurch es zu Scheinbestätigungen kommt. Die Kritik dürfte sich im ersten Fall hauptsächlich auf die verwendeten Hilfshypothesen bezüglich der unabhängigen Variablen (z. B. »Induziert der Film *wirklich* die Emotion ›Freude‹?«) oder der abhängigen Variablen beziehen (z. B. »Misst die Menge der vergebenen Schocks *wirklich* Aggression?«) sowie auf Fragen der Reaktivität von Messungen, Hypothesenbildung der Versuchspersonen und so weiter. Dies sind nach Bredenkamp (1979; 1980) Aspekte der Variablenvalidität oder ökologischen Validität. Der zweite Fall ist von mindestens ebenso großer Bedeutung und scheint die wesentliche Befürchtung von Kritikern der »Künstlichkeit« von Laborforschung zu sein: Die in der speziellen Laborsituation replizierbaren Effekte oder Bestätigungen von Hypothesen gelten nur für diese Bedingungen, nicht aber für andere. Meines Erachtens ist es sehr wichtig, in diesem Zusammenhang eine *triviale* Variante dieser Kritik von einer *ernst zu nehmenden* Variante zu unterscheiden, was in den üblichen Debatten um die externe Validität nicht geschieht und daher viel Verwirrung stiftet.

Die triviale Variante beklagt, dass Laboreffekte sich »in der Welt da draußen« – wie immer sie definiert sein mag – nicht replizieren ließen und daher für diese keine Bedeutung hätten. Diese Kritik geht offenbar davon aus, dass nur solche kausalen Variablen der Betrachtung wert seien, deren Einfluss sich auch in einem beliebigen Umfeld unsystematisch wirksamer Störfaktoren 1. Art nachweisen lässt. Diese Kritik verwechselt ein vordergründiges Kriterium der Nützlichkeit mit dem eigentlichen Ziel der Hypothesenprüfung, nämlich dem Schaffen einer soliden empirischen Basis für kausale Erklärungen. Jedes wissenschaftliche Kausalgesetz bezieht sich »auf vollständig kontrollierbare und gegenüber den möglichen Einflussfaktoren abgeschlossene Situationen ... also auf ideale Systeme, die in der natürlichen und sozialen Umwelt gewöhnlich nicht vorkommen« (Westermann, 2000, S. 157). Das gilt auch in besonderem Maß für die »harte« Wissenschaft Physik, worauf schon Ebbinghaus (1885) verwies: »Der freie Fall, die reibungslosen Maschinen u.s.w., mit denen sich die Physik beschäftigt, sind auch im Vergleich mit dem,

was im wirklichen Geschehen der Natur vorkommt und für uns wichtig ist, nur Abstraktionen. Ein annäherndes Verständnis des Komplicierten und Wirklichen ermöglichen wir uns fast nirgendwo direkt, sondern auf Umwegen, durch successive Zusammensetzung von Erfahrungen, deren jede an künstlich zurechtgemachten und von der Natur selbst in dieser Weise selten oder nie erzeugten Fällen gewonnen wurde« (S. 36). Es wäre absurd zu behaupten, das Newtonsche Fallgesetz sei irrelevant oder ein Laborartefakt, nur weil sein kausaler Einfluss auf einen Schwarm umherfliegender Blätter in einem Herbststurm nicht gerade leicht zu erkennen ist. Man kann sich zwar *wünschen*, dass die Psychologie kausale Faktoren solcher Durchschlagkraft identifiziert, dass sie sich auch in der realen Welt gegen die meisten Störbedingungen durchsetzen, aber wenn wir sie nicht finden, ist dies mit Sicherheit nicht dem Labor anzulasten. Im Gegenteil: Wenn die in der Psychologie wirksamen kausalen Faktoren nur kleine Effekte haben, dann können sie *nur* unter Bedingungen rigider Kontrolle und Ausschaltung der Störfaktoren *überhaupt* identifiziert werden. Die (durchaus mögliche) Alternative besteht dann lediglich darin, die Psychologie unter Nützlichkeitsgesichtspunkten als rein technologische Disziplin zu betreiben und den Erklärungsanspruch aufzugeben (s. Herrmann, 1979).

Die ernster zu nehmende Kritik gegen die Forschung im Labor argumentiert, dass die dort identifizierten Effekte nur durch die Kombination der spezifischen Situation oder Methoden mit den Antezedenzbedingungen hervorgerufen werden, nach Gadennes (1976) Terminologie also Störbedingungen 2. Art vorliegen. Hier geht es nicht um die Frage, ob die kausalen Faktoren stark genug sind, um sich durchzusetzen, sondern um Zweifel daran, dass sie außerhalb der spezifischen Bedingungen überhaupt *wirksam* sind. Diese Zweifel mögen theoretisch begründet sein oder einem generellen Vorurteil gegenüber Labormethoden entspringen, ihre Bewertung kann jedenfalls nur mit empirischen Mitteln versucht werden, und aus einer deduktivistischen Perspektive ist dies auch unbedingt notwendig. Die Kritik mangelnder externer Validität eines großen Teils der psychologischen Laborforschung ist durchaus berechtigt, sofern diese Forschung »paradigmenzentriert« betrieben wird. Häufig werden zur Untersuchung von Effekten die Durchführungsmodalitäten, Stimuli, Versuchspersonen verwendet, mit denen sich Effekte erfahrungsgemäß »am zuverlässigsten« replizieren lassen. Dazu greift man gerne auch auf das »bewährte« Material aus anderen Labors zurück. Dies birgt natürlich die Gefahr großer Serien von Scheinbestätigungen, die zu einem völlig falschen Bild über die geprüfte Hypothese führen. Eine nicht zu unterschätzende Schuld an dieser Paradigmenzentriertheit hat sicherlich die Herausgeberpolitik vieler Fachzeit-

schriften, die in der Tradition des statistischen Hypothesentestens nach R.A. Fisher nur »signifikante« Ergebnisse für informativ und publikationswürdig hält, selbst wenn mit hoher Teststärke die Nullhypothese beibehalten wurde (Bredenkamp, 1972). Forscher, die auf Publikationen angewiesen sind, können es sich daher im Allgemeinen nicht leisten, Ressourcen und Energie in »unsichere« Verfahren zu investieren. Diese verzerrende Praxis *verhindert* aber die strenge Prüfung von Theorien und somit die externe Validität von Forschungsprogrammen.

Die Forderung nach höherer Validität umfasst natürlich auch aus deduktivistischer Perspektive die größere Strenge der Hypothesenprüfung, die durch konzeptuelle Replikationen erreicht werden kann, in denen der kausale Zusammenhang zwischen theoretischen Variablen für viele ihrer konkreten Operationalisierungen sowie unter variierenden Umständen zu prüfen ist (Bredenkamp, 1980). Pointiert beschreibt Abelson (1995) diese Strategie: »You can't see the dust if you don't move the couch« (zit. in Smith, 2000, S. 26). Was ist nun in diesem Zusammenhang von der Forderung zu halten, das Labor zu verlassen, um die externe Validität zu erhöhen? Dies kann *eine* Maßnahme sein, um entsprechende Moderatorvariable zu identifizieren, wobei jedoch zu bedenken ist, dass dies im Allgemeinen eine geringere Kontrollierbarkeit anderer Störfaktoren mit sich bringt und somit eine weniger strenge (oder faire) Hypothesenprüfung impliziert. Im Gegensatz zur landläufigen Auffassung externer Validität ist das *umgekehrte* Vorgehen angebracht: Im »Feld« replizierbare Effekte, die theoretische Zusammenhänge zwischen Variablen suggerieren, müssen sich *auch im Labor* bewähren, da dieses eine bessere Kontrolle von Störvariablen und damit eine strengere Prüfung der Hypothese ermöglicht.[3] Unabhängig davon ist aber der grobe Vergleich zwischen dem »Labor« und dem »Feld« vermutlich nicht zielführend, da eine Fülle von Unterschieden zwischen beiden besteht und eventuell abweichende Resultate in beiden Situationen mysteriös bleiben, sodass der Informationsgewinn für die Hypothese minimal ist. Deutlich informativer ist es, vermutete Störbedingungen 2. Art (je nach Fragestellung z. B. Alter oder Bildungsstand der Probanden, An- versus Abwesenheit eines weiblichen Versuchsleiters, Tageszeit oder Ort der Unter-

3 Diese Aussage kann natürlich nur für Phänomene gelten, die im Prinzip (und pragmatisch) im Labor untersuchbar sind. Die von Bahrick et al. (1975) untersuchten Vergessenskurven über mehrere Jahrzehnte gehören vermutlich nicht dazu. Andererseits besteht aber gerade eine Herausforderung der Psychologie darin, neue Methoden zu finden, um psychologische Hypothesen einer experimentellen Prüfung zugänglich zu machen, die zunächst als ununtersuchbar galten, zum Beispiel die Hypothese »Freudscher Versprecher« (s. Bröder u. Bredenkamp, 1996; Schüttauf et al., 1997; Ufer et al., 2000).

suchung etc.) *systematisch* zu variieren und als Kontrollfaktoren in die Untersuchung aufzunehmen. Dies ermöglicht einerseits die Entdeckung disordinaler Interaktionen mit der unabhängigen Variablen (die den Geltungsbereich der Hypothese einschränken) und hat noch den positiven Nebeneffekt, im Allgemeinen die Teststärke des verwendeten statistischen Tests zu erhöhen, da Fehlervarianz gebunden wird, sofern es sich bei dem Kontrollfaktor um einen Störfaktor 1. Art handelt. Dieses Vorgehen verlangt allerdings eine theoretisch oder intuitiv begründete Vorstellung davon, welche Variablen relevant sein könnten.[4]

Zusammenfassend bleibt festzuhalten, dass auch aus deduktivistischer Sicht die externe Validität extrem wichtig ist und im Prinzip kein echter Dissens mit Vertretern einer ernst zu nehmenden Kritik der gängigen Laborforschung über die Natur des Problems besteht. Aus deduktivistischer Sicht ist es jedoch erstens falsch, dieses Problem dem »Labor« an sich vorzuwerfen, und zweitens wenig zielführend, undifferenziert eine größere Ähnlichkeit zum »echten Leben« zu verlangen, da dies – abgesehen von definitorischen Schwierigkeiten – oft eine geringere Strenge der Hypothesenprüfung sowie fehlende Informationen über die vermittelnden Moderatorvariablen zur Folge hat.

Validitätskriterien nach Bredenkamp und ihre Beziehung zu interner und externer Validität

Im Folgenden werden die von Bredenkamp (1980) vorgeschlagenen Validitätskriterien in aller Kürze skizziert und zu den erläuterten Begriffen der internen und externen Validität in Beziehung gesetzt.

Populationsvalidität

Die Forderung nach Populationsvalidität verlangt trivialerweise, dass die Untersuchungseinheiten, an denen die Gültigkeit der zugrunde liegenden Hypothese getestet wird, aus der Population stammen, für die die Hypo-

4 Es spricht selbstverständlich nichts dagegen, auch post hoc den Einfluss in einer Untersuchung »zufällig« erhobener Variablen (Versuchsleiter, Alter der Versuchsperson etc.) auf das Ergebnis zu untersuchen und Datensätze, die zur Hypothesenprüfung erhoben wurden, auch heuristisch zu nutzen (Erdfelder, 1994), um potenzielle Moderatorvariable zu entdecken, die dann später systematisch untersucht werden können.

these Gültigkeit beansprucht. Da dies bei allgemeinen Gesetzesaussagen meist offene Populationen sind, die zeitlichen Veränderungen unterliegen, sind »Zufallsstichproben« oder »repräsentative Stichproben« nicht möglich. Dies ist aber auch gar nicht nötig, da alle zum beanspruchten Geltungsbereich der Hypothese gehörenden Versuchspersonen zur Prüfung der Hypothese geeignet sind. Nun unterscheiden sich Menschen aber glücklicherweise auf den verschiedensten Dimensionen und es könnte sein, dass die Hypothese nur für einen Teil der Personen mit definierten Kriterien zutrifft (z. B. Psychologiestudenten), was einer Störbedingung 2. Art entspricht. Folglich wird die externe Validität von Untersuchungen zur Hypothese erhöht, wenn nicht nur diese Personengruppe, sondern auch andere untersucht werden. Allerdings besteht die Methode der Wahl hier *nicht* darin, einfach eine nach beliebigen Kriterien heterogenere Stichprobe zu untersuchen, da dies eventuell aufgrund unsystematischer Störfaktoren 1. Art lediglich die nicht auf die unabhängige Variable zurückzuführende Kriteriumsvarianz erhöht und somit eventuell die Strenge oder Fairness der Prüfung sogar mindert (je nach implizierter statistischer Hypothese). Sinnvoller ist es, als relevant vermutete Persönlichkeitsunterschiede systematisch als Kontrollfaktor zu variieren, da dies erstens ihre Rolle als potenzielle Störvariable 2. Art erst sichtbar machen kann (i. S. einer Interaktion mit der UV) und zweitens die unsystematische Fehlervarianz gegebenenfalls reduziert.

Variablenvalidität

Die Variablenvalidität umfasst eine Reihe von Aspekten, die als Konstruktvalidität, konzeptuelle Replizierbarkeit, Durchschlagskraft der unabhängigen Variablen und Konfundierungsfreiheit bezeichnet werden können, wenngleich sich die Bedeutungen zum Teil überlappen.

Die Konstruktvalidität meint die Frage, ob die empirischen Realisierungen der theoretischen Variablen diese auch tatsächlich repräsentieren: Misst die abhängige Variable wirklich das betreffende Konstrukt und beeinflusst die experimentelle Manipulation wirklich die interessierende Variable? Es geht also darum, ob die Hilfshypothesen zur Überbrückung zwischen Theorie und Empirie gültig sind. Diese Validitätsart lässt sich nicht auf eine Erörterung der Störbedingungen 1. und 2. Art reduzieren, vielmehr ist der Prozess der Konstruktvalidierung ein komplexer Vorgang, der theoretische Überlegungen und empirische Argumente gleichermaßen umfasst und auf diese zurückwirkt (s. Cronbach u. Meehl, 1955). Nebenbei stellt sich aber auch hier

das Problem der externen Validität, da manche Hilfshypothesen möglicherweise nur unter bestimmten Bedingungen Gültigkeit besitzen. Beispielsweise ist es durchaus denkbar, dass Intelligenztests zwar in normalen Testsituationen mit persönlicher Bedeutung für den Testanden tatsächlich Intelligenz messen, während sie dagegen als Bestandteil eines psychologischen Experiments, in dem die Versuchspersonenvergütung unabhängig von der Testleistung erfolgt, eher Motivation oder Leistungsorientierung messen.

Die konzeptuelle Replizierbarkeit wurde oben bereits erörtert und bezieht sich hier auf die Frage, ob die Hypothese auch für andere Realisierungen der theoretischen Variablen Gültigkeit besitzt. An dieser Stelle sollte noch einmal deutlich werden, dass interne und externe Validität unterschiedlichen Ebenen angehören. Interne Validität ist ein Attribut, das man in unterschiedlichem Ausmaß einzelnen Untersuchungen zuschreiben kann. Externe Validität kann dagegen sinnvoller Weise nur ganzen Forschungsprogrammen und Versuchsserien attestiert werden. Schon deshalb ist es verfehlt, die externe Validität einzelner Untersuchungen durch singuläre Maßnahmen wie »größere Lebensnähe« erhöhen zu wollen.

Die Durchschlagkraft der unabhängigen Variablen kann gegebenenfalls unter den gewählten Bedingungen zu gering sein, da in *U* viele unsystematische Störbedingungen wirksam sind, sodass ein im Prinzip vorhandener Effekt auf die abhängige Variable in diesen Faktoren »untergeht«. Maßnahmen zur Reduzierung dieser unsystematischen Fehler beinhalten beispielsweise die Konstanthaltung oder Elimination derselben, Standardisierung oder Automatisierung des Vorgehens, Versuchsleitertraining sowie den Einbezug varianzbindender Kovariaten als Kontrollfaktoren oder in Kovarianzanalysen (s. dazu Bredenkamp, 1969, S. 349 ff.).

Die Konfundierungsfreiheit betrifft den Aspekt der internen Validität, die immer dann in Gefahr ist, wenn Störfaktoren 1. Art mit der interessierenden unabhängigen Variablen statistisch assoziiert sind.

Ökologische Validität

Die Forderung nach der ökologischen Validität verlangt, dass die Untersuchungsbedingungen die relevanten Merkmale des Geltungsbereichs der psychologischen Hypothese aufweisen. Im Prinzip lässt sich diese Frage unter die oben definierten Aspekte Konstruktvalidität und externe Validität subsumieren, aber eine eigenständige Bezeichnung wird dadurch gerechtfertigt, dass unter diese Rubrik viele *bekannte* Störbedingungen 2. Art fallen, etwa die Sensitivierung von Versuchspersonen durch Vortests und ge-

nerell die Reaktivität der Messung, die Hypothesenbildungen der Versuchspersonen (»demand characteristics« nach Orne, 1962), Versuchsleitereffekte oder das Vernachlässigen langfristiger Effekte. Alle diese Faktoren können zu Scheinfalsifikationen oder -bestätigungen führen.

Die Validität des statistischen Schlusses hängt nicht von Störbedingungen 1. oder 2. Art ab, ist aber ebenfalls ein wichtiger Problemkreis, der bei Bredenkamp (1972, 1980, 1986), Erdfelder und Bredenkamp (1994) sowie Westermann (2000) ausführlich behandelt wird.

In Tabelle 1 sind die Validitätskriterien und ihr Bezug zu den Arten von Störbedingungen noch einmal zusammengefasst. Daraus wird deutlich, dass die meisten Überlegungen sich auf Störbedingungen 2. Art und damit auf die externe Validität von Untersuchungen beziehen. Ebenso wird klar, dass diese Klassifikation ohne die Verwendung der Begriffe »interne« und »externe« Validität den Vorteil birgt, dass es aus deduktivistischer Sicht kein Spannungsverhältnis zwischen der Kontrolle von Störfaktoren 1. und 2. Art gibt.

Tabelle 1: Validitätsarten nach Bredenkamp (1980)

		Störbedingungen		
		1. Art		2. Art
		unsystematisch	systematisch	
Populationsvalidität				X
Variablenvalidität	Konstruktvalidität			(X)
	konzeptuelle Replizierbarkeit			X
	Durchschlagkraft	X		
	Konfundierungsfreiheit		X	
ökologische Validität				X

Schlussfolgerungen

In Hypothesen testenden Untersuchungen ist Sorge zu tragen, Störbedingungen 1. und 2. Art möglichst zu vermeiden, was sich nicht gegenseitig ausschließt. Beides erhöht die Validität von Untersuchungen, oft als interne und externe Validität bezeichnet. Die Wahl »repräsentativer Stichproben« oder »alltagsnaher Untersuchungsbedingungen« ist (abgesehen von der Schwierigkeit, diese zu definieren) kein geeignetes Mittel, um Störfaktoren 2. Art zu kontrollieren oder auszuschalten. Inhomogene Stichproben können im Gegenteil sogar zu einer geringeren Strenge oder Fairness der Untersuchung führen. Wesentlich sind dagegen konzeptuelle Replikationen, die die Identifizierung von Störfaktoren 2. Art auf der Ebene der Variablen-

validität möglich machen, sowie gezielte Variationen der Untersuchungsstichproben bezüglich als Störfaktoren 2. Art vermuteter Variablen, um gegebenenfalls den Allgemeinheitsgrad der Hypothesen einzuschränken. Nur insoweit die »Laborforschung« *diese* Prinzipien nicht beachtet, ist ihr aus deduktivistischer Perspektive mangelnde externe Validität vorzuwerfen.

Literatur

Aanstos, C.M. (1991). Experimental psychology and the challenge of real life. *American Psychologist, 46,* 77–78.

Bahrick, H.P. (1991). A speedy recovery from bankruptcy for ecological memory research. *American Psychologist, 46,* 76–77.

Bahrick, H.P., Bahrick, P.O., Wittlinger, R.P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104,* 54–75.

Banaji, M.R., Crowder, R.G. (1989). The bancruptcy of everyday memory. *American Psychologist, 44,* 1185–1193.

Banaji, M.R., Crowder, R.G. (1991). Some everyday thoughts on ecologically valid methods. *American Psychologist, 46,* 78–79.

Bredenkamp, J. (1969). Experiment und Feldexperiment. In C.F. Graumann (Hg.), *Handbuch der Psychologie. Band: Sozialpsychologie, 1. Halbband: Theorien und Methoden* (S. 332–374). Göttingen: Hogrefe.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1979). Das Problem der externen Validität pädagogisch-psychologischer Untersuchungen. In J. Brandstätter, G. Reinert u. K.A. Schneewind (Hg.), *Pädagogische Psychologie: Probleme und Perspektiven* (S. 267–289). Stuttgart: Klett-Cotta.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1986). Dürfen wir psychologische Hypothesen statistisch prüfen? *Berichte aus dem Psychologischen Institut der Universität Bonn, 12*(2).

Bredenkamp, J. (2001). Laboratory experiment: Methodology. In N.J. Smelser u. P.B. Baltes (Hg.), *International encyclopedia of the social & behavioral sciences* (S. 8226–8232). Oxford: Elsevier/Pergamon.

Brewer, M.B. (2000). Research design and issues of validity. In H.T. Reis u. C.M. Judd (Hg.), *Handbook of research methods in social and personality psychology* (S. 3–16). New York: Cambridge University Press.

Bröder, A., Bredenkamp, J. (1996). SLIP-Technik, Prozeßdissoziationsmodell und multinomiale Modellierung: neue Werkzeuge zum experimentellen Nachweis »Freudscher Versprecher?«. *Zeitschrift für Experimentelle Psychologie, 43,* 175–202.

Campbell, D.T., Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage, *Handbook of research on teaching* (S. 171–245). Chicago, IL: Rand McNally.

Ceci, S.J., Bronfenbrenner, U. (1991). On the demise of everyday memory. *American Psychologist, 46,* 27–31.

Conway, M.A. (1991). In defense of everyday memory. *American Psychologist, 46,* 19–26.

Cooksey, R.W. (2001). Pursuing an integrated decision science: Does ›naturalistic decision making‹ help or hinder? *Journal of Behavioral Decision Making, 14,* 361–362.

Cronbach, L.J., Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Ebbinghaus, H. (1885). *Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie.* Leipzig: Duncker und Humblot. [Nachdruck 1966, Amsterdam: Bonset.]

Erdfelder, E. (1994). Erzeugung und Verwendung empirischer Daten. In T. Herrmann u. W.H. Tack (Hg.), *Enzyklopädie der Psychologie. Band: Methodologische Grundlagen der Psychologie* (S. 47–97). Göttingen: Hogrefe.

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann u. W. Tack (Hg.), *Enzyklopädie der Psychologie. Band: Methodologische Grundlagen der Psychologie* (S. 604–648). Göttingen: Hogrefe.

Gadenne, V. (1976). *Die Gültigkeit psychologischer Untersuchungen.* Stuttgart: Kohlhammer.

Gadenne, V. (1994). Theorien. In T. Hermann u. H. Tack (Hg.), *Enzyklopädie der Psychologie. Band: Forschungsmethoden der Psychologie* (Band 1, S. 295–342). Göttingen: Hogrefe.

Gruneberg, M.M., Morris, P.E., Sykes, R.N. (1988). *Practical aspects of memory: Current research and issues* (Vol. 1 u. 2). New York: Wiley.

Gruneberg, M.M., Morris, P.E., Sykes, R.N. (1991). The obituary on everyday memory and its practical applications is premature. *American Psychologist, 46,* 74–76.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung psychologischer Hypothesen. In G. Lüer (Hg.), *Allgemeine experimentelle Psychologie* (S. 43–264). Stuttgart: G. Fischer.

Hermann, T. (1979). Pädagogische Psychologie als psychologische Technologie. In J. Brandstätter, G. Reinert u. K.A. Schneewind (Hg.), *Pädagogische Psychologie: Probleme und Perspektiven* (S. 209–236). Stuttgart: Klett-Cotta.

Huber, O. (1995). *Das psychologische Experiment. Eine Einführung.* Göttingen: Hogrefe.

Jungermann, H. (2001). One mode of decision making (research) – no less, no more. *Journal of Behavioral Decision Making, 14,* 367–370.

Klatzky, R.L. (1991). Let's be friends. *American Psychologist, 46,* 43–45.

Lakatos, I. (1974). Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme. In I. Lakatos u. A. Musgrave (Hg.), *Kritik und Erkenntnisfortschritt* (S. 89–189). Braunschweig: Vieweg.

Lipshitz, R., Klein, G., Orasanu, J., Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making, 14,* 331–352.

Loftus, E.F. (1991). The glitter of everyday memory ... and the gold. *American Psychologist, 46,* 16–18.

Mook, D.G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Morton, J. (1991). The bankruptcy of everyday thinking. *American Psychologist, 46,* 32–33.

Orne, M.T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776–783.

Popper, K.R. (1966). *Logik der Forschung.* Tübingen: Mohr.

Popper, K.R. (1984). *Objektive Erkenntnis. Ein evolutionärer Entwurf.* Hamburg: Hoffmann und Campe.

Prinz, W., Müsseler, J. (2002). Einleitung: Psychologie als Wissenschaft. In J. Müsseler u. W. Prinz (Hg.), *Allgemeine Psychologie* (S. 1–12). Heidelberg: Spektrum.

Roelofsma, P.H.M.P. (2001). Evaluating ten years of naturalistic decision making: Welcome back in the lab! *Journal of Behavioral Decision Making, 14,* 377–379.

Schüttauf, K., Bredenkamp, J., Specht, E.K. (1997). Induzierte »Freudsche Versprecher« und zwangsneurotischer Konflikt. *Sprache & Kognition, 16,* 3–13.

Smith, E.R. (2000). Research design. In H.T. Reis u. C.M. Judd (Hg.), *Handbook of research methods in social and personality psychology* (S. 17–39). New York: Cambridge University Press.

Ufer, N., Bredenkamp, J., Jacoby, G.E. (2000). Induzierte »Freudsche« Versprecher und Bulimie. *Zeitschrift für klinische Psychologie und Psychotherapie, 29,* 180–186.

Vissers, G., Heyne, G., Peters, V., Guerts, J. (2001). The validity of laboratory research in social and behavioral science. *Quality & Quantity, 35,* 129–145.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur Psychologischen Methodenlehre.* Göttingen: Hogrefe.

Westmeyer, H. (1973). *Kritik der psychologischen Unvernunft.* Stuttgart: Kohlhammer.

Rainer Westermann

Strukturalistische Theorienkonzeption und deduktive Forschungsmethodologie

Zusammenfassung

Die deduktive Methodologie für die psychologische Forschung ist von Bredenkamp auf der Basis des Falsifikationskonzepts von Popper begründet worden. In neueren Darstellungen spielt diese wissenschaftsphilosophische Basis jedoch zu Recht nur noch eine untergeordnete Rolle. Die Poppersche Methodologie konzentriert sich zu sehr auf die Falsifikation von Theorien und die Strenge von Prüfungen, sie verkennt die notwendige Rolle induktiver Argumente im Wissenschaftsprozess und sie analysiert nicht den Aufbau von Theorien und die Genese von Hypothesen. Prinzipien und Konsequenzen deduktiver Methodologien lassen sich jedoch konsistent innerhalb der strukturalistischen Wissenschaftsphilosophie begründen.

Deduktive Methodologien

Grundlagen

Jürgen Bredenkamp (1972) hat in seiner wegweisenden Schrift über den Signifikanztest in der psychologischen Forschung einen Widerspruch betont, der damals wie heute viel zu wenig beachtet wird. Als Ergebnisse psychologischer Forschung werden meist Ergebnisse von Signifikanztests berichtet. Diese prüfen *statistische Hypothesen*, beispielsweise über die Gleichheit von Mittelwerten in verschiedenen Personengruppen. Die interessierenden *psychologischen Hypothesen*, Theorien oder Modelle dagegen beziehen sich zumindest in der Allgemeinen Psychologie auf Zusammenhänge oder Prozesse »innerhalb« jeder einzelnen Person.

Diese Diskrepanz ist unproblematisch, falls sich aus der psychologischen Hypothese oder Theorie T eine statistische Hypothese SH deduzieren lässt, das heißt, falls die Implikation $T \Rightarrow SH$ gilt (Bredenkamp, 1972, S. 150 ff.). In diesem Fall gilt nach dem *Modus tollens* der Aussagenlogik zwangsläufig $\neg SH \Rightarrow \neg T$. Damit kann auf die Falschheit einer psychologischen Theorie

geschlossen werden, wenn eine statistische Hypothese abgelehnt wird, die logisch aus ihr folgt.

Die Annahme einer implikativen Beziehung zwischen der interessierenden Theorie und einer prüfbaren statistischen Hypothese ist der Kern der ursprünglichen *deduktiven Methodologie* der psychologischen Forschung, die von Bredenkamp (1972, 1980) begründet und von anderen Autoren weiter entwickelt worden ist (Gadenne, 1976, 1984; Hager, 1984; Hager u. Westermann, 1983). Sie ist stringent der Methodologie von Karl Popper untergeordnet (Bredenkamp, 1980, S. VI).

Popper (1934/1994) hat seine *deduktive Methodik der Nachprüfung* in der Auseinandersetzung mit dem Logischen Empirismus entwickelt. Nach Carnap (1926, 1986) ist eine empirisch-wissenschaftliche Aussage T nur dann sinnvoll, wenn sie verifizierbar ist, das heißt, wenn sie sich logisch zwingend aus einer Aussage B über unmittelbare Beobachtungen ergibt: $B \Rightarrow T$. Dementsprechend sollen Wissenschaften induktiv vorgehen und aus ihren Daten allmählich Theorien ableiten.

Wie Popper zu Recht betont, sollen wissenschaftliche Hypothesen und Theorien in der Regel für prinzipiell unbegrenzte Mengen von Personen, Objekten, Zeitpunkten oder Situationen gelten. Derartige *unbegrenzte* Allaussagen lassen sich aber nie zwingend aus Aussagen über stets *begrenzte* Mengen von Beobachtungen ableiten. Der Schritt von den konkreten Beobachtungen zu den abstrakteren Hypothesen ist vielmehr stets induktiv und damit unsicher.

Popper ersetzt deshalb das Verifikations- durch ein Falsifikationskriterium. Eine Theorie T ist genau dann falsifizierbar, wenn es einen *Basissatz* B über eine mögliche Beobachtung geben kann, aus der die Falschheit von T logisch folgt: $B \Rightarrow \neg T$. Nach dem *Modus tollens* ist dies äquivalent zu $T \Rightarrow \neg B$. Wissenschaft soll deshalb deduktiv vorgehen: Aus dem theoretischen System werden empirische Folgerungen abgeleitet und kritisch überprüft (Popper, 1994, S. 7 f.). Induktive Elemente sollen in diesem Verfahren nicht auftreten.

Probleme

Bei der Übertragung der logischen Grundüberlegungen auf die tatsächliche Forschung erkannte Popper verschiedene Schwierigkeiten. So sind auch Basissätze keine sicheren Aussagen, sondern haben immer etwas hypothetisches, fallibles, auch wenn sie einen viel geringeren Allgemeinheitsgrad haben als die Hypothesen oder Theorien.

Nachprüfbare Folgerungen sind meist nicht aus der Theorie allein abzuleiten, sondern nur mit *Hilfshypothesen* oder Zusatzannahmen Z (Popper, 1994, S. 51), beispielsweise über die Abwesenheit von Störfaktoren oder die Adäquatheit von Operationalisierungen. Dann gilt nicht mehr $T \Rightarrow \neg B$, sondern $[T \wedge Z] \Rightarrow \neg B$. Tritt vorhersagekonträr B ein, kann über den *Modus tollens* nicht mehr eindeutig auf $\neg T$ geschlossen werden, sondern nur noch darauf, dass T oder Z oder beide falsch sind.

Wenn die abgeleitete Vorhersage eintritt, ist die Theorie oder Hypothese nach Popper (1965, 1994) vorläufig gestützt oder bewährt. Die Bewertung ist jedoch nicht logisch zwingend: Schon bei der Implikation $T \Rightarrow \neg B$ folgt aus der Wahrheit der Konklusion $\neg B$ logisch zwingend weder die Wahrheit noch die Falschheit der Prämisse T.

In verschärfter Form treten diese Probleme auf, wenn aus der Theorie statistische Hypothesen abgeleitet werden. Erstens lassen sich statistische aus wissenschaftlichen Hypothesen nur unter Zusatzannahmen deduzieren, die über die genannten noch hinausgehen: die Unabhängigkeit der Beobachtungen sowie bei parametrischen Tests Normalverteilung und Varianzgleichheit. Zweitens lässt sich nicht definitiv feststellen, ob eine statistische Hypothese wahr oder falsch ist, vielmehr kann man über die Annahme oder Ablehnung von statistischen Hypothesen nur entscheiden (Bredenkamp, 1972; Hays, 1981).

Konventionen und Modifikationen

Um mit diesen Problemen umzugehen, schlägt Popper verschiedene methodologische Konventionen vor. So werden Basissätze »durch Beschluss, durch Konvention anerkannt, sie sind *Festsetzungen*« (Popper, 1994, S. 71). Außerdem können in jeder empirischen Untersuchung Operationalisierungen und andere Zusatzannahmen per Konvention als unproblematisches Hintergrundwissen akzeptiert werden (Gadenne, 1994, S. 312; Popper, 1994, S. 51). Dadurch wird die Falsifizierbarkeit der Hypothese (wieder) hergestellt. Dies ist sinnvoll, wenn das einzige Ziel des Forschungsprozesses darin gesehen wird, Hypothesen zu falsifizieren, um zu empirisch gehaltvolleren und besser bewährten Modifikationen zu gelangen. Man erhöht allerdings auch unweigerlich die Gefahr ungerechtfertigter Falsifikationen. Dass man auch diese fälschlichen Ablehnungen von Hypothesen vermeiden muss, wird vom Standpunkt der strukturalistischen Theorienkonzeption aus deutlich werden.

Häufig werden Zusatzannahmen auch deshalb akzeptiert, weil sie sich in vorangegangenen Untersuchungen gut bewährt haben. Ein derartiger »Schluss« von bisherigen Beobachtungen auf einen ähnlichen neuen Fall gehört allerdings zu den induktiven Argumenten, die aus der deduktiven Methodologie definitiv ausgeschlossen werden sollten. Aus strukturalistischer Sicht kann man allerdings auch mit den unvermeidlichen induktiven Komponenten des Forschungsprozesses gut begründet umgehen.

Die Vorstellung von implikativen Verbindungen zwischen Theorien und empirischen Vorhersagen beziehungsweise statistischen Hypothesen können auch dadurch aufrechterhalten werden, dass man die Zusatzannahmen mit zur Theorie rechnet (Gadenne, 1994, S. 303). Man kann auch jede psychologische Hypothese als eine Konjunktion eines Strukturmodells und eines geeigneten Messmodells auffassen (Erdfelder u. Bredenkamp, 1994, S. 643). In beiden Fällen ist die Ableitung einer Vorhersage oder statistischen Hypothese dann eine logische Implikation und aus der Falschheit der abgeleiteten Vorhersage folgt per Modus tollens zwangsläufig die Falschheit der Theorie. Dabei wird jedoch ein Begriff einer wissenschaftlichen Theorie zugrunde gelegt, der sehr umfassend und wenig ausdifferenziert ist. Strukturalistische Rekonstruktionen zeigen überdies, dass selbst einfache Theorien keine einheitlichen Gebilde sind und dass empirische Untersuchungen meist vielfältige theoretische Bezugspunkte haben. Auch ist es sinnvoll, zwischen der Theorie mit ihren substanziellen Begriffen und Zusammenhängen und den Zusatzannahmen zur adäquaten Operationalisierung oder Störfaktorenkontrolle deutlich zu unterscheiden.

Angesichts der beschriebenen Schwierigkeiten können statt strenger implikativer Beziehungen schwächere Ableitungs- oder Umsetzungsbeziehungen zwischen psychologischen und statistischen Hypothesen unterstellt werden (Hager, 1992; Meehl, 1967). Die Entscheidung über die statistische Hypothese soll die Entscheidung über die psychologische Hypothese nicht determinieren, sondern nur systematisch beeinflussen. Die Ablehnung der abgeleiteten statistischen Hypothese ist zwar in der Regel notwendig, aber keineswegs hinreichend für die Ablehnung der psychologischen Hypothese. Eine bestimmte psychologische Hypothese kann dabei je nach Zusatzannahmen und Randbedingungen zu unterschiedlichen statistischen Hypothesen führen. Die Adäquatheit der Ableitungen wird daran gemessen, wie gut sie geeignet sind, eine *valide* Entscheidung über die psychologische Hypothese herbeizuführen (Hager, 1992).

In neueren Darstellungen deduktiver Methodologien (Erdfelder u. Bredenkamp, 1994; Hager, 1992, 2000) spielt die Poppersche Falsifikationsmethodologie als wissenschaftsphilosophische Basis nur noch eine unterge-

ordnete Rolle. Wie wir in den folgenden Abschnitten sehen werden, können liberale Formen der deduktiven Methodologie mit Gewinn in eine strukturalistische Methodologie der wissenschaftlichen Forschung eingebettet werden (Westermann, 1987, 2000a).

Struktur von Theorien

Das Hauptziel der *strukturalistischen Wissenschaftstheorie* (Balzer et al., 1987; Stegmüller, 1986) besteht darin, Aufbau, Funktion und Entwicklung wissenschaftlicher Theorien zu rekonstruieren, das heißt, sie so genau wie möglich zu beschreiben und sie so weit wie möglich zu verbessern. Angewendet worden ist sie zur Analyse verschiedener psychologischer Theorien (Gerjets, 1995; Heise, 1991, 1998; Stephan, 1990; Westermann, 1987) sowie auf nahezu alle Wissenschaftsbereiche von der Physik über die Biologie bis zur Ökonomie (Balzer et al., 2000).

Die *strukturalistische Wissenschaftstheorie* beruht auf anderen Ausgangspunkten und setzt andere Akzente als die *deduktive Methodik* von Popper. Erstens ist sie in erster Linie eine Metatheorie und keine Methodologie. Zweitens ist sie weniger normativ und präskriptiv als der Kritische Rationalismus, sonder eher deskriptiv und rekonstruierend. Im Folgenden stelle ich zunächst die metatheoretischen Konzeptionen und Erkenntnisse der strukturalistischen Wissenschaftstheorie vor und gehe dann zu den methodologischen Konsequenzen über.

Theorien als Ordnungsstrukturen

Eine wissenschaftliche Theorie kann insgesamt durch ein *Theorie-Netz* rekonstruiert werden. Es besteht aus verschiedenen *Theorie-Elementen*, die eine teilweise geordnete, baumartige Struktur bilden.

Für jede Theorie können wir ein *Basiselement* identifizieren, das die Grundbegriffe und -annahmen enthält. Im Basiselement zur Dissonanztheorie (Festinger, 1978) beispielsweise wird vor allem ausgedrückt, dass der Druck zur Reduktion einer Dissonanz umso höher ist, je stärker die Dissonanz ist (Westermann, 1987, 2000b). Basiselemente sind in der Regel *empirisch gehaltlos*, das heißt so allgemein und abstrakt, dass ihnen keine Beobachtungen entgegenstehen können.

Bei der Darstellung eines Theorie-Netzes wird das Basiselement an die

Spitze gesetzt, während die anderen Theorie-Elemente darunter aufgefächert werden. Einige dieser Theorie-Elemente enthalten spezifische Annahmen für bestimmte Anwendungsbereiche, beispielsweise für Dissonanz nach Entscheidungen. Andere Theorie-Elemente stellen Modifikationen der Theorie nach abweichenden Befunden dar. Einige Elemente sind abstrakt und allgemein, andere repräsentieren die konkrete theoretische Basis von Prüfexperimenten oder praktischen Anwendungen.

In einem Theorie-Netz sind tiefer angeordnete Elemente differenzierter als hierarchisch darüber angeordnete. Diese *Differenzierungsrelation* kann auf unterschiedliche Weise zustande kommen. Wenn die Zusammenhangsannahmen eines Theorie-Elements verschärft oder um zusätzliche Annahmen ergänzt werden, entsteht ein neues Theorie-Element, das eine *Spezialisierung* des alten ist. Werden auch zusätzliche Begriffe eingeführt, liegt eine *Erweiterung* vor. Die speziellen Annahmen zur Dissonanz nach Entscheidungen beispielsweise beziehen sich auf Konzepte wie die Attraktivität von Alternativen oder Entscheidungssicherheit, von der im Basiselement noch nicht die Rede war.

Häufig stammen diese neuen Begriffe und Annahmen aus anderen Elementen des Theorie-Netzes oder aus ganz anderen Theorien. Diese Verbindungen zwischen den Theorien einer wissenschaftlichen Disziplin können als *intertheoretische Bänder* rekonstruiert werden. Durch sie werden Begriffsbedeutungen spezifiziert und Zusammenhangsannahmen importiert, Operationalisierungen begründet und Anwendungssituationen erschlossen. Bei der Anwendung der Dissonanztheorie beispielsweise müssen andere Theorien herangezogen werden, um Begriffe wie Kognition oder Einstellung hinreichend zu spezifizieren.

Nach der Standardtheorienkonzeption aus dem Logischen Empirismus besteht jede Theorie aus einer bestimmten Menge von Axiomen, aus denen alle anderen Aussagen der Theorie (die Theoreme) logisch abgeleitet werden können (Carnap, 1986; Gadenne, 1994). Die strukturalistischen Rekonstruktionen zeigen dagegen, dass die Übergänge von einem Theorie-Element zu einem anderen nicht einfach logische Implikationen sind, sondern Entscheidungen, Ergänzungen und kreative Modifikationen beinhalten.

Theorie-Elemente und Empirische Hypothesen

Für das Verständnis von Forschungsprozessen wichtig ist die Unterscheidung zwischen dem konzeptuellen *Kern* eines Theorie-Elements mit seinen Begriffen und Annahmen einerseits und seinen *intendierten Anwendungen*

andererseits. Sie verdeutlicht erstens, dass ein System aus Begriffen und Zusammenhangsannahmen erst dann zu einer *empirischen* Theorie wird, wenn es systematisch mit realen Systemen konfrontiert wird. Ansonsten bleibt es eine *formale* (logische oder mathematische) Theorie. Zweitens sind empirische Theorien nicht universell anwendbar, sondern nur auf bestimmte Arten von Systemen. Drittens soll eine Theorie in der Regel bei weitem nicht auf alle Systeme angewendet werden, auf die sie angewendet werden könnte.

Der Kern eines Theorie-Elements kann mithilfe von mengen- und relationstheoretischen Axiomatisierungen sehr genau rekonstruiert werden. Systematisch werden dabei die Begriffe charakterisiert und die angenommenen Zusammenhänge spezifiziert. Dadurch gelingt es, verbal vorformulierte Theorien mathematisch zu präzisieren, ohne in eine Pseudoexaktheit abzugleiten (Bredenkamp u. Graumann, 1973).

Zum Kern eines Theorie-Elements gehört eine Menge von so genannten *Partialmodellen*. Das sind Beschreibungen von *möglichen Anwendungen* des Theorie-Elements, das heißt von Systemen, die in der Begrifflichkeit der Theorie beschreibbar sind.

Der *Geltungsbereich* eines Theorie-Elements ist die Menge der möglichen Anwendungen, für die auch die Zusammenhangsannahmen zutreffen. Wenn ein Theorie-Element empirischen Gehalt hat, kann es auch mögliche Anwendungen geben, bei denen mindestens eine der Zusammenhangsannahmen nicht zutreffend ist, die also außerhalb des Geltungsbereichs liegen.

Eine wissenschaftliche Theorie wird nicht von allen Personen oder Gruppen und nicht zu allen Zeitpunkten in der gleichen Weise vertreten. Die Vertreter einer Theorie unterscheiden sich darin, welche Elemente eines Theorienetzes sie akzeptieren und als Grundlagen ihrer Arbeit verwenden, welche Begriffe und Annahmen für sie in den einzelnen Theorie-Elementen enthalten sind und wie die Mengen der intendierten Anwendungen jeweils zusammengesetzt sind. Solange diese individuellen Theorienetze auf einem gemeinsamen Basiselement aufgebaut sind, wird die gleiche Theorie vertreten. Unterschiedliche Basiselemente stellen dagegen immer verschiedene Theorien dar.

Die Menge I der *Intendierten Anwendungen* eines Theorie-Elements ist formal eine Teilmenge der Menge der Partialmodelle. Sie kann unendlich viele Elemente umfassen, sich beständig in ihrer Zusammensetzung ändern und nur unscharf abgegrenzt werden. Die Menge I teilen wir in vier disjunkte Untermengen ein. Die ersten beiden umfassen die *paradigmatischen* Beispiele im Sinne von Kuhn (1977) und die sonstigen *erfolgreichen* Prü-

fungen und Anwendungen. Eine dritte Teilmenge bilden die *vermuteten* Anwendungen, denn eine Theorie soll auch erfolgreich auf all die Fälle anwendbar sein, deren relevante Aspekte *hinreichend ähnlich* zu den bereits bekannten positiven Anwendungsfällen sind. Vorläufig zur Menge I gezählt werden schließlich die *zweifelhaften* Anwendungen, von denen man zwar denkt, dass sie erfolgreiche Anwendungen der Theorie sind, aber empirische Prüfungen noch für angezeigt hält.

In den offenen und unscharfen Teilmengen der vermuteten und zweifelhaften Anwendungen schlägt sich eine wichtige Eigenschaft von Theorien nieder, die man häufig etwas irreführend als *Allgemeingültigkeit* bezeichnet: Eine Theorie muss mehr sein als eine Zusammenfassung bekannter Befunde, mit ihr ist der Anspruch verbunden, dass sie auch für nicht untersuchte oder gar nicht bekannte Anwendungsfälle gilt. Dies bedeutet allerdings nicht, dass eine Theorie einen universellen Geltungsanspruch haben und »alles« erklären muss. Eine Theorie ist stets auf die Beschreibung und Erklärung bestimmter Systeme und Phänomene beschränkt, innerhalb dieses Bereichs möglicher Anwendungen muss sie allerdings für eine unendliche Menge von Fällen Geltung beanspruchen.

Jedes Theorie-Element T mit seinem Kern K und seiner intendierten Anwendungsmenge I ist mit der *empirischen Behauptung* verbunden, dass alle intendierten Anwendungen auch erfolgreiche sind. Alle Elemente der Anwendungsmenge sollen also zum Geltungsbereich G gehören: $I \subseteq G$. Während ein Theorie-Element ein Instrumentarium von Mengen und Strukturen ist, ist die mit ihm verbundene empirische Behauptung eine Aussage, die im Prinzip wahr oder falsch sein kann.

Ist ein Partialmodell a Element der Menge I, so folgt aus der empirischen Behauptung $I \subseteq G$ die *empirische Hypothese* $a \in G$: Die mögliche Anwendung der Theorie, die durch das Partialmodell a eines Theorie-Elements beschrieben wird, soll eine erfolgreiche Anwendung sein. Empirische Untersuchungen in der Psychologie dienen dazu, derartige empirische Hypothesen zu überprüfen. Für die zweifelhaften Anwendungen stehen diese Prüfungen noch aus, für die anderen intendierten Anwendungen werden die empirischen Hypothesen entweder auf Grund der vorliegenden Untersuchungsergebnisse oder der vermuteten Ähnlichkeit als zutreffend angenommen.

Induktive Argumente im Forschungsprozess

Mit einer Methodologie, die explizit als deduktiv bezeichnet wird, ist der Anspruch verbunden, ohne induktive Argumente auszukommen (Popper, 1994, S. 8) oder auf induktive Elemente so weit wie möglich zu verzichten (Erdfelder u. Bredenkamp, 1994, S. 605). Die strukturalistische Rekonstruktion von Theorien und Forschungsprozessen lässt die Zweifel daran wachsen, ob dies erreichbare und sinnvolle Ziele sind (Westermann u. Gerjets, 1994).

Ein deduktiv-logischer Schluss ist eine logische Implikation von einer Aussage auf eine andere: $A \Rightarrow B$. Er zeichnet sich dadurch aus, dass er *wahrheitskonservierend* und *nicht gehaltserweiternd* ist: Aus der Wahrheit der ersten Aussage (der Prämisse A) folgt zwangsläufig die Wahrheit des zweiten Satzes (der Konklusion B) und der Gehalt der Konklusion geht nicht über den Gehalt der Prämisse hinaus. (Wenn der Gehalt des zweiten Satzes größer wäre als der des ersten Satzes, könnte man aus der Wahrheit des ersten nicht mit Sicherheit auf die Wahrheit des zweiten schließen.)

Demgegenüber sind induktive Argumente gehaltserweiternd (*ampliativ*) und deshalb nicht wahrheitskonservierend. Nicht alle gehaltserweiternden Argumente werden jedoch als induktiv betrachtet. Vielmehr sprechen wir von Induktion nur dann, wenn die Prämisse die Konklusion in irgendeiner Weise unterstützt, ihr Gewicht verleiht oder sie wahrscheinlicher macht.

Induktive Argumente treten in Forschungsprozessen vor allem in Zusammenhang mit *empirischen Vermutungen* auf, das heißt bei empirischen Hypothesen, die sich auf vermutete Anwendungen eines Theorie-Elements beziehen. Wenn eine Theorie sich in einer paradigmatischen oder sonstigen erfolgreichen Anwendung bewährt hat (Prämisse), vermuten wir, dass sie auch in einem anderen, ähnlichen Fall erfolgreich angewendet werden kann (Konklusion). Die Konklusion dieses »Schlusses« wird durch die Prämisse zwar nicht impliziert, aber durchaus gestützt. Das Argument ist deshalb kein deduktives, sondern ein induktives.

Empirische Vermutungen finden sich sowohl bei der wissenschaftlichen Prüfung als auch bei der praktischen Anwendung von Theorien. Wie bereits erwähnt, erfordert die wissenschaftliche Überprüfung einer Theorie in einer bestimmten Situation stets bestimmte Zusatzannahmen, beispielsweise über adäquate Operationalisierungen oder erfolgreiche Kontrollen von Störfaktoren. Diese zusätzlichen Hypothesen oder Theorien können zwar gut bewährt sein, in der Regel können sie aber für die vorliegende Situation nicht erneut überprüft werden. Dass sie auch in der vorliegenden Situation zutreffend sind, kann nur induktiv aus den positiven Erfahrungen in ähn-

lichen Situationen abgeleitet werden. Selbst wenn eine empirische Prüfung von (einigen) Zusatzannahmen realisiert wird, beruht diese zwangsläufig auf weiteren Zusatzannahmen usf.

Auch im Alltag und in der Berufspraxis wenden wir Theorien an, um Phänomene zu erklären und Maßnahmen zu begründen, ohne dass wir deren Gültigkeit im jeweiligen Einzelfall überprüfen können.

Wenn wir akzeptieren, dass induktive Argumente im Forschungsprozess unvermeidlich sind, entsteht das Bedürfnis, mehr oder minder berechtigte Induktionen zu unterscheiden. Da alle analytischen Explikationen der Stärke einer induktiven Beziehung durch bedingte subjektive oder logische Wahrscheinlichkeiten (Carnap, 1971) unbefriedigend bleiben, kann die Berechtigung eines induktiven Arguments nur pragmatisch bewertet werden. Allgemein soll sie umso höher sein, je bessere Gründe es für die Erwartung gibt, dass die Konklusion wahr ist oder sich bei einer Überprüfung als wahr herausstellen wird. Eine Analyse der philosophischen und psychologischen Konzeptionen zum induktiven Denken hat zu vier Induktionsfaktoren geführt, deren Einfluss auch in empirischen Untersuchungen demonstriert worden ist (Gerjets et al., 1994; Lütkemeier et al., 2003). Die Berechtigung eines induktiven Schlusses ist (jeweils unter sonst gleichen Bedingungen) umso größer, je größer die *Anzahl* der bisherigen positiven Fälle ist, die in die Prämisse eingehen, je größer ihre *Validität* ist, je größer die *Variabilität* zwischen ihnen ist und je mehr relevante *Ähnlichkeit* es zwischen den neuen und den bisherigen positiven Fällen gibt.

Ziele und Kriterien empirischer Forschung

In der deduktiven Methodologie sensu Popper besteht das Ziel der wissenschaftlichen Forschung darin, die aufgestellten Theorien zu falsifizieren. Die Falsifikation einer Theorie soll aber nicht dazu führen, dass sie einfach aufgegeben wird. Sie soll vielmehr dazu dienen, die Theorie so zu verbessern, dass sie auch die abweichenden Befunde mit erklären kann und dadurch der Wahrheit näher kommt (Popper, 1965). Wenn die Falsifikation das Ziel der wissenschaftlichen Forschung ist, dann müssen alle wissenschaftlichen Untersuchungen möglichst strenge Prüfungen der Theorie sein, das heißt, sie müssen so gestaltet sein, dass falsche Theorien mit hoher Wahrscheinlichkeit scheitern. Dazu sollten aus der Theorie möglichst mutige Hypothesen *(bold conjectures)* abgeleitet werden. Das können Vorhersagen sein, die nach dem bisherigen Wissen sehr unwahrscheinlich sind, die

quantitativ sehr präzise sind, die sich auf viele unterschiedliche Operationalisierungen beziehen usf.

Je strenger eine Prüfung ist, desto eher scheitern aber auch brauchbare Theorien. Im Extremfall kann eine Prüfung so streng sein, dass auch eine sehr gute Theorie praktisch mit Sicherheit scheitert. Mit den empirischen Ergebnissen wäre dann aber kein Erkenntnisgewinn mehr verbunden und wir hätten keine bewährten Theorien mehr. Deshalb müssen Untersuchungen so gestaltet werden, dass Theorien auch eine faire Bewährungschance haben (Gadenne, 1984; Hager u. Westermann, 1983).

Aus der Sicht der strukturalistischen Theorienkonzeption und nach den Ergebnissen strukturalistischer Rekonstruktionen sind Theorien allerdings gar nicht falsifizierbar, sondern weitgehend *immun gegenüber aufsässigen Erfahrungen* (Stegmüller, 1980, 1986; Westermann, 1987, 2000a). Dies liegt vor allem darin begründet, dass sowohl die Menge der Theorie-Elemente als auch die Mengen der intendierten Anwendungen unscharf, offen und variabel sind. Empirisch prüfbar sind deshalb immer nur bestimmte empirische Behauptungen und Hypothesen.

Ziel empirischer Forschungsprozesse ist es aus strukturalistischer Sicht, unser Wissen sowohl über empirische Zusammenhänge als auch über ihre theoretischen Erklärungen zu mehren. Jede erfolgreiche Anwendung der Theorie auf eine neue Situation ist ein empirischer und jede Weiterentwicklung der Theorie ein theoretischer Fortschritt. Auch in empirischen oder theoretischen Rückschlägen liegt Erkenntnisgewinn: Wir erfahren, dass die erwarteten Zusammenhänge nicht bestehen oder dass ein Teil der Theorie nicht aufrechtzuerhalten ist.

Ist nach den Untersuchungsergebnissen ein empirisch gehaltvolles Theorie-Element nicht erfolgreich auf eine bestimmte Situationsklasse anwendbar, können die entsprechenden Partialmodelle aus der Menge der intendierten Anwendungen entfernt werden. Die empirische Behauptung des Theorie-Elements wird dann entsprechend eingeschränkt und man muss nach alternativen theoretischen Erklärungen suchen. Die Theorie-Elemente bestimmen ihren Anwendungsbereich also teilweise selber (partielle *Autodetermination*).

Wenn sich für ein Theorie-Element nicht die erhofften erfolgreichen Anwendungen finden lassen und dieses Element aufgegeben wird, müssen dadurch die anderen Elemente des Theorie-Netzes nicht tangiert werden, zumal sie durchaus erfolgreiche Anwendungen haben können.

Das unmittelbare Ziel der empirischen Forschung besteht damit darin, die erfolgreichen Anwendungen der einzelnen Elemente einer wissenschaftlichen Theorie gegen die möglichen Anwendungen abzugrenzen, die

nicht zum Geltungsbereich gehören. Dies bedeutet, dass man nach Interaktionen und insbesondere nach *disordinalen Interaktionen* sucht (vgl. Bredenkamp, 1979): zwischen den interessierenden Ursachenvariablen und Eigenschaften von Personen oder Situationen, sodass unter mindestens einer Eigenschaftsausprägung die von der Theorie vorhergesagten Ursachen-Wirkungs-Zusammenhänge gelten, unter mindestens einer anderen jedoch nicht.

Jede empirische Studie, in der untersucht wird, ob eine Theorieform, eine Hypothese, ein Modell oder ein Zusammenhang für eine bestimmte Personen-Situations-Konfiguration gilt, trägt zur Abgrenzung erfolgreicher und nicht-erfolgreicher Anwendungen bei. Aus strukturalistischer Sicht wird in ihr mindestens eine empirische Hypothese (EH) geprüft, dass ein bestimmtes Partialmodell a zum Geltungsbereich eines bestimmten Theorie-Elements T gehört: $a \in G(T)$.

Die Validität einer Untersuchung bezeichnet ihre Eignung, Güte oder Fähigkeit, zu einer richtigen Entscheidung über die empirische Hypothese zu führen. Sie ist umso größer, je kleiner die Wahrscheinlichkeit für falsche Entscheidungen über die EH ist (Westermann, 1987). Die bedingte Wahrscheinlichkeit für eine *fälschliche Akzeptierung* der EH wird mit e bezeichnet, die bedingte Wahrscheinlichkeit für eine *fälschliche Ablehnung* mit f. Die numerischen Werte von e und f sind nicht zu bestimmen, man kann aber Faktoren benennen, die sie unter sonst gleichen Bedingungen erhöhen oder verringern.

Erhöht werden können die Fehlerwahrscheinlichkeiten durch alle Störfaktoren der Untersuchungsvalidität (interne Validität, statistische Validität, Variablenvalidität), die an verschiedenen Stellen aus induktiver, deduktiver und strukturalistischer Sicht beschrieben und interpretiert worden sind (Bredenkamp, 1980; Cook u. Campbell, 1979; Hager u. Westermann, 1983; Westermann, 2000a). Umgekehrt können e und/oder f durch die entsprechenden Techniken zur Kontrolle der Störfaktoren reduziert werden.

Poppers Kriterium möglichst strenger Prüfungen entspricht der Forderung, dass die Wahrscheinlichkeit e möglichst klein sein soll. Dies wird (jeweils unter sonst gleichen Bedingungen) durch ein kleineres Signifikanzniveau erreicht oder beispielsweise dadurch, dass man die EH nur akzeptiert, falls sie sich in allen Untergruppen oder für alle abhängigen Variablen bewährt.

Unter sonst gleichen Bedingungen führen Maßnahmen zur Verringerung von e zu einer Erhöhung von f. Je kleiner f ist, desto *wohlwollender* nennen wir die Prüfung.

Prinzipiell lässt sich jede Untersuchung so gestalten, dass sie extrem

streng oder extrem wohlwollend ist. In beiden Fällen sagt das Untersuchungsergebnis aber wenig über die tatsächliche Gültigkeit der geprüften Hypothese aus. Aus strukturalistischer Sicht sind Untersuchungen in der Regel so zu gestalten, dass Strenge und Wohlwollen etwa gleich hoch sind. Wenn dies der Fall ist, bezeichnen wir die Untersuchung als eine *faire Prüfung* der empirischen Hypothese. Sie hat dann den höchsten Informationswert für die Abgrenzung der erfolgreichen von den nicht-erfolgreichen Anwendungsgebieten.

Des Weiteren ergeben sich aus strukturalistischer Sicht stringente Begründungen für methodische Desiderate, auf die schon Bredenkamp (1972) aufmerksam gemacht hat. Die erste wichtige Forderung ist die nach Kontrolle der Teststärke: Sie ist (bei den üblichen Vorhersagen von Mittelwertsrangordnungen) gerade dann wichtig, wenn nicht nur Falsifikationen angezielt werden, sondern auch erfolgreiche Theorieanwendungen identifiziert werden sollen (Westermann u. Hager, 1983). Der zweite wichtige Punkt ist die Interpretation von Signifikanztests als (approximative) Randomisierungstests ohne Populationsannahmen: Sie ist besonders angezeigt, wenn eine Untersuchung nicht zur Prüfung einer Theorie insgesamt verwendet werden soll, sondern strukturalistisch als Prüfung einer empirischen Hypothese über ihre Anwendbarkeit in einer konkreten Situation mit bestimmten Personen, Bedingungen und Ergebnissen (Westermann u. Siemer, 2001).

Von der deduktiven zur theoriegeleiteten Methodologie

Der Grundgedanke der deduktiven Methodologien kann folgendermaßen zusammengefasst werden: Aus der interessierenden Theorie (oder wissenschaftlichen Hypothese) wird eine Aussage über eine Beobachtung oder eine statistische Hypothese abgeleitet und anhand empirischer Daten überprüft. Trifft diese deduzierte Vorhersage nicht zu, kann auf die Falschheit der übergeordneten Theorie geschlossen werden.

Mit einer deduktiven Methodologie ist der Anspruch verbunden, dass im Forschungsprozess deduktiv-logische Argumente die zentrale Rolle spielen und alle Formen von induktiven Argumenten vermieden werden. Strenge Formen der deduktiven Methodologie wollen Induktionen vollständig ausschalten und unterstellen eine implikative Beziehung zwischen Theorien oder wissenschaftlichen Hypothesen einerseits und Daten oder statistischen Hypothesen andererseits. Pragmatischere Formen akzeptieren

auch schwächere Formen von Ableitungsbeziehungen und versuchen lediglich, die Rolle induktiver Argumente zu minimieren.

Für sich genommen impliziert eine wissenschaftliche Theorie jedoch weder bestimmte beobachtbare Ereignisse noch bestimmte statistische Hypothesen. Es sind vielmehr immer Zusatzannahmen notwendig. Die Ableitung einer Beobachtung aus einer Theorie zusammen mit Zusatzannahmen ($T \wedge Z \Rightarrow B$) ist zwar ein deduktiver Schluss, die Beobachtung folgt aber gar nicht, wie eigentlich beabsichtigt, deduktiv aus der Theorie allein.

Die für die Ableitung von Vorhersagen notwendigen Zusatzannahmen sind keineswegs nur technische Hilfshypothesen (wie beispielsweise Annahmen über die richtige Funktionsweise von Messinstrumenten), die problemlos als erfüllt angesehen werden können. Sowohl zur praktischen Anwendung als auch zur experimentellen Prüfung von Theorien müssen vielmehr häufig zusätzliche Annahmen herangezogen werden, die eigenständigen Theorie-Elementen aus dem betrachteten oder anderen Theorie-Netzen entsprechen. Die Annahme, dass diese zusätzlichen Theorie-Elemente auch in der vorliegenden Untersuchungssituation erfüllt sind, kann in der Regel nicht empirisch geprüft werden, sondern ist eine empirische Vermutung, beruht also auf einem induktiven Argument.

Aus strukturalistischer Sicht ist das wesentliche Charakteristikum wissenschaftlicher Forschungsprozesse nicht das Deduktiv-Logische, das gegen das Induktive abgegrenzt werden muss. Wichtig ist vielmehr, dass erfolgreiche Forschung durchgängig theoriebasiert ist und sich nicht in der Sammlung von empirischen Beobachtungen verliert.

Literatur

Balzer, W., Moulines, C.U., Sneed, J.D. (1987). *An architectonic for science: The structuralist program.* Dordrecht: Reidel.

Balzer, W., Sneed, J.D., Moulines, C.U. (Hg.) (2000). *Structuralist knowledge representation: Paradigmatic examples.* Amsterdam: Rodopi.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1979). Das Problem der externen Validität pädagogisch-psychologischer Untersuchungen. In J. Brandtstädter, G. Reinert u. K.A. Schneewind (Hg.), *Pädagogische Psychologie: Probleme und Perspektiven* (S. 267–289). Stuttgart: Klett-Cotta.

Bredenkamp, J., Graumann, C.F. (1973). Möglichkeiten und Grenzen mathematischen Verfahrens in den Verhaltenswissenschaften. In P. Vogler (Hg.), *Psychologische Anthropologie* (Neue Anthropologie, Band 5) (S. 51–93). Stuttgart: Thieme.

Carnap, R. (1926). *Physikalische Begriffsbildung.* Karlsruhe: Braun.

Carnap, R. (1986). *Einführung in die Philosophie der Naturwissenschaften.* Frankfurt a. M.: Ullstein.

Carnap, R. (1971). A basic system of inductive logic, part I. In R. Carnap u. R.C. Jeffrey (Hg.), *Studies in inductive logic and probability* (Bd. I, S. 33–165). Berkeley, CA: University of California Press.

Cook, T.D., Campbell, D.T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann u. W.H. Tack (Hg.), *Methodologische Grundlagen der Psychologie* (Enzyklopädie der Psychologie, Serie Forschungsmethoden der Psychologie, Band 1, S. 604–648). Göttingen: Hogrefe.

Festinger, L. (1978). *Theorie der kognitiven Dissonanz.* Bern: Huber.

Gadenne, V. (1976). *Die Gültigkeit psychologischer Untersuchungen.* Stuttgart: Kohlhammer.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung.* Tübingen: Mohr.

Gadenne, V. (1994). Theorien. In T. Herrmann u. W.H. Tack (Hg.), *Methodologische Grundlagen der Psychologie* (Enzyklopädie der Psychologie, Serie Forschungsmethoden der Psychologie, Band 1, S. 295–342). Göttingen: Hogrefe.

Gerjets, P. (1995). *Zur Verknüpfung psychologischer Handlungs- und Kognitionstheorien.* Frankfurt a. M.: Lang.

Gerjets, P., Westermann, R., Lütkemeier, E. (1994). Induktive Erwartungsbildung: Einflußfaktoren auf die Berechtigung empirischer Vermutungen über die Anwendbarkeit handlungsbezogener Theorien. *Sprache & Kognition, 13,* 52–64.

Hager, W. (1984). Aspekte eines deduktiven Forschungsansatzes in der Empirischen Pädagogik: Fragen der Validität der Hypothesenableitung, der Untersuchungsplanung und der Hypothesenbeurteilung. *Zeitschrift für Empirische Pädagogik und Pädagogische Psychologie, 8,* 55–75.

Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen.* Göttingen: Hogrefe.

Hager, W. (2000). A principle of derivation and decisions based on statistical tests and effect sizes. *MPR-Online, 5,* 9–41.

Hager, W., Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp u. H. Feger (Hg.), *Hypothesenprüfung* (Enzyklopädie der Psychologie, Serie Forschungsmethoden der Psychologie: Band 5, S. 24–238). Göttingen: Hogrefe.

Hays, W.L. (1981). *Statistics* (3. Aufl.). New York: Holt, Rinehart & Winston.

Heise, E. (1991). *Strukturalistische Rekonstruktion der ACT*-Theorie von Anderson.* Frankfurt a. M.: Lang.

Heise, E. (1998). *Volitionale Handlungskontrolle: Theoretische und empirische Analysen auf strukturalistischer Basis.* Münster: Waxmann.

Kuhn, T.S. (1977). Neue Überlegungen zum Begriff des Paradigma. In L. Krüger (Hg.), T.S. Kuhn. *Die Entstehung des Neuen. Studien zur Struktur der Wissenschaftsgeschichte* (S. 389–415). Frankfurt a. M.: Suhrkamp.

Lütkemeier, E., Westermann, R., Gerjets, P. (2003). Induktive Vermutungen über die Anwendbarkeit von Theorien und Methoden nach erfolgreichen und gescheiterten Anwendungsfällen. *Zeitschrift für Psychologie, 211,* 38–56.

Meehl, P.E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Popper, K.R. (1965). *Conjectures and refutations. The growth of scientific knowledge* (2. Aufl.). London: Routledge and Kegan Paul.

Popper, K.R. (1994). *Logik der Forschung* (10. Aufl.). Tübingen: Mohr.

Stegmüller, W. (1980). *Neue Wege der Wissenschaftsphilosophie.* Berlin: Springer-Verlag.

Stegmüller, W. (1986). *Theorie und Erfahrung: Dritter Teilband. Die Entwicklung des neuen Strukturalismus seit 1973 (Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie: Band 2).* Berlin: Springer-Verlag.

Stephan, E. (1990). *Zur logischen Struktur psychologischer Theorien.* Berlin: Springer-Verlag.

Westermann, R. (1987). *Strukturalistische Theorienkonzeption und empirische Forschung in der Psychologie.* Berlin: Springer-Verlag.

Westermann, R. (2000a). *Wissenschaftstheorie und Experimentalmethodik: Ein Lehrbuch zur psychologischen Methodenlehre.* Göttingen: Hogrefe.

Westermann, R. (2000b). Festinger's theory of cognitive dissonance: A structuralist theory-net. In W. Balzer, J.D. Sneed u. C.U. Moulines (Hg.), *Structuralist knowledge representation: Paradigmatic examples* (S. 189–217). Amsterdam: Rodopi.

Westermann, R., Gerjets, P. (1994). Induktion. In T. Herrmann u. W.H. Tack (Hg.), *Methodologische Grundlagen der Psychologie* (Enzyklopädie der Psychologie, Serie Forschungsmethoden, Band 1, S. 428–471). Göttingen: Hogrefe.

Westermann, R., Hager, W. (1983). On severe tests of trend hypotheses in psychology. *Psychological Record, 33,* 201–211.

Westermann, R., Siemer, M. (2001). Datenauswertung ohne Annahmen über Populationen und Zufallsauswahl: Permutationstests und Bootstrapping. In R.K. Silbereisen u. M. Reitzle (Hg.), *Psychologie 2000. Bericht über den 42. Kongreß der Deutschen Gesellschaft für Psychologie in Jena 2000* (S. 724–735). Lengerich: Pabst.

Willi Hager

Psychologische und statistische Hypothesen: Ableitung von Vorhersagen und Entscheidungen

Zusammenfassung

Dieser Beitrag befasst sich mit der Ableitung von statistischen aus psychologischen Hypothesen über aus diesen abgeleiteten psychologischen Vorhersagen sowie mit möglichen Entscheidungen über die psychologischen Hypothesen. Es werden für die statistische Prüfung von psychologischen Hypothesen vier Ebenen konstruiert: die der psychologischen Hypothese, die der psychologischen Vorhersage, die der statistischen Vorhersage und die der statistischen Tests. Es wird dann an verschiedenen einfachen Beispielen demonstriert, wie eine Ableitung von Vorhersagen aus einer psychologischen Hypothese auf diesen Ebenen erfolgen kann. Während die Entscheidung auf den statistischen Ebenen und der psychologischen Vorhersage ein weitgehend formaler Akt ist, soll die Entscheidung über die vorgeordnete psychologische Hypothese niemals formal sein, sondern soll neben den Versuchsergebnissen vor allem von der Gesamt-Validität des Versuchs und den Kenntnissen des Experimentators abhängen.

Betrachtungsebenen

Eine der wichtigsten Aufgaben einer sich als empirisch verstehenden Psychologie stellt die Prüfung von Theorien und psychologischen Hypothesen dar. Solche sollten für jede psychologische Untersuchung explizit gemacht werden und den Versuch leiten. Ohne vorangestellte psychologische Hypothese ist kein Versuch vernünftig planbar. Psychologische Hypothesen werden über aus ihnen abgeleitete Vorhersagen geprüft. Die Ableitung von Vorhersagen relativ zu einer vorgegebenen psychologischen Hypothese soll zweckmäßigerweise immer vier Ebenen betreffen, die sich durch die Begrifflichkeiten unterscheiden, mit denen auf ihnen operiert wird.

Ebene der psychologischen Hypothese (PH)

Psychologische Hypothesen können Kausalhypothesen von der Form »Wenn . . ., dann . . .« darstellen und sind dann üblicherweise in eine Theorie eingebettet (in diesem Fall wäre eine Ebene der psychologischen Theorie vorzuordnen). Psychologische Hypothesen können lediglich einen Zusammenhang zwischen zwei oder mehr Variablen behaupten und stellen dann Zusammenhangshypothesen dar. Bei psychologischen Hypothesen kann es sich auch um »eine vorläufig unbegründete Antizipation« oder einen »Einfall« handeln (Popper, 1984, S. 7). Eine psychologische Hypothese kann wesentlich Begriffe, Variable oder Konstrukte der Psychologie enthalten, die nicht-beobachtbare oder theoretische Größen darstellen und die für die empirische Prüfung beobachtbar gemacht oder *konkretisiert* beziehungsweise operationalisiert werden müssen. Psychologische Hypothesen können sich auf eine qualitative unabhängige Variable (Bsp.: »Bildhaftigkeit«) beziehen oder auf eine organismische (Bsp. »Geschlecht«). Bei beiden Hypothesenarten können die Ausprägungen der unabhängigen Variablen monoton geordnet sein (Bsp.: »Bildhaftigkeit«; Ausnahme: z. B. »Geschlecht«) und die gleichen Beziehungen zwischen unabhängiger (UV) und abhängiger Variablen (AV) behauptet werden. Ferner kann die (empirische) unabhängige Variable der psychologischen Hypothese quantitativ sein, sodass neben der Rangordnung auch die Abstände zwischen je zwei Stufen der unabhängigen Variablen interpretierbar sind; es kann dann eine funktionale Beziehung zwischen empirischer unabhängiger und abhängiger Variablen behauptet werden (Bsp.: »Darbietungszeiten in einem Lernexperiment«). Eine einfaktorielle psychologische Hypothese behauptet eine Beziehung zwischen zwei Variablen, einer unabhängigen und einer abhängigen. Diese Behauptung kann gerichtet oder ungerichtet sein; zuweilen wird auch das Fehlen einer Beziehung behauptet. Von mehrfaktoriellen psychologischen Hypothesen werden mehrere unabhängige Variable angesprochen.

Die meisten psychologischen Hypothesen beziehen sich auf einzelne Personen; sie werden dennoch in aller Regel (zumeist implizit) nur probabilistisch oder »im Mittel« geprüft, indem man zur Prüfung einen der statistischen Tests heranzieht, die fast durchgängig mit mittleren Werten operieren. Wenn man dies explizit machen will, versieht man die psychologische Hypothese mit Wahrscheinlichkeitsklauseln der Art »im Mittel gilt, dass . . .« (Probabilisierung; Westermann, 2000, S. 148–150). Die übliche Prüfung »im Mittel« dient aus der Sicht der experimentellen Validität der Minimierung des Einflusses von möglichen Störfaktoren bei der Hypothesenprüfung und der Konstanthaltung aller Faktoren über die Faktorstufen

(»Ceteris-paribus-Bedingung«; Westermann, 2000, S. 156–157), die nicht explizit von der Hypothese angesprochen werden und die im Allgemeinen bei Einzelfallprüfungen nicht befriedigend kontrolliert werden können (vgl. Erdfelder u. Bredenkamp, 1994, S. 607–608). Um mit *probabilistischen Gesetzen*, die mit den psychologischen Hypothesen assoziiert sind, ein wichtiges Ziel psychologischen Forschens, die theoretische *Erklärung* psychischer Phänomene, Sachverhalte, Prozesse und/oder psychischen Erlebens, zu erreichen, kann man sich allerdings nicht mehr des bekannten Hempel-Oppenheim-Schemas der Erklärung bedienen, das deterministische Gesetzesaussagen voraussetzt (Westermann, 2000, S. 166–168). An seine Stelle treten die *induktiv-statistischen Systematisierungen* (Westermann, 2000, S. 176–178). Genau wie beim Hempel-Oppenheim-Schema müssen sich allerdings die zur theoretischen Erklärung herangezogenen Gesetze auf theoretische Begriffe beziehen, die über das Beobachtbare hinausgehen (Westermann, 2000, S. 165–182; S. 245). Sind psychologische Hypothesen über Beobachtungsbegriffe formuliert, können sie dem Ziel der *Beschreibung* dienen.

Nach Popper (1984) sollen alle Prüfungen von (psychologischen) Hypothesen *kritisch* und *streng* erfolgen, um die Hypothese einem hohen Falsifikationsrisiko auszusetzen, und nach Westermann (1987, S. 109) sollen diese Prüfungen auch *wohlwollend* gestaltet werden, um der psychologischen Hypothese eine Chance zur Bewährung zu geben. Dabei sind zwei Bedeutungen von »Strenge« zu unterscheiden. Einmal handelt es sich um die *Strenge der Prüfbarkeit*, die eine vom Versuchsleiter (Vl) nicht beeinflussbare Eigenschaft der Hypothesen darstellt (Popper, 1984, S. 85–86). Eine psychologische Hypothese ist umso strenger prüfbar, je höher ihr empirischer Gehalt im Sinne Poppers (a. a. O.) ist, je mehr sie also verbietet. Und zum anderen handelt es sich um die *Strenge der Prüfungen* von Hypothesen, die eine Eigenschaft der Prüfsituation darstellt und die durch entsprechende Maßnahmen der Versuchsplanung vonseiten des Versuchsleiters sehr wohl beeinflussbar ist. Die beiden Forderungen »Strenge und Wohlwollen von Prüfungen« können über bedingte »Risiken« definiert werden. Dabei bezeichnet e das »Risiko«, eine nicht-zutreffende psychologische Hypothese irrtümlich für bewährt zu halten, während f für das »Risiko« steht, eine zutreffende psychologische Hypothese irrtümlich für nicht bewährt zu halten. e und f werden hier »Risiken« genannt und nicht Fehler*wahrscheinlichkeiten*, da unklar ist, inwiefern sie mit dem Kalkül der Wahrscheinlichkeitstheorie vereinbar sind. Die Bezeichnung »Risiko« in diesem Kontext stammt von A. Iseler, Berlin. Mit dem »Gegenrisiko« $1 - e$ wird eine zutreffende Entscheidung *gegen* und mit dem »Gegenrisiko« $1 - f$ eine zutreffende

Entscheidung *für* die psychologische Hypothese getroffen. Die *e-Validität* oder *Strenge einer Untersuchung* ist umso größer, je größer das »Gegenrisiko« $1 - e$ ist. Das *Wohlwollen* oder die *f-Validität einer Untersuchung* ist umso größer, je größer das »Gegenrisiko« $1 - f$ ist. Diese beiden Validitätsarten sind der klassischen experimentellen Validität von Versuchen vorgeordnet. Mittels Maßnahmen der Versuchsplanung können die Gegenrisiken systematisch beeinflusst werden.

Ebene der psychologischen Vorhersage (PV)

Auf der Ebene der psychologischen Vorhersage werden zunächst die theoretischen Konstrukte oder Variablen der psychologischen Hypothese konkretisiert, das heißt beobachtbar gemacht. Bei der Wahl dieser *Konkretisierungen* spielen unter anderem die Theorie, aus der die psychologische Hypothese gewonnen wurde, und das verfügbare Hintergrundwissen sowie Hilfshypothesen ohne empirischen Gehalt (also empirisch nicht prüfbare Hilfshypothesen) eine große Rolle; man benutzt häufig die paradigmatischen Konkretisierungen, für die sich die psychologische Hypothese bereits bewährt hat beziehungsweise die auch bei den ersten erfolgreichen Anwendungen der Theorie benutzt wurden, oder doch zumindest solche, die einen hohen Plausibilitätsgrad haben, oder solche, auf die sich eine Forschergemeinschaft, die im selben Bereich forscht, explizit oder implizit geeignet hat. Diese Hilfshypothesen ohne empirischen Gehalt fungieren also als eine Art »Zuordnungsregel«, wobei die »Richtigkeit« oder »Güte« der Zuordnung in keinem Fall empirisch überprüft werden kann. Auch kann man neuartige Konkretisierungen erproben, um versuchsweise den Anwendungsbereich der psychologischen Hypothese zu erweitern. Die empirische Konkretisierung der theoretischen abhängigen Variablen hat ein bestimmtes *Skalenniveau*, das bei den weiteren Ableitungen zu berücksichtigen ist.

Zusätzlich müssen auf dieser Ebene die Maßnahmen zur Kontrolle von potenziellen Störfaktoren festgelegt und ein *Versuchsplan* sowie die Art Bedingungsvariation gewählt werden (inter- vs. intraindividuell), etwa ein einfaktorieller Plan mit K Faktorstufen. *Ohne den Bezug auf einen Versuchsplan ist das Ableiten von Vorhersagen nicht möglich.* Erst wenn die Randbedingungen des durchzuführenden Versuchs einschließlich des Versuchsmaterials vollständig spezifiziert sind, wird es möglich, aus der psychologischen Hypothese eine psychologische Vorhersage abzuleiten, die sich immer auf die konkrete und vollständig spezifizierbare Versuchssituation, auf die konkretisierten oder beobachtbaren Variablen sowie auf die gerichtete oder ungerichtete

Behauptung der psychologischen Hypothese bezieht. Werden bereits in der psychologischen Hypothese empirische Variablen angesprochen, dann unterscheidet sich die psychologische Vorhersage nur durch den Bezug auf den vollständig spezifizierten Versuch von der psychologischen Hypothese. Die Beziehung zwischen psychologischer Vorhersage und psychologischer Hypothese wird hier im Sinne Meehls (1967, S. 104) nur »in rather loose sense of derive« rekonstruiert, wird aber durch die Zusatzspezifikationen der Adäquatheit und der Erschöpfendheit verschärft, die beide zusammen die Ableitungsvalidität ausmachen.

Ebene der statistischen Vorhersage (SV)

Wenn eine statistische Prüfung mittels statistischer Tests vorgesehen ist, muss die psychologische Vorhersage in eine statistische Vorhersage umgesetzt werden. Dazu werden die individuellen Werte der beobachtbaren abhängigen Variablen als Realisierungen einer Zufallsvariablen interpretiert. Über diese Zufallsvariable formuliert man dann verschiedene Hilfshypothesen und Zusatzannahmen, etwa über ihre theoretische Verteilung (z. B. Normalverteilung) und ein bestimmtes Datenmodell. Aus der psychologischen Vorhersage wird dann eine statistische Vorhersage abgeleitet, und zwar erneut »in rather loose sense of derive«, denn: »Zwischen psychologischen und statistischen Hypothesen besteht . . . keine Implikationsbeziehung, sondern allenfalls eine Ableitungsbeziehung in einem relativ lockeren Sinn« (Westermann, 2000, S. 388). Die statistische Vorhersage umfasst mindestens eine statistische Hypothese, die das Skalenniveau der beobachtbaren abhängigen Variablen berücksichtigt und die sich nur auf solche statistischen Kennwerte bezieht, die mit planbaren Tests getestet werden können (siehe dazu unten). Die statistische Vorhersage kann Wahrscheinlichkeiten oder Proportionen π, erwartete mittlere Ränge und Mittelwerte (μ_k), Varianzen (σ^2_k), einfache (ϱ) oder quadrierte multiple Korrelationen (η^2) betreffen; für Hypothesen über diese statistischen Kennwerte kann durchgängig eine Testplanung erfolgen (vgl. etwa Cohen, 1988; Bredenkamp, 1980; Hager, 2004). Die Ableitung einer statistischen Vorhersage kann dabei *stets nur aus einer psychologischen Vorhersage, nicht aber direkt aus einer psychologischen Hypothese erfolgen, weil zum Ableiten einer statistischen Vorhersage stets Bezug auf einen Versuchsplan und auf die empirischen Variablen genommen werden muss.*

Die statistische Prüfung einer psychologischen Hypothese über eine psychologische Vorhersage und eine statistische Vorhersage führt stets zu *a priori* geplanten Kontrasten, Vergleichen und statistischen Tests, bei denen es

unerheblich ist, ob sie orthogonal zueinander sind oder nicht. Wesentlich ist allein die »Passung« zwischen psychologischer Hypothese und psychologischer sowie statistischer Vorhersage.

Für die statistische Vorhersage werden die bedingten Fehlerwahrscheinlichkeiten ε und φ eingeführt, wobei ε die bedingte Wahrscheinlichkeit bezeichnet, eine nicht-zutreffende statistische Vorhersage irrtümlich anzunehmen, und φ die bedingte Wahrscheinlichkeit, eine zutreffende statistische Vorhersage irrtümlich abzulehnen.

Die statistische Vorhersage besteht also aus einer oder mehreren statistischen Hypothesen, die mittels eines verfügbaren Tests direkt testbar sein können oder nicht (z. B. ist so gut wie jede statistische Hypothese über einen oder über zwei Parameter direkt testbar, während die Hypothese eines monotonen Trends bei $K \geq 3$ nicht mit nur einem Test testbar ist).

Ebene der statistischen Tests (ST)

Im letzteren Fall werden dann auf der Ebene der statistischen Tests (ST) aus der statistischen Vorhersage direkt testbare Null- und/oder Alternativhypothesen (H_0, H_1) abgeleitet (Zerlegung der statistischen Vorhersage in testbare Einzelhypothesen). Auf dieser Ebene sind die bedingten Fehlerwahrscheinlichkeiten α, die Wahrscheinlichkeit für eine irrtümliche Annahme der H_1, und β definiert, die Wahrscheinlichkeit einer fälschlichen Beibehaltung der H_0. Bei der statistischen Prüfung psychologischer Hypothesen ist es unabdingbar, die beiden Fehlerwahrscheinlichkeiten α und β simultan zu kontrollieren (z. B. Bredenkamp, 1969, 1972, 1980; Erdfelder u. Bredenkamp, 1994, S. 609). Alle Tests, die diese Simultankontrolle mittels Teststärkenanalyse ermöglichen, heißen planbar. Wenn die psychologische Hypothese, die psychologische und die statistische Vorhersage eine gerichtete oder ungerichtete Zusammenhangsaussage umfasst, wird/werden aus der psychologischen und aus der statistischen Vorhersage im Allgemeinen eine oder $R > 1$ $H_{1,r}$ ($r = 1, 2, \ldots, R$) abgeleitet. In diesem Fall gilt: SV-A $\Leftrightarrow$ H_1: $\varepsilon = \alpha$ und $\varphi = \beta$ (»$\Leftrightarrow$«: »ist äquivalent zu«). Ein β-Fehler führt mit der Wahrscheinlichkeit φ zu einer falschen Annahme der SV-A, und ein α-Fehler führt mit der Wahrscheinlichkeit ε zu einer falschen Ablehnung der SV-A.

Wird kein Zusammenhang oder ein präziser funktionaler Zusammenhang behauptet, führen die psychologische Hypothese und die psychologische Vorhersage üblicherweise zu einer oder zu $S > 1$ ungerichteten $H_{0,s}$ ($s = 1, 2, \ldots, S$). Bei SV-B $\Leftrightarrow$ H_0 gilt: $\varepsilon = \beta$ und $\varphi = \alpha$. Ein α-Fehler führt mit der Wahrscheinlichkeit φ zu einer falschen Annahme der SV-B, und ein

β-Fehler führt mit der Wahrscheinlichkeit ε zu einer falschen Ablehnung der SV-B.

Bei der Simultankontrolle von α und β kann α auf die gewohnte Weise auf einen kleinen Wert α_{krit} festgelegt werden, während man zur Kontrolle von β zusätzlich eine *Effektgröße* (*EG*) benötigt. Kann man eine solche für einen vorgesehenen statistischen Test identifizieren oder definieren, wird es möglich, eine *Testplanung* (oder Teststärkenanalyse) durchzuführen, das heißt, einen Ausgleich zwischen den vier Determinanten eines statistischen Tests unter der Theorie von Neyman und Pearson (Neyman, 1942) herbeizuführen, also zwischen der Effektgröße (EG), den statistischen Fehlerwahrscheinlichkeiten α und β und dem Stichprobenumfang n (N). Hierzu werden spezielle Formeln und/oder Tabellen benötigt, die man etwa in Cohen (1988), in Hager (2004) und in Kraemer und Thiemann (1987) findet; auch gibt es Computerprogramme zur Durchführung von Teststärkenanalysen (etwa Erdfelder et al., 1996). Die Testplanung bezieht sich immer auf eine H_1. Sie findet stets auf der Ebene ST der testbaren statistischen Hypothesen und der statistischen Tests statt. Sie kann bei festliegendem Stichprobenumfang auch dazu genutzt werden, die eine der anderen Größen (*EG*, α oder β) zu berechnen. Im Zuge der Testplanung kommt es zu »methodisch-technischen« Festlegungen, die weitere Vorhersagen implizieren können. Zum Beispiel kann man bei der Vorhersage eines strikt monotonen Trends über K = 3 Versuchsbedingungen die Kriteriumseffektgrößen $\delta_{B,t}$ auf die gleichen Werte festlegen; dies bedeutet dann Gleichabständigkeit der Mittelwerte. Diese Tatsache könnte im Sinne der Vorhersage eines linearen Trends interpretiert werden, doch wäre diese Interpretation falsch, weil die unabhängige Variable qualitativ und nicht quantitativ ist. Sich auf der Ebene der statistischen Tests ergebende Folgerungen aus den »methodisch-technischen« Festlegungen sind also weder explizit aufzustellen noch zu prüfen; zu prüfen sind einzig und allein die auf den vorgeordneten Ebenen der psychologischen Vorhersage und der statistischen Vorhersage abgeleiteten Vorhersagen, die von den Festlegungen auf der Ebene der statistischen Tests unbeeinflusst bleiben.

Da mehrheitlich gerichtete statistische Hypothesen aus psychologischen Hypothesen und Vorhersagen abgeleitet werden können, sollten auch statistische Tests gewählt werden, die die Richtung berücksichtigen, also etwa *z*- und *t*-Tests. Demgegenüber testen die sehr häufig anzutreffenden χ^2- und *F*-Tests ungerichtete statistische Hypothesen über Wahrscheinlichkeiten π oder über Erwartungswerte μ_k. Diese Teststatistiken kommen unter bestimmten Bedingungen also in Frage, wenn die psychologische Hypothese, die psychologische und die statistische Vorhersage ungerichtet sind.

Bei allen Ableitungen von mehreren testbaren Einzelhypothesen aus einer statistischen Vorhersage *kumulieren* die statistischen Fehlerwahrscheinlichkeiten α und β in Abhängigkeit von der Anzahl, der Art der statistischen Hypothesen (H_0, H_1) und ihrer Verknüpfung pro statistischer Vorhersage und erhöhen ε und/oder φ auf der Ebene der statistischen Vorhersage. Die Fehlerwahrscheinlichkeiten sollten dann entsprechend pro statistischer Vorhersage adjustiert werden. Dazu werden ε_{krit} und φ_{krit} vor dem Versuch in der gleichen Weise festgelegt wie sonst α_{krit} und β_{krit} bei einzelnen Tests. Zur Adjustierung wird/werden ε_{krit} und/oder φ_{krit} durch die Anzahl der Einzelhypothesen dividiert (Bonferroni-Adjustierung; vgl. Westermann u. Hager, 1986), woraus sich die adjustierten Werte für α_{krit} und/oder β_{krit} ergeben, auf denen die Testplanung beruht.

Der statistische Test in der Neyman-Pearson-Tradition fungiert als Hilfsmittel zur Prüfung von psychologischen Hypothesen, dient als Entscheidungsregel und wird als Hilfstheorie ohne empirischen Gehalt interpretiert, für den also die Voraussetzungen als empirisch nicht prüfbar angesehen werden und dem bedarfsweise die Voraussetzungen zugeschrieben werden, die für seine Durchführung erforderlich sind (Westermann, 1987). Auch ist es möglich, die gängigen parametrischen Tests als Permutationstest zu interpretieren (Bredenkamp, 1980, S. 17–23), um ihre restriktiven Voraussetzungen zu vermeiden; dass dies möglich ist, konnte für parametrische Tests wie dem *t*- und dem *F*-Test der Varianzanalysen gezeigt werden. Damit man mit Permutationstests die üblichen Mittelwertshypothesen prüfen kann, muss ihnen Symmetrie oder Homomerität der zugrunde liegenden Verteilungen zugeschrieben werden – eine relativ schwache Annahme. Zusätzlich muss für jeden statistischen Test die Bedingung erfüllt sein, dass die empirischen Daten voneinander unabhängig sind. Die statistischen Hypothesen beziehen sich nur auf den durchzuführenden Versuch; Schlüsse auf reale oder fiktive Populationen von Daten oder Datentragende werden nicht angestrebt.

Betrachten wir zwei Beispiele für die Beziehungen zwischen den Fehlerwahrscheinlichkeiten und den Risiken, und zwar der Einfachheit halber unter Außerachtlassung der psychologischen Vorhersage.

Beispiel 1: Die psychologische Hypothese behauptet einen Unterschied und sie ist »zutreffend«. Man leitet aus ihr für eine gewählte Versuchssituation und für die beobachtbare abhängige Variable das Zutreffen einer H_1 ab: PH $\approx>$ PV $\approx>$ SV $\Leftrightarrow$ H_1 (»A $\approx>$ B«: »B wird als vorhersagekonform aus A abgeleitet«). Auf Grund der Daten wird jedoch die vorhersagekonträre H_0 angenommen (β-Fehler), und das heißt, die statistische Vorhersage wird abgelehnt. Dann entscheidet man *fälschlich* für die Nicht-Bewährung der psy-

chologischen Hypothese. Das Risiko dieser falschen Entscheidung über die psychologische Hypothese wird durch die Irrtumswahrscheinlichkeit β maßgeblich mit beeinflusst, sodass gilt: f = fkt(φ) und φ = β (fkt: »ist eine nicht weiter spezifizierbare, aber mutmaßlich monotone Funktion von«).

Beispiel 2: Es wird eine neue, das heißt nicht-paradigmatische beobachtbare abhängige Variable gewählt, um versuchsweise den Anwendungsbereich der psychologischen Hypothese auszuweiten. Diese abhängige Variable gehört aber nicht zum Anwendungsbereich der psychologischen Hypothese, die daher für das Experiment nicht gilt. Unter der Annahme des Zutreffens der psychologischen Hypothese wird erneut eine statistische H_1 als vorhersagekonform abgeleitet. Irrtümlich wird die vorhersagekonträre H_0 beibehalten (β-Fehler). Daraus wird gefolgert, dass die psychologische Hypothese sich nicht bewährt hat. Diese richtige Entscheidung über die psychologische Hypothese kommt aber auf Grund einer Fehlentscheidung auf der statistischen Ebene zustande: $1 - e$ = fkt(φ) und φ = β.

Bei jedem Übergang von einer Ebene zur nächsten ist die *Ableitungsvalidität* zu beachten, die gewährleisten soll, dass die Ableitung nicht »*zu* locker« erfolgt und dass von der Aussage der psychologischen Hypothese so viel auf die jeweils nächstniedrige Ebene übertragen wird, wie dies im konkreten Fall möglich ist. Jeder Übergang bedeutet aber eine Einschränkung, zum Beispiel, weil die Konkretisierungen auf der Ebene der psychologischen Vorhersage immer eingeschränkter sind als die theoretischen Begriffe der psychologischen Hypothese und weil statistische Kennwerte und Zufallsvariable etwas ganz anderes sind als psychologische Variablen, Begriffe oder Konstrukte. Die Ableitungsvalidität setzt sich aus den Kriterien Adäquatheit und Erschöpfendheit zusammen. Die *Adäquatheit* bezieht sich dabei auf die Richtung der Beziehungsaussage der psychologischen Hypothese. Adäquat ist eine Ableitung dann, wenn aus einer gerichteten psychologischen Hypothese eine gerichtete psychologische Vorhersage, eine gerichtete statistische Vorhersage und eine (oder mehrere) gerichtete testbare statistische Hypothese(n) abgeleitet werden; für ungerichtete psychologische Hypothesen gilt sinngemäß dasselbe. Wird – was häufig der Fall ist – aus einer gerichteten psychologischen Hypothese über die gerichtete psychologische Vorhersage eine ungerichtete statistische Hypothese abgeleitet, ist die Adäquatheit verletzt. Zuweilen führt eine gewählte Operationalisierung auch dazu, dass sich die Richtung der psychologischen Hypothese umkehrt. Prüft man die *PH-1:* »Mit der Dauer des Schlafentzugs (beobachtbare UV) sinkt die Konzentrationsleistung (nicht-beobachtbare AV)« über die Fehlerzahlen in einem Konzentrationstest (empirische AV), die mit der Dauer des Schlafentzugs ansteigen sollen, wird die Gerichtetheit

der PH-1 auf der Ebene der PV-1 beibehalten; die Ableitung ist soweit adäquat.

Die *Erschöpfendheit* bezieht sich auf die Menge an Informationen, die in der psychologischen Hypothese enthalten sind. Dieses Kriterium ist verletzt, wenn eine der Ebenen »unterhalb« der psychologischen Hypothese nicht alle Behauptungen der psychologischen Vorhersage enthält. Daher sind bei der Formulierung der psychologischen Vorhersage so viele Einzelvorhersagen wie erforderlich abzuleiten, und die Einzelvorhersagen sowie die daraus abgeleiteten statistischen Hypothesen sind zur Sicherung der Erschöpfendheit stets dann konjunktiv zusammenzufassen, wenn sie sich auf *eine* Prüfinstanz beziehen. Mit *Prüfinstanz* werden die minimalen formalen Anfangsbedingungen bezeichnet, die ein Versuchsplan erfüllen muss, um die Prüfung einer psychologischen Hypothese zu ermöglichen. Ein Versuchsplan muss aus einer minimalen Anzahl (min) von unabhängigen Variablen bestehen ($UV_{min} = 1$ bei einfaktoriellen und $UV_{min} = 2$ bei zweifaktoriellen sowie $AV_{min} = 1$ bei univariaten PHn) und aus einer Minimalzahl K_{min} von Stufen für diese unabhängigen Variablen. Bei den meisten einfaktoriellen psychologischen Hypothesen ist $K_{min} = 2$; bei psychologischen Hypothesen über einen bitonen Trend und bei quantitativen psychologischen Hypothesen über einen linearen oder quadratischen quantitativen Zusammenhang ist $K_{min} = 3$ und $(JK)_{min} = 4$ bei zweifaktoriellen Interaktionshypothesen und so weiter. In seltenen Fällen gilt $K = 1$, beispielsweise bei der Prüfung der *PH-2*: »Im Mittel haben Psychologie-Studierende (beobachtbare organismische UV) eine überdurchschnittliche Intelligenz (nicht-beobachtbare AV)«. Die PV-2 besagt: »Psychologie-Studierende haben in einem Intelligenztest im Mittel einen IQ von mindestens $\mu_0 = 116$«. Die PV-2($K = 1$) legt also auch den unverzichtbaren Vergleichswert als untere Grenze μ_0 fest. Aus ihr wird die SV-2($K = 1$) $\Leftrightarrow$ *ST*-2($K = 1$): $H_{0,1}$: $\mu - \mu_0 \geq 0$ eines einseitigen Ein-Stichproben-*t*-Tests abgeleitet, der auf der empirischen Streuung s beruht ($H_{0,1}$: Nullhypothese Nr. 1). Die Effektgröße lautet: $\delta_{B,1} = (\mu - \mu_0)/\sigma$ (Cohens *d*, 1988, S. 20; σ: theoretische Binnenstreuung) und ihre empirische Entsprechung $d_{B,1} = (M - \mu_0)/s$ (M: empirischer Mittelwert; s: empirische Streuung). Die Stichprobengröße bestimmt sich nach:

$$(1)\ n = (z_{1-\alpha} + z_{1-\beta})^2/\delta^2_{B,krit},$$

wobei $z_{1-\alpha}$ und $z_{1-\beta}$ die z-Werte der Normalverteilung für die vorgewählten Fehlerwahrscheinlichkeiten α_{krit} und β_{krit} bezeichnen. Die Effektgröße $\delta_{B,krit,1}$ ist hier als Höchsteffekt zu interpretieren, den man als Bereich unter der $H_{0,1}$ in der vorhersagekonträren Richtung noch zu akzeptieren bereit ist. Dieser

Effekt sollte wie bei allen abgeleiteten $H_{0,s}$ klein gewählt werden (etwa $0{,}20 \geq \delta_{B,krit,1} \geq 0{,}40$), und je geringer er ausfällt, desto mehr Versuchspersonen werden benötigt. Er ist die Grundlage der Testplanung für die *vorhersagekonträre* $H_{1,1}$: $\mu - \mu_0 < 0$. Unter der $H_{0,1}$ können kleine oder große Effekte auftreten, deren Größe erst nach dem Versuch bestimmbar ist. Die A-priori-Festlegung eines für bedeutsam erachteten Werts für die Effektgröße unter einer abgeleiteten H_1 beziehungsweise eines Höchsteffekts unter einer abgeleiteten H_0 ist oft nicht einfach, aber dennoch unverzichtbar; man kann sich dabei auf ähnliche Vorläuferuntersuchungen beziehen oder auch auf Cohens (1988) Konventionen (zu weiteren Möglichkeiten siehe Hager, 2004). Bereits 1972 hat Bredenkamp (S. 72–73) darauf hingewiesen, welche unerwünschten Konsequenzen es nach sich zieht, wenn man auf diese Festlegung verzichtet (vgl. auch Bredenkamp, 1980, S. 22; Cohen, 1988, S. 285).

Prüfinstanzen selbst können konjunktiv (»∧«) miteinander verknüpft werden, das heißt, alle Tests müssen vorhersagekonform ausfallen – dies erhöht die e-Validität –, oder disjunktiv (»∨«), das heißt, der Test oder die Tests mindestens einer Prüfinstanz muss/müssen vorhersagekonform ausgehen – dies erhöht die f-Validität. Die Art der Verknüpfung von Prüfinstanzen legt eines der Entscheidungskriterien fest. Wenn man einen zusätzlichen Faktor A mit J Stufen etwa zur Kontrolle einer bekannten Störvariablen einführt, können Prüfinstanzen auf der Ebene der Zellen AB_{jk} definiert werden, sodass sich die Vorhersagen auf die Zellenebene beziehen. Die Testplanung stellt sicher, dass die Vpn-Anzahl n_{jk} pro Zelle zum Nachweis des festgelegten Kriteriumswerts für die Effektgröße, EG_{krit}, bei α_{krit} und $1 - \beta_{krit}$ genügend groß ist; die Ebene des Haupteffekts B ist dann ohne Belang. Bei $K > K_{min} = 2$ ergeben sich $J(K - 1)$ Prüfinstanzen über benachbarte Faktorstufen des Faktors B beziehungsweise unter Einbezug der nicht-benachbarten Stufen $J[K(K - 1)]/2$ Prüfinstanzen. Fasst man die Prüfinstanzen unter B bei $K > K_{min}$ testbedingt konjunktiv oder disjunktiv zu einer zusammen, so resultieren J Prüfinstanzen. Die testbedingte Zusammenfassung von Prüfinstanzen ist *nur* bei abgeleiteten statistischen Hypothesen möglich, die adäquat und erschöpfend mit einem Globaltest wie der Varianzanalyse getestet werden können. Die konjunktive Zusammenfassung der Prüfinstanzen erhöht die e-Validität, die disjunktive die f-Validität. Alternativ kann man die Prüfinstanzen auf der Ebene des Haupteffekts B festlegen und die Testplanung auf diese Ebene beziehen; die Zellen dienen bei varianzanalytischer Sichtweise nur der Bestimmung der Testvarianz. Die Prüfung auf der Zellenebene ist bei erfolgter Testplanung e-valider als die alternative Prüfung auf der Ebene des Haupteffekts B, weil auf der Zellenebene mehr Prüfinstanzen definiert werden können.

Ableitungen

Einfaktorielle psychologische Hypothesen nehmen nur auf eine theoretische oder empirische unabhängige Variable und eine theoretische oder empirische abhängige Variable Bezug. Die Duale-Kode-Theorie (DKT; Paivio, 1971; vgl. auch Wippich u. Bredenkamp, 1979) postuliert zwei unabhängig voneinander arbeitende Kodiersysteme. Das verbale Kodiersystem ist für die serielle oder sequenzielle Verarbeitung von Informationen zuständig. Das nonverbale oder imaginale System ist auf die räumlich-parallele Verarbeitung von Informationen spezialisiert; es soll Informationen analog-synchron repräsentieren. Eine zentrale Annahme der Duale-Kode-Theorie besagt dann, dass man die beiden Kodiersysteme durch unterschiedlich bildhaftes Lernmaterial in unterschiedlicher Weise aktivieren kann. Aufgrund dieser Annahme kann die folgende typische qualitative psychologischen Hypothesen über theoretische Begriffe aus der Duale-Kode-Theorie abgeleitet werden: *DKT* ≈> *PH-3:* »Wenn sehr bildhaftes Material gelernt wird, dann erhöht dies im Mittel die Gedächtnisleistung im Vergleich zum Lernen von wenig bildhaftem Material« (in Anlehnung an Erdfelder u. Bredenkamp, 1994, S. 606; die Probabilisierung wurde hinzugefügt). Die PH-3 benennt die theoretische unabhängige Variable, das sehr bildhafte Material, deren Konkretisierung zu variieren ist, und die theoretische abhängige Variable, die Gedächtnisleistung, deren Konkretisierung zu erfassen ist. Um die nicht-beobachtbare unabhängige Variable Bildhaftigkeit konkretisieren zu können, werden in einem Vorversuch mittels einer siebenstufigen Ratingskala Wörter und/oder Bilder entweder als »sehr bildhaft« (»erzeugt leicht bildliche Vorstellungen«) eingeschätzt oder als »wenig bildhaft« (»erzeugt wenig oder gar keine bildlichen Vorstellungen«); diese Einschätzungen stellen die Konkretisierung der unabhängigen Variablen Bildhaftigkeit dar. Aus den Einschätzungen werden $K = 2$ Wortlisten mit hoher und niedriger eingeschätzter mittlerer Bildhaftigkeit konstruiert, die die Variation der beobachtbaren unabhängigen Variablen »eingeschätzte Bildhaftigkeit« (B) darstellen. Die Hilfshypothese mit empirischem Gehalt, die diesem Vorgehen zugrunde liegt, besagt, dass die unterschiedlichen Einschätzungen in interindividuell einheitlicher Weise unterschiedliche Ausprägungen der Bildhaftigkeit repräsentieren. Bei der Konstruktion der Wortlisten werden bekannte konfundierende Variable wie die Bedeutungshaltigkeit oder die Wortlänge durch Konstanthaltung kontrolliert; gegenüber unbekannten potenziellen Störfaktoren wird randomisiert. Die Konkretisierung der abhängigen Variablen Gedächtnisleistung erfolgt über die »Anzahl der richtig reproduzierten

Wörter«, wobei die Hilfshypothese ohne empirischen Gehalt herangezogen wird, dass eine höhere Anzahl richtig reproduzierter Wörter einer besseren Gedächtnisleistung entspricht. Diese Konkretisierung wird von der in diesem inhaltlichen Gebiet arbeitenden Forschergemeinschaft als eine der Möglichkeiten angesehen, die Gedächtnisleistung zu erfassen. Eine andere Möglichkeit besteht im richtigen Wiedererkennen von Wörtern (vgl. Wippich u. Bredenkamp, 1979). Es handelt sich um paradigmatische, das heißt typische Konkretisierungen, die bereits in den ersten Prüfungen der Duale-Kode-Theorie (Paivio, 1971) Verwendung fanden, die plausibel und die allgemein akzeptiert sind.

Bei der PH-3 besteht eine Prüfinstanz aus der empirischen unabhängigen Variablen mit $K = 2$ Ausprägungen und der empirischen abhängigen Variablen. Sie behauptet eine gerichtete Beziehung zwischen der theoretischen unabhängigen Variablen und der theoretischen abhängigen Variablen; für die Konkretisierungen und den gewählten Versuchsplan mit $K = 2$ Faktorstufen müssen aus ihr also gerichtete Vorhersagen abgeleitet werden. Die PV-3 lautet daher: »Die mittlere Anzahl richtig reproduzierter Wörter (MWR) ist unter einer Liste mit hoch bildhaften Wörtern (mittlerer eingeschätzter Bildhaftigkeitswert: $B_1 = 6{,}3$) höher als unter einer Liste mit niedrig bildhaften Wörtern (mittlerer eingeschätzter Bildhaftigkeitswert: $B_2 = 2{,}5$)«. Wenn im Mittel mehr hoch bildhafte als wenig bildhafte Wörter richtig reproduziert werden sollen, dann heißt das auch, dass – statistisch gesehen – die erwartete (E) Anzahl von richtig reproduzierten bildhaften Wörtern höher sein soll als die erwartete Anzahl von richtig reproduzierten wenig bildhaften Wörtern [E(WR)]. Im nächsten Schritt werden die Reproduktionsleistungen pro Person als Realisierungen einer Zufallsvariablen interpretiert, und über diese Zufallsvariable wird eine weitere definiert, die die Realisierungen der ersten Zufallsvariablen pro Versuchsgruppe zu einem Wert zusammenfasst, dem Mittelwert M_k. Seine theoretische Entsprechung ist m_k. Man erhält also (vgl. Tab. 1):

Tabelle 1: Ebenen der Vorhersage bei der Prüfung der PH-3

Vorhersagenebene	UV: eingeschätzte Bildhaftigkeit (beobachtbar) Stufe B_1	V	Stufe B_2	Bemerkungen
PV-3: UV B	$B_1 = 2{,}5$	(<)	$B_2 = 6{,}3$	Ebene der PV: Operationalisierung der UV
PV-3: AV MWR	MWR_1	<	MWR_2	Ebene der PV: Operationalisierung der AV
SV-3: E(WR)	$E(WR)_1$	<	$E(WR)_2$	Ebene der SV: Übersetzung in statistische
μ_k	μ_1	<	μ_2	Begriffe; hier: Erwartungswerte
ST-3: $H_{1,2}$: μ_k	μ_1	<	μ_2	Ebene der ST; ggf. Zerlegung der SV

Anmerkung: »V« steht für »Vorhersage«.

(2) DKT $\approx>$ PH-3 $\approx>$ PV-3: $(MWR_1 < MWR_2) \approx>$ SV-3: $(\mu_1 < \mu_2) \Leftrightarrow$ ST-3: $(H_{1,2}: \mu_2 - \mu_1 > 0)$.

Bei der $H_{1,2}$ handelt sich um die gerichtete Alternative eines einseitigen t-Tests. Ein F-Test wäre hier nicht adäquat, weil er zwei ungerichtete Mittelwertshypothesen gegeneinander testet.

Die Effektgröße ist als standardisierter Mittelwertsabstand $\delta_B = (\mu_2 - \mu_1)/\sigma$ definiert [Stichprobenentsprechung: $d_B = (M_2 - M_1)/s$] und betrifft den erwarteten Mindesteffekt in der vorhersagekonformen Richtung. Dieser kann relativ groß gewählt werden ($\delta_{B,krit,2} = 0{,}8$ oder größer), sodass die nach

(3) $n_k = 2(z_{1-\alpha} + z_{1-\beta})^2/\delta^2_{B,krit}$

bestimmte Anzahl an Versuchspersonen vergleichsweise gering ist (siehe zur Ableitung dieser Formel Bredenkamp, 1969; vgl. auch die zugehörige allgemeine Formel in Erdfelder u. Bredenkamp, 1994, S. 629). Je größer der Kriteriumswert EG_{krit} gewählt wird, desto geringer ist unter sonst gleichen Bedingungen der Vpn-Bedarf oder desto höher ist die Teststärke $1 - \beta$.

Die unabhängige Variable »eingeschätzte Bildhaftigkeit« kann in mehr als $K = 2$ Stufen variiert werden. Bei einer Prüfung der PH-3 mit $K = 3$ ($B_1 = 2{,}5$; $B_2 = 4{,}3$ und $B_3 = 6{,}3$) folgt ein strikt monotoner Trend über benachbarte Faktorstufen B_k:

(4) DKT $\approx>$ PH-3 $\approx>$ PV-3($K = 3$): $(MWR_1 < MWR_2 < MWR_3) \approx>$
$\approx>$ SV-3($K = 3$): $(\mu_1 < \mu_2 < \mu_3) \approx>$ ST-3($K = 3$): $[(H_{1,3}: \mu_2 - \mu_1 > 0) \wedge (H_{1,4}: \mu_3 - \mu_2 > 0)]$.

$R = K - 1 = 2$ Prüfinstanzen sind ausreichend, weil aus $A(H_{1,3} \wedge H_{1,4})$ die $AH_{1,5}$: $(\mu_3 > \mu_1)$ (AH: Annahme einer statistischen Hypothese) über die nicht-benachbarten Faktorstufen folgt, und bei $A(H_{0,3} \vee H_{0,4})$ wird die SV-3($K = 3$) ungeachtet des Status' der $H_{1,5}$ ohnehin abgelehnt. Bei konjunktiver Verknüpfung der abgeleiteten Alternativhypothesen gilt bzgl. der Kumulierung: $\varepsilon \leq \max(\alpha_r)$ und $\varphi \leq \Sigma\beta_r$, und die Adjustierung lautet entsprechend: $\alpha_{r,krit} = \varepsilon_{krit}$ und $\beta_{r,krit} = \varphi_{krit}/R$.

Eine Varianzanalyse wäre hier nicht adäquat, weil sie die vorhergesagten Richtungen außer Acht lässt. Monotone-Trend-Tests wären zwar adäquat, aber nicht erschöpfend, da sie bereits bei *mindestens einer* vorhersagekonformen Rangordnung signifikant werden können, die nicht über benachbarte Faktorstufen vorliegen muss wie vorhergesagt; dabei können sogar Ranginversionen vorliegen (Hager, 2004).

Zur Bestimmung des benötigten Stichprobenumfangs pro Faktorstufe B_k findet für Paarkontraste erneut die Formel (3) Anwendung; die empirische Entsprechung von $\delta_{B,t}$ lautet: $d_{B,t} = (M_k - M_{k'})/s$. Für die abgeleiteten $R = 2$ $H_{1,r}$ sollten die *E*Gn als $\delta_{B,krit,3} = \delta_{B,krit,4} = \delta_{B,krit,2}/2$ festgelegt werden.

Bei disjunktiver Verknüpfung wird auch über die nicht-benachbarten Faktorstufen (B_1 und B_3) eine Prüfinstanz vorgesehen, sodass $R = K(K - 1)/2 = 3$ Prüfinstanzen resultieren: SV-3($K = 3$) ≈> ST-3($K = 3$): ($H_{1,3} \vee H_{1,4} \vee H_{1,5}$), wobei $H_{1,5} \Leftrightarrow H_{1,2}$. Bei disjunktiver Verbindung der R $H_{1,r}$ kumuliert α: $\varepsilon \leq \Sigma\alpha_r$ und $\varphi = \max(\beta_r)$ und wird folglich adjustiert: $\alpha_{krit,r} = \varepsilon_{krit}/R$ und $\beta_{krit,r} = \varphi_{krit}$.

Die Paarkontraste sind nonorthogonal. Die Festlegung des Kriteriumswerts der Effektgröße, $EG_{krit,r}$, erfolgt für jede $H_{1,r}$ separat, wobei lediglich berücksichtigt wird, ob es sich um eine Festlegung für benachbarte oder nicht-benachbarte Faktorstufen handelt. Der sich aus den einzelnen Festlegungen ergebende Gesamteffekt über alle *K* Faktorstufen bleibt unberücksichtigt. Erdfelder u. Bredenkamp (1994, S. 629–630) legen dagegen die Einzeleffekte nur unter Beachtung des Gesamteffekts fest. Zur Berechnung des Gesamteffekts $\varphi^2_{VA,G}$ siehe Erdfelder u. Bredenkamp (1994, S. 630).

Man kann mit der »eingeschätzten Abstraktheit-Konkretheit« einen zweiten Faktor A mit *J* Stufen einführen, um auf diese Weise einen mit der »eingeschätzten Bildhaftigkeit« bekanntermaßen konfundierten Faktor zu kontrollieren. Dadurch können Einzelvorhersagen auf der Zellenebene AB_{jk} abgeleitet werden. Für $J = 2$ Stufen von A und $K = 3$ Stufen von B ergibt sich:

(5) DKT ≈> PH-3 ≈> PV-3($J = 2$; $K = 3$): ($MWR_{11} < MWR_{12} < MWR_{13}$) ∧ ($MWR_{21} < MWR_{22} < MWR_{23}$) ≈> SV-3($J = 2$;$K = 3$): [($\mu_{11} < \mu_{12} \approx< \mu_{13}$) ∧ ($\mu_{21} < \mu_{22} \approx< \mu_{23}$)] ≈> ST-3($J = 2$;$K = 3$): [($H_{1,6}$: $\mu_{12} - \mu_{11} > 0$) ∧ ($H_{1,7}$: $\mu_{13} - \mu_{12} > 0$) ∧ ($H_{1,8}$: $\mu_{22} - \mu_{21} > 0$) ∧ ($H_{1,9}$: $\mu_{23} - \mu_{22} > 0$)].

Analoge Hypothesen resultieren, wenn man statt dessen *alternativ* den Haupteffekt B als Prüfinstanz wählt. Da die Einführung eines zweiten Faktors in den Versuchsplan in aller Regel die Präzision erhöht und die Testvarianz verringert wird, sollte man diese Präzisionserhöhung auch bei der Testplanung berücksichtigen. Dazu wird ein Präzisionsindex Π_{BA} wie folgt definiert:

(6) $\Pi_{BA} = \sigma^2_{I,BA}/\sigma^2_{I,B}$,

wobei $\sigma^2_{I,BA}$ die Testvarianz im zweifaktoriellen Plan bezeichnet und $\sigma^2_{I,B}$ die Testvarianz im einfaktoriellen Plan. Es gilt: $0 < \Pi_{BA} \leq 1$, und je geringer dieser

Präzisionsindex ausfällt, desto größer ist der Präzisionsgewinn. Für zweifaktorielle Pläne können a priori Werte zwischen 0,50 und 0,90 angesetzt werden. Dann resultiert für die Effektgröße $\delta_{BA,B}$ im zweifaktoriellen Plan:

(7) $\delta^2_{BA,B} = \delta^2_B/\pi_{BA}$.

Hierdurch ergibt sich eine Vergrößerung der Effektgröße, sodass vergleichsweise weniger Versuchspersonen zu ihrem Nachweis erforderlich sind. Auf die gleiche Weise kann im Übrigen auch bei höherfaktoriellen Varianzanalysen verfahren werden.

Sofern sich die folgenden einfaktoriellen qualitativen psychologischen Hypothesen auf Begriffe der Duale-Kode-Theorie beziehen, stellen sie lediglich *Typen* oder *Aussageformen* für psychologische Hypothesen dar, bei denen die Frage außer Acht bleibt, ob diese Hypothesen ernsthaft von jemanden vertreten werden.

PH-4: »Bei imaginaler Kodierung *verändert* sich im Mittel die Gedächtnisleistung im Vergleich zur verbalen Kodierung«. Die Hypothese ist ungerichtet; es besteht also *irgendeine* Beziehung zwischen unabhängiger und abhängiger Variablen. Für einen Zwei-Gruppen-Plan gelangt man zu:

(8) DKT: PH-4 $\approx>$ PV-4: $(MWR_1 \neq MWR_2) \approx>$ SV-4: $(\mu_1 \neq \mu_2) \Leftrightarrow$ ST-4: $(H_{1,10}: \mu_1 - \mu_2 \neq 0)$.

Die ungerichtete $H_{1,10}$ über einen Paarkontrast wird mittels zweiseitigem *t*- oder mittels einseitigem *F*-Test getestet. Die Strenge der Prüfbarkeit der PH-4 ist geringer als die der PH-3.

Für die Testplanung wird die Effektgröße als Mindesteffekt $EG_{krit} = \pm\delta_{B,krit,10}$ für den *t*-Test in beiden Richtungen spezifiziert; die Anzahl der Versuchspersonen wird nach Formel

(9) $n_k = 2(z_{1-\alpha/2} + z_{1-\beta})^2/\delta^2_{B,krit}$

für einen zweiseitigen *t*-Test ermittelt. Der Kriteriumswert für die Effektgröße $\delta_{B,krit,10}$ kann groß gewählt werden, zum Beispiel: $\delta_{B,krit,10} \geq |\pm 0{,}8|$. Bei der Prüfung der PH-4 über die PV-4($K = 3$) werden $R = K(K-1)/2 = 3$ $H_{1,r}$ abgeleitet:

(10) PV-4($K = 3$): $(MWR_1 \neq MWR_2 \neq MWR_3) \approx>$ SV-4($K = 3$): $(\mu_k \neq \mu_{k'}$ für alle k, k') $\approx>$ ST- 4($K = 3$): $[(H_{1,11}: \mu_1 - \mu_2 \neq 0) \wedge (H_{1,12}: \mu_2 - \mu_3 \neq 0) \wedge (H_{1,13}: \mu_1 - \mu_3 \neq 0)]$.

Ist die $H_{1,13} \Leftrightarrow H_{1,10}$ erforderlich? Aus $AH_{1,11} \wedge AH_{1,12}$ folgt nicht auch $AH_{1,13}$, sondern die Mittelwertskonfiguration könnte auch so sein, dass die $H_{0,13}$ angenommen wird ($AH_{0,13}$). In Bezug auf die getesteten statistischen Hypothesen folgt aus $AH_{1,11} \wedge AH_{1,12}$ zwar $\mu_1 \neq \mu_2$ und $\mu_2 \neq \mu_3$, aber nicht auch zwingend $\mu_1 \neq \mu_3$. Damit kann auf die $H_{1,13}$ nicht verzichtet werden. Für einen einfaktoriellen Plan mit K Versuchsbedingungen geben Erdfelder und Bredenkamp (1994, S. 629) die folgende Effektgröße an, bei der der Index »VA« für Varianzanalyse steht:

$$(11)\ \Phi^2_{VA} = \frac{(\Sigma c_{k,t}\mu_k)^2/(K\Sigma c^2_{k,t})}{\sigma^2_{I,B}} = \frac{\delta^2_{B,t}}{K\Sigma c^2_{k,t}}.$$

$c_{k,t}$ sind Kontrastkoeffizienten, deren Summe Null ergibt und die nicht alle gleich Null sein dürfen. Nach Cohen (1988, S. 276) gilt:

$$(12)\ \Phi^2_{VA} = \delta^2_{B,t} / 4.$$

Ferner geben Erdfelder und Bredenkamp (1994, S. 630) die folgende Formel für die Stichprobengrößenbestimmung für Kontraste im Rahmen eines einfaktoriellen Versuchsplans mit $K \geq 3$ Versuchsbedingungen an:

$$(13)\ n = \frac{z_{1-\alpha} + z_{1-\beta})^2}{K\Phi^2_{VA}}.$$

Einsetzen führt dann zu:

$$(14)\ n = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{K\delta^2_{B,t}/(K\Sigma c^2_{k,t})} = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\delta^2_{B,t}/(K\Sigma c^2_{k,t})} = \frac{z_{1-\alpha} + z_{1-\beta})^2\Sigma c^2_{k,t}}{\delta^2_{B,t}}$$

Die Kriteriumswerte für die Effektgrößen, $EG_{krit,r}$, sollten für die *t*-Tests nach $\delta_{B,krit,11} = \delta_{B,krit,12} = \delta_{B,krit,13}/2$ festgelegt werden. Um bei ungleichen Effekten für die einzelnen Prüfinstanzen gleiche Stichprobenumfänge für alle Faktorstufen zu erreichen, wird man unter Berücksichtigung der Kumulation die Fehlerwahrscheinlichkeiten $\alpha_{krit,r}$ und $\beta_{krit,r}$ entsprechend unterschiedlich wählen. Die Adjustierung erfolgt wie bei der PH-3. – Eine Varianzanalyse wäre hier nicht erschöpfend, weil drei Ungleichheiten vorhergesagt werden, während die Varianzanalyse bereits bei nur einer Ungleichheit signifikant werden kann. Sie führt daher bei einer disjunktiven Verknüpfung der Prüfinstanzen zu einer adäquaten und erschöpfenden Prüfung der PH-4:

(15) SV-4($K = 3$): $[(\mu_1 \neq \mu_2) \vee (\mu_2 \neq \mu_3) \vee (\mu_1 \neq \mu_3)] \approx>$ ST- 4($K = 3$): [$H_{1,14}$ der Varianzanalyse: ($\mu_k \neq \mu_{k'}$ für *mindestens ein Paar* k, k' mit $k \neq k'$)].

Bei der Varianzanalyse ist die Effektgröße $\varphi_{VA,B}$ wie folgt definiert (Cohen, 1988, S. 281):

(16) $\varphi^2_{VA,B} = \sigma^2_B/\sigma^2_{I,B}$,

wobei σ^2_B die Treatment- und $\sigma^2_{I,B}$ die Binnenvarianz bezeichnet. Diese Effektgröße steht in einer einfachen Beziehung zum multiplen Korrelationsquadrat $\eta^2_{Y.B}$ (Cohen, 1988, S. 281):

(17) $\eta^2_{Y.B} = \varphi^2_{VA,B}/(1 + \varphi^2_{VA,B})$.

Im Beispielfall kann ein relativ großer Wert für diese Effektgröße gewählt werden, also etwa 0,35 bis 0,40. – Alternativ kann man die $R = K(K-1)/2 = 3$ disjunktiv verknüpften Prüfinstanzen auch einzeln testen und dadurch die e-Validität im Vergleich zur Varianzanalyse erhöhen (wegen der α-Adjustierung wird es den Einzelhypothesen erschwert, angenommen zu werden):

(18) SV-4($K = 3$) $\approx>$ ST-4($K = 3$): $(H_{1,15} \vee H_{1,16} \vee H_{1,17})$.

Bei disjunktiver Zusammenfassung der $H_{1,r}$ gilt für die Kumulation: $\varepsilon \leq \Sigma\alpha_r$ und $\varphi \leq \max(\beta_r)$ sowie für die Adjustierung $\alpha_{krit} = \varepsilon_{krit}/R$ und $\beta_{krit,r} = \varphi_{krit}$.

Bei zweifaktoriellen Varianzanalysen wir dann im Übrigen mit der wie folgt adaptierten Effektgröße operiert:

(19) $\Phi^2_{VA,BA,B} = \varphi^2_{VA,B}/\pi_{BA}$.

PH-5: »Bei imaginaler Kodierung im Vergleich zu verbaler Kodierung *verändert* sich im Mittel die Gedächtnisleistung *nicht*«. Die qualitative Hypothese ist ungerichtet; es besteht *kein* Zusammenhang zwischen unabhängiger und abhängiger Variablen:

(20) DKT: PH-5 $\approx>$ PV-5: $(MWR_1 = MWR_2) \approx>$ SV-5: $(\mu_1 = \mu_2) \Leftrightarrow$
ST-5: $(H_{0,18}: \mu_1 - \mu_2 = 0)$.

Die Strenge der Prüfbarkeit der PH-5 ist höher als die der PH-3, weil die PH-5 zu exakten Vorhersagen führt. Die $H_{0,18}$ kann mittels zweiseitigem t- oder einseitigem F-Test getestet werden. Hier stellt die Effektgröße wieder einen klein zu wählenden Höchsteffekt ($\pm\delta_{B,krit,18}$ bzw. $\varphi_{VA,B,krit,18}$) dar: $|\pm 0{,}20| \leq \delta_{B,krit,18} \leq |\pm 0{,}40|$ bzw. $0{,}10 \leq \varphi_{VA,B,krit,18} \leq 0{,}20$. Je größer EG_{krit} als

Höchsteffekt gewählt wird, desto mehr widerspricht er der abgeleiteten $H_{0,18}$, die einen Nulleffekt behauptet. Die Testplanung für die $H_{1,18}$ erfolgt für diesen Höchsteffekt.

Bei $K = 3$ resultiert:

(21) PV-5($K = 3$): ($MWR_1 = MWR_2 = MWR_3$) $\approx>$ SV-5($K = 3$) $\Leftrightarrow$
ST-5($K = 3$): ($H_{0.19}$: $\mu_k = \mu_{k'}$ für *alle Paare* k, k' mit $k \neq k'$).

Die Varianzanalyse ist hier adäquat und erschöpfend und bewirkt testbedingt eine konjunktive Verknüpfung der $K(K-1)/2$ Prüfinstanzen zu $S = 1$. Hier wird die Effektgröße wegen der abgeleiteten Nullhypothese gering gewählt, also etwa $0{,}08 \leq \varphi_{VA,B,krit,19} \leq 0{,}16$.

Alternativ können aus der SV-5($K = 3$) auch $H_{0,s}$ abgeleitet werden. Aus $AH_{0,20} \wedge AH_{0,21}$ folgt nicht auch $AH_{0,22}$, sodass auch $H_{0,22}$: ($\mu_1 = \mu_3$) $\Leftrightarrow H_{0,15}$ aufgestellt wird [$S = K(K-1)/2 = 3$]. Andererseits gilt: Aus $AH_{0,20} \wedge AH_{0,21}$ folgt, dass $\mu_1 = \mu_3$; und aus dieser Sicht ist die $H_{0,22}$ überflüssig ($S = K - 1 = 2$). Die separate Betrachtung der Prüfinstanzen führt dabei im Vergleich zur testbedingten Zusammenfassung zu einer Erhöhung der f-Validität, weil bei gleichem Φ_{VA} für die Varianzanalyse und die separaten Tests letztere auf einem durch die Adjustierung erniedrigten Signifikanzniveau $\alpha_{s,krit}$ beruhen, das dazu führt, dass die vorhersagekonformen Einzelhypothesen länger beibehalten werden, als dies bei der Varianzanalyse der Fall ist.

(22) SV-5($K = 3$) $\approx>$ ST-5($K = 3$): $\{(H_{0,17}: \mu_1 - \mu_2 = 0) \wedge (H_{0,18}:$
$\mu_2 - \mu_3 = 0)$ $[\wedge (H_{0,19}: \mu_1 - \mu_3 = 0)]\}$.

Bei disjunktiver Verknüpfung entstehen $S = K(K-1)/2 = 3$ Prüfinstanzen:

(23) ST-5($K = 3$) $\approx> (H_{0,20} \vee H_{0,21} \vee H_{0,22})$.

Es gilt bei konjunktiver Verknüpfung der Nullhypothesen $H_{0,s}$: $\varepsilon \leq \max(\beta_s)$ und $\varphi \leq \Sigma\alpha_s$ sowie $\beta_{krit,s} = \varepsilon_{krit}$ und $\alpha_{krit,s} = \varphi_{krit}/S$ und bei disjunktiver Verbindung der $H_{0,s}$: $\varepsilon \leq \Sigma\beta_s$ und $\varphi \leq \max(\alpha_s)$ sowie $\beta_{krit,s} = \varepsilon_{krit}/S$ und $\alpha_{krit,s} = \varphi_{krit}$.

PH-6: »Zeitschätzungen fallen im Mittel umso ungenauer aus, je länger die zu schätzende Zeitdauer ist.« Diese psychologische Hypothese nimmt bereits auf empirische Variablen Bezug, die von der PV-6 aufgegriffen werden. Zusätzlich spezifiziert diese die Anzahl K der Faktorstufen der zu schätzenden Zeitdauer (empirische UV) und damit den Versuchsplan; empirische abhängige Variable sind die Zeitschätzungen: »Ungenauer« wird als »stärker streuen« interpretiert, sodass die PV-6 lautet: »Die mittleren

Zeitschätzungen streuen umso stärker, je länger die zu schätzende Zeitdauer ist (k = 1: 10 sec; k = 2: 60 sec und k = 3: 100 sec)«. Die SV-6 besagt dann, dass sich die drei Varianzen (σ^2_k) mit der zu schätzenden Zeitdauer (empirische UV) vergrößern; es wird ein monotoner Trend über drei Varianzen vorhergesagt.

(24) PH-6 $\approx>$ PV-6 $\approx>$ SV-6: $(\sigma^2_1 < \sigma^2_2 < \sigma^2_3) \approx>$ ST-6:
$[(H_{1,23}: \sigma^2_2/\sigma^2_1 > 1) \wedge (H_{1,24}: \sigma^2_3/\sigma^2_2 > 1)]$.

Für die gerichteten $H_{1,23}$ und $H_{1,24}$ ist jeweils ein einseitiger F-Test vorzusehen. – Als Effektgröße wird

(25) $\Delta_{KT,V} = (\sigma^2_2 - \sigma^2_1)/(\sigma^2_2 + \sigma^2_1)$

verwendet (vgl. Kraemer u. Thiemann, 1987, S. 68) und als großer Mindesteffekt spezifiziert, etwa als $\delta_{KT,V,krit,r} = 0{,}55$; der Vpn-Bedarf wird nach der Tabelle in Kraemer und Thiemann (1987, S. 105–112) bestimmt. Die empirische Effektgröße lautet: $D_{KT,V,r} = (s^2_2 - s^2_1)/(s^2_2 + s^2_1)$.

QPH–7: Üblicherweise beziehen sich quantitative psychologische Hypothesen (QPHn) auf beobachtbare Variable (UV und AV) und auch auf eine bestimmte Funktion, so dass sich die quantitative psychologische Vorhersage (QPV) nur durch die Angabe der Anzahl K der Faktorstufen des Versuchsplans von der quantitativen psychologischen Hypothese unterscheidet. Die QPH-7 von Erdfelder und Bredenkamp (1994, S. 607) lautet: »Die Gesamtlernzeit T_i ist für ein bestimmtes Lernmaterial eine individuelle Konstante, und zwar unabhängig davon, in welche Darbietungszeiten t_k diese Gesamtlernzeit unterteilt wird: $T_i = y_{ik}\, t_k$« (»y_{ik}«: Anzahl der Lernversuche pro Darbietungszeit t_k). Als empirische abhängige Variable wird von Erdfelder und Bredenkamp (a. a. O.) die *mittlere* Gesamtlernzeit gewählt; dies entspricht einer Probabilisierung. Die QPV-7 bezieht sich auf eine bestimmte Anzahl K von Versuchsbedingungen, und aus ihr wird die QSV-7($K \geq 2$) abgeleitet: »Für den Erwartungswert der Gesamtlernzeiten, $E(T)$, gilt bei $K \geq 2$ Faktorstufen t_k: $E(T) = E(Y_k)\, t_k$ = konstant über alle k, k' mit $k \neq k'$«.

(26) QPH-7 $\approx>$ QPV-7 $\approx>$ QSV-7($K \geq 2$) $\approx>$ ST-7($K \geq 2$): $H_{0,25}$:
$(\mu_k = \mu_{k'}$ für alle k, k' mit $k \neq k')$.

Die $H_{0,25}$ der Varianzanalyse stellt eine adäquate und erschöpfende Ableitung aus der psychologischen Vorhersage dar. Sofern $K \geq 3$, führt die Va-

rianzanalyse testbedingt zu einer konjunktiven Verknüpfung der $K(K-1)/2$ Prüfinstanzen zu einer einzigen; dadurch wird die e-Validität bei einer ohnehin streng prüfbaren Hypothese erhöht. Quantitative Hypothesen sind strenger prüfbar als qualitative, weil sie infolge des spezifizierten quantitativen Zusammenhangs zwischen unabhängiger und abhängiger Variablen einen höheren empirischen Gehalt aufweisen als letztere. Die Effektgröße für die Varianzanalyse, $\varphi_{VA,B,krit}$, wird auf einen möglichst geringen Wert festgelegt; der benötigte Vpn-Bedarf ist hoch.

QPH-8: Die gemessen an Lehrbuchdarstellungen typische quantitative Hypothese QPH-8 behauptet einen bestimmten quantitativen Trend zwischen empirischer unabhängiger und empirischer abhängiger Variablen: »Die Anzahl der Lernversuche wächst im Mittel ausschließlich linear mit sinkenden Darbietungszeiten an« (modifiziert nach Erdfelder u. Bredenkamp, 1994, S. 607). Die QPV-8($K \geq 3$) lautet: »Die mittlere Anzahl der Lernversuche wächst bei $K \geq 3$ sinkenden Darbietungszeiten t_k ausschließlich linear an«. Dies ist auch die verbale Umschreibung der SV-8($K \geq 3$):

(27) QPH-8 $\approx>$ QPV-8($K \geq 3$) $\approx>$ SV-8($K \geq 3$) $\approx>$ ST-8($K \geq 3$):
$[(H_{1,26}\colon \psi_{Lin} = \Sigma c'_{k,Lin}\, \mu_k > 0) \wedge (H_{0,27}\colon \Sigma\psi^2_{,Abw,24} = 0)]$.

Hierbei bedeutet: »ψ«: Kontrast (empirische Entsprechung: $D_{Lin} = \Sigma c'_{k,Lin}\, M_k$), $c'_{k,Lin}$: tabellierte lineare Kontrastkoeffizienten.

Die QPH-8 ist weniger streng prüfbar als die QPH-7, da aus der QPH-8 auch eine unexakte Vorhersage abgeleitet wird. Die komplexen Trendkontraste beruhen auf Kontrastkoeffizienten $c'_{Trend,k}$, bei denen die Summe der Beträge in den meisten Fällen ungleich 2 ist. Cohens δ setzt allerdings voraus, dass diese Summe gleich 2 ist. Um dies bei komplexen Kontrasten zu erreichen, werden sie durch die halbe Summe ihrer Beträge ($0{,}5\,|\Sigma c'_{k,Trend}|$) dividiert, und für die neuen Koeffizienten $c_{k,Trend}$ ist die Summe der Beträge gleich 2. Für sie kann man sich dann bei der Testplanung wieder an den Konventionen Cohens (1988, S. 25–27) orientieren. Einem großen Effekt entspricht $\delta_{B,krit,t} = 0{,}80$, und die entsprechende Testplanung für den einseitigen t-Test der $H_{0,26}$ erfolgt mit der Formel

(28) $n_k = (z_{1-\alpha} + z_{1-\beta})^2\, \Sigma c^2_k / \delta^2_{B,krit,t}$.

Bei der Berechnung des *t*-Tests und der Bestimmung der empirischen Effektgröße $d_{B,Lin} = \Sigma c_{k,Lin}\, M_k$ sind die transformierten Koeffizienten zu benutzen, mit denen auch bei der Testplanung operiert wurde. Die $H_{0,27}$ besagt, dass alle Abweichungen von der Linearität simultan gleich Null sind. Diese

K – 2 orthogonalen nichtlinearen Trendkontraste lassen sich quadriert zu einem Vergleich mit $FG_Z = K - 2$ zusammenfassen (FG_Z: Zählerfreiheitsgrade). Für die Testung der $H_{0,27}$ mittels F-Test einer Varianzanalyse bei $FG_Z = K - 2$ wird die Kriteriumseffektgröße EG_{krit} als möglichst geringer Höchsteffekt festgelegt: $0{,}10 \leq \Phi_{VA,Abw,B,krit,27} \leq 0{,}25$. Es gilt hier: $f^2_{Abw,B,27} = \sigma QSAbw/QSI$ und $\sigma QSAbw = QSB - QSLin$. (QSLin: lineare Quadratsumme). Für den F-Test gilt: $F = [(QSTreat - QSLin)/(K - 2)]/s^2_1$. – Der signifikante F-Test gibt keinen Aufschluss über die Abweichungen. Interessieren diese, so wird für jeden der $K - 2$ nonlinearen Trendkontraste eine separate $H_{0,Trend,u}$: $\psi_{Trend,u} = \Sigma c_{k,Trend,u}\, \mu_k = 0$ ($u = 2, \ldots, K - 1$; u: Ordnung des Trends) vorgesehen; die Effektgrößen lauten: $\varphi_{VA,B,Trend,u,krit} = \delta_{B,Trend,u,krit}/(K \Sigma c^2_{k,Trend,u})^{1/2}$ beziehungsweise $\varrho_{Trend,u,krit}$ und $r_{Trend,u}$; $\sigma QSTrend,u = \sigma QSAbw$ und $\Sigma r^2_{Trend,u} = R^2_{Abw}$; der Vpn-Bedarf wird wie oben beschrieben bestimmt. Wenn man die Faktorstufen identifizieren will, in denen gegebenenfalls Abweichungen vorliegen, wird für jede Faktorstufe die $H_{0,k}$: $(\mu_k - \mu_{k,Lin} = 0)$ eines Ein-Stichproben-t-Tests abgeleitet; die Effektgrößen lauten:

(29) $\delta_{B,krit,k} = (\mu_k - \mu_{k,Lin})/\sigma_k$

sowie

(30) $d_{B,k} = (M_k - M_{k,Lin})/s_k$.

(s_k: Streuung in Bedingung B_k). Der Vpn-Bedarf wird nach

(31) $n_k = (z_{1-\alpha} + z_{1-\beta})^2/\delta^2_{B,krit,k}$

bestimmt. Bei der Testplanung für die $H_{1,26}$ und die $H_{0,27}$ beziehungsweise für die beiden anderen Vorgehensweisen ($R = 1$ und $S = K - 2$ bzw. $S = K$) werden erneut sehr unterschiedliche Stichprobenumfänge resultieren, sodass man wie bei der Prüfung der PH-4 über die PV-4($K = 3$) oben verfahren wird. Die SV-8($K \geq 3$) wird über eine Konjunktion von $R = 1$ H_1 und $S = 1$ H_0 getestet; die Kumulierung lautet dabei: $\varepsilon \leq \max(\alpha_r, \beta_s)$ und $\varphi \leq (\Sigma\beta_r + \Sigma\alpha_s)$, woraus für die Adjustierung folgt: $\alpha_{krit,r} = \beta_{krit,s} = \varepsilon_{krit}$ und $\beta_{krit,r} = \alpha_{krit,s} = \varphi_{krit}/(R + S)$ (siehe zu weiteren Vorgehensweisen bei quantitativen unabhängigen Variablen und Trends Bredenkamp, 1968, sowie Hager, 2004).

Sieht man eine oder beide genannten Strategien *anstelle* der $H_{0,27}$ vor, müssen deren Tests *a priori* geplant und die Fehlerwahrscheinlichkeiten adjustiert werden; bei ihrer Anwendung *nach* $AH_{1,26} \wedge AH_{1,27}$ bedarf es keiner Planung und Adjustierung wie immer bei einer der Hypothesenprüfung

folgenden testgestützten Inspektion der Daten, denn die zusätzlichen Informationen sind für die *Hypothesenprüfung irrelevant*. Für diese notwendig, adäquat und erschöpfend sind unabhängig von der Anzahl K der Faktorstufen nur die aus Gründen der Erschöpfendheit konjunktiv verknüpften $H_{0,26} \wedge H_{0,27}$, die sich testbedingt auf *eine* Prüfinstanz mit $K \geq 3$ beziehen und dabei zu einer konjunktiven Verknüpfung möglicher Prüfinstanzen zu einer führen. Je größer man K wählt, desto mehr vorhersagewidrige Trends höherer Ordnung können auftreten. Dies erhöht die e-Validität.

Entscheidungen

Nach den Ableitungen erfolgt die Testplanung, dann werden die Daten erhoben und im Anschluss daran die *a priori* geplanten statistischen Tests durchgeführt. Die Entscheidung über die jeweils gegeneinander getesteten statistischen Hypothesen wird auf Grund des vorgewählten Signifikanzniveaus $\alpha_{krit,t}$ ($t = 1, \ldots, T$; T: Anzahl der Tests) getroffen. Es handelt sich um eine rein formale Entscheidung relativ zu vorher festgelegten Entscheidungskriterien mit zwei möglichen Ausgängen pro Hypothese: »Annahme der H_0 (AH_0)« oder »Annahme der H_1 (AH_1)«. Die Entscheidung über die vorgeordnete statistische Vorhersage erfolgt auf der Grundlage der Entscheidungen auf der Ebene der statistischen Tests (ST). Die statistische Vorhersage wird bei nur einer abgeleiteten Hypothese dann angenommen, wenn das Ergebnis auf der nachgeordneten Ebene vorhersagekonform ausgefallen ist. Bei mehreren Tests auf der Ebene der statistischen Tests wird die statistische Vorhersage jeweils nur dann angenommen, wenn bei konjunktiver Verknüpfung der Prüfinstanzen alle Tests vorhersagekonform ausgehen oder wenn bei disjunktiver Verknüpfung der oder die Tests mindestens einer Prüfinstanz vorhersagekonform ausgegangen sind.

Bei der als Nächstes zu fällenden Entscheidung über die psychologische Vorhersage werden die Ergebnisse der statistischen Tests und bei abgeleiteten Alternativhypothesen auch die Effektgrößen berücksichtigt, da die letzteren von allen Determinanten des statistischen Tests noch den relativ engsten Bezug zu den psychologischen Inhalten aufweisen. Bei PV-B $\approx>$ SV-B $\Leftrightarrow H_{1,B}$ wird der empirische Wert der Effektgröße (EG_{emp}) mit dem zum Zeitpunkt der Testplanung festgelegten (oder bestimmten) Kriteriumswert EG_{krit} verglichen. Dieser Vergleich erfolgt rein deskriptiv, also ohne Zuhilfenahme statistischer Tests. Würde man letztere in diesem Kontext einsetzen, dann müsste auch für diese Tests wieder eine Testplanung erfolgen, was

leicht in einen unendlichen Regress münden kann. Die Annahme der hypothesenkonformen H_1 bei $EG_{B,emp} \geq EG_{B,krit}$ führt zu der Entscheidung, die PV-B als eingetreten zu bewerten, wobei man geringfügig kleinere Effekte $EG_{B,emp}$ als $EG_{B,krit}$ nicht zum Anlass nehmen sollte, die PV-B als nicht eingetreten zu bewerten, sofern der statistische Test vorhersagekonform ausgegangen ist. Je größer der bei der Testplanung festgelegte Wert der Effektgröße ist, desto schwerer wird es der PV-B gemacht, als eingetreten beurteilt zu werden. Andererseits: Je geringer dieser Mindestwert festgelegt worden ist, desto eher wird die PV-B als eingetreten beurteilt, aber desto mehr Versuchspersonen werden auch benötigt, um ihn nachzuweisen. Daher sollte eine Verringerung von EG_{krit} in der Planungsphase, aber auch seine Vergrößerung nur »mit Augenmaß« vorgenommen werden. Die Beibehaltung der hypothesenkonträren $H_{0,B}$ führt bei genügender Teststärke dazu, die PV-B als nicht eingetreten anzusehen.

Bei PV-A $\approx>$ SV-A $\Leftrightarrow$ $H_{0,A}$ nimmt bei $AH_{0,A}$ $EG_{A,emp}$ zwangsläufig einen geringeren Wert an als der für die vorhersagekonträre $H_{1,A}$ festgelegte $EG_{A,krit}$. Lediglich bei Annahme einer vorhersagekonformen und gerichteten $H_{0,A}$ enthält die unter dieser bestimmbare $EG_{A,emp}$ zusätzliche Information und sollte deshalb als deskriptives Maß stets angegeben werden. Bei genügender Teststärke spricht $AH_{0,A}$ für das Eintreten der PV-A; $AH_{1,A}$ bedeutet ihr Nicht-Eintreten.

Für die Entscheidung über die Bewährung oder die Nicht-Bewährung der psychologischen Hypothese ist die Entscheidung über die psychologische Vorhersage eine unverzichtbare Grundlage. Bei Nicht-Eintreten der psychologischen Vorhersage sollte in aller Regel auf *Nicht-Bewährung der psychologischen Hypothese* entschieden werden. Für eine *Bewährung der psychologischen Hypothese* stellt das Eintreten der psychologischen Vorhersage jedoch nur eine *notwendige, aber keineswegs auch hinreichende* Voraussetzung dar. Denn bei der nicht formalisierbaren Entscheidung über die Bewährung der psychologischen Hypothese ist jeweils auch die *experimentelle Validität* des Versuchs zu berücksichtigen. Waren die Anfangsbedingungen zur Prüfung der Hypothese gegeben, das heißt, konnte die psychologische Hypothese überhaupt geprüft werden? Wurde randomisiert (interne Validität)? Wurden die theoretischen Begriffe der psychologischen Hypothese über paradigmatische oder zumindest plausible und allgemein akzeptierte empirische Begriffe operationalisiert (Variablenvalidität)? Wurde mit einer oder mit mehreren Prüfinstanzen operiert, und wie wurden mehrere Prüfinstanzen zusammengefasst (konjunktiv oder disjunktiv; e- und/oder f-Validität)? Wurden bei den Ableitungen die Adäquatheit und die Erschöpfendheit berücksichtigt (Ableitungsvalidität)? Wurde bei der Ableitung der

statistischen Vorhersage das Skalenniveau der empirischen abhängigen Variablen (Variablenvalidität) und die Planbarkeit der resultierenden Tests beachtet? Wurde eine Testplanung zur Kontrolle der Fehlerwahrscheinlichkeiten relativ zu EG_{krit} durchgeführt (statistische sowie e- und/oder f-Validität)? Wie wurden bekannte Störfaktoren kontrolliert (e-und/oder f-Validität)? Welche unerwarteten Störfaktoren sind aufgetreten (e- und/oder f-Validität)? Waren die Hilfshypothesen bewährt oder zumindest plausibel und allgemein akzeptiert? Welche Hilfshypothesen haben sich als nicht aufrecht zu erhalten erwiesen? Waren die Randbedingungen des Versuchs so wie geplant? Erfolgte eine Dateninspektion zur Entdeckung von Auffälligkeiten? Sind nachträglich nicht mehr zu beseitigende Validitätsmängel festzustellen?

Die Antworten auf diese Fragen hängen vom tatsächlichen Verlauf des Versuchs, vom verfügbaren Hintergrundwissen, vom Sachverstand und vom Expertenwissen des Versuchsleiters ab. Bei gravierenden Validitätsmängeln sollte die Entscheidung über die psychologische Hypothese negativ ausfallen oder nicht getroffen werden, selbst wenn alle Entscheidungen auf den nachgeordneten Ebenen erwartungskonform erfolgt sind.

Jede positive Entscheidung über eine psychologische Hypothese mit theoretischen Begriffen ist *gehaltserweiternd*, vor allem, weil in einem genau umschriebenen Versuch mit Konkretisierungen gearbeitet wird, die eingeschränkter sind als die theoretischen Begriffe der psychologischen Hypothese.

Literatur

Bredenkamp, J. (1968). F-Tests zur Prüfung von Trends und Trendunterschieden. *Zeitschrift für Experimentelle und Angewandte Psychologie, 15,* 239–272.

Bredenkamp, J. (1969). Über die Anwendung von Theorie-testenden Experimenten. *Psychologische Beiträge, 11,* 275–285.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Wiesbaden: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

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

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann u. W. Tack (Hg.), *Methodologische Grundlagen der Psychologie* (S. 604–648). Göttingen: Hogrefe.

Erdfelder, E., Faul, F., Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers, 28,* 1–11.

Hager, W. (2004). *Testplanung bei der statistischen Prüfung von psychologischen Hypothe-*

sen. Die Ableitung von Vorhersagen und die Kontrolle der Determinanten des statistischen Tests. Göttingen: Hogrefe.

Kraemer, H.C., Thiemann, S. (1987). *How many subjects?* Newbury Park, CA: Sage.

Meehl, P.E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society, 105,* 292–327.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Popper, K.R. (1984). *Logik der Forschung* (8. Aufl.). Tübingen: Mohr.

Westermann, R. (1987). *Strukturalistische Theorienkonzeption und empirische Forschung in der Psychologie.* Berlin: Springer-Verlag.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur Psychologischen Methodenlehre.* Göttingen: Hogrefe.

Westermann, R., Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11,* 117–146.

Wippich, W., Bredenkamp, J. (1979). *Bildhaftigkeit und Lernen.* Darmstadt: Steinkopff.

Bayerische
Staatsbibliothek
München

Edgar Erdfelder[1]

Angewandte deduktivistische Methodologie in der empirischen Psychologie

Zusammenfassung

In der Forschungspraxis stellt sich ständig das Problem, zwischen unterschiedlichen, auf den ersten Blick oft gleichermaßen plausiblen Methoden zur Untersuchung einer wissenschaftlichen Fragestellung wählen zu müssen, wobei die Methodenwahl die Ergebnisse beeinflussen kann. Eine angewandte Methodologie ist ein System von normativen Regeln, das bestimmt, wie diese Methodenentscheidungen zu treffen sind oder welche Wahlen vermieden werden sollten. In diesem Kapitel wird eine angewandte Methodologie dargestellt, die sich einer deduktivistischen Forschungskonzeption der empirischen Psychologie verpflichtet fühlt (Bredenkamp, 1972, 1980; Gadenne, 1976, 1984; Hager u. Westermann, 1983a, 1983b). Auf Kritik an diesem Ansatz (Westmeyer, 1998) wird abschließend eingegangen.

Einleitung

Welche Methodologie liefert eine geeignete Grundlage wissenschaftlichen Arbeitens? Diese Frage wird von Wissenschaftlerinnen und Wissenschaftlern einer Disziplin selten einhellig beantwortet. Zudem scheint die Mehrheitsmeinung historischem Wandel unterworfen zu sein. Während zum Beispiel die Falsifikationsmethodologie Poppers (1934/1982) sowie deren Modifikationen durch Lakatos (1970) und andere kritische Rationalisten vor 20 bis 30 Jahren recht viele Anhänger in der empirischen Psychologie fanden (s. Bredenkamp, 1972, 1980; Gadenne, 1976, 1984; Herrmann, 1976, 1979), dominiert derzeit eher die Kritik an diesem Ansatz bei gleichzeitiger Präferenz für weniger normative, sondern eher historisch-rekonstruktive Ansätze in der Wissenschaftstheorie (z. B. Westermann, 1987, 2000, in diesem Band; Westmeyer, 1989, 1992, 1998).

1 Für hilfreiche Kommentare und Kritik an einer Vorform dieser Arbeit bin ich Volker Gadenne, Albrecht Iseler und Rolf Steyer sehr dankbar.

Ein Aspekt, der in der Debatte um das Pro und Kontra verschiedener wissenschaftstheoretischer Konzeptionen erstaunlicherweise oft unberücksichtigt blieb, ist die Frage nach dem praktischen Nutzen einer Methodologie im Forschungsalltag. Das vorliegende Kapitel geht von der Prämisse aus, dass eine gute Methodologie auch eine angewandte Methodologie beinhalten sollte, die konkrete Entscheidungshilfen gibt, wenn es im Forschungsprozess darum geht, zwischen mehreren optionalen Methoden eine optimale Wahl zu treffen. Aufbauend auf einer früheren Arbeit (Erdfelder u. Bredenkamp, 1994) werde ich Grundbausteine einer angewandten Methodologie darstellen, die durch Ideen von Jürgen Bredenkamp (1972, 1980), Volker Gadenne (1976) sowie Willi Hager und Rainer Westermann (1983a, 1983b) angeregt wurde.[2] Ebenso wie der methodologische Falsifikationismus (Popper, 1934/1982; Lakatos, 1970), der das wissenschaftstheoretische Fundament bildet, kann diese Methodologie im doppelten Sinne »deduktivistisch« genannt werden. Zum einen versucht sie, induktive Argumente wegen der ungelösten Problematik der Rechtfertigung gehaltserweiternder Schlüsse zu vermeiden und durch deduktive Argumente zu ersetzen (Gadenne, 1976, 1984, in diesem Band). Zum anderen ist die Methodologie selbst deduktiv aufgebaut, da das Regelsystem, das diese Methodologie charakterisiert, dadurch ausgezeichnet ist, dass sich alle Regeln aus einer einzigen »obersten methodologischen Regel« (Popper, 1934/1982) ableiten lassen. Poppers Überlegungen hierzu lassen sich jedoch wegen der Besonderheiten des Forschungsprozesses in der empirischen Psychologie nicht unmittelbar auf die Psychologie übertragen, so dass Poppers oberste methodologische Regel geeignet zu reformulieren ist.

Die statistische Prüfung psychologischer Hypothesen

Seit etwa fünfzig Jahren zeichnet sich die psychologische Forschungspraxis dadurch aus, dass substanzwissenschaftliche Hypothesen mehr oder minder routinemäßig mit statistischen Mitteln überprüft werden (Bredenkamp, 1969, 1972; Gigerenzer, 1987, 1993; Gigerenzer u. Murray, 1987, Kap. 1). Konkret bedeutet dies, dass aus einer interessierenden substanz-

2 In der konkreten Umsetzung der Ideen entferne ich mich allerdings von den genannten Arbeiten, sodass alle etwaigen Unzulänglichkeiten der vorgenommenen Erweiterungen selbstverständlich ausschließlich dem Verfasser anzulasten sind.

wissenschaftlichen Hypothese – im Folgenden kurz »psychologische Hypothese« (PH) genannt – zumeist auf nicht näher spezifiziertem Wege eine oder mehrere statistische Hypothesen gewonnen werden, über die sodann anhand einer Datenstichprobe mit inferenzstatistischen Mitteln entschieden wird (Meehl, 1967). Die zugrunde gelegte statistische Entscheidungstheorie wird dabei selten explizit gemacht; anhand von gebräuchlichen Lehrtexten der Statistik lässt sie sich am ehesten als hybride Mischung der Ideen von R. A. Fisher einerseits und J. Neyman sowie E. S. Pearson andererseits rekonstruieren (Gigerenzer, 1987, 1993). In jedem Fall zieht die statistische Entscheidung Konsequenzen für die vorgeordnete psychologische Hypothese nach sich: Die Annahme oder Verwerfung der als Prüfinstanz dienenden statistischen Hypothese (SH) wird interpretiert als Bestätigung oder Nichtbestätigung der psychologischen Hypothese, aus der die statistische Hypothese gewonnen wurde.

Bemerkenswert ist dieses Vorgehen vor allem deshalb, weil die geprüfte statistische Hypothese im Regelfall keineswegs als formale Explikation der psychologischen Hypothese gelten kann, die im Zentrum des Interesses steht. Vielmehr ist oftmals ohne Mühe erkennbar, dass die sie einen geringeren empirischen Gehalt (i. S. von Popper, 1934/1982) als die psychologische Hypothese aufweist. In der Psychologie handelt es sich bei der geprüften statistischen Hypothese zum Beispiel häufig um eine Mittelwertshypothese über Aggregate von Individuen (Versuchsgruppen), wohingegen die psychologische Hypothese oft allgemeinpsychologischer Natur ist und sich somit auf einzelne, prinzipiell beliebige Individuen eines offenen Individuenbereichs bezieht. Die Gültigkeit der psychologischen Hypothese kann unter diesen Umständen nicht aus dem Zutreffen der statistischen Hypothese gefolgert werden (vgl. Bredenkamp, 1972, 1980; Gadenne, 1976). Man kann es auch noch schärfer formulieren: Selbst wenn sich statistische Entscheidungsfehler völlig ausschließen ließen, wäre im Fall einer Entscheidung für die abgeleitete Aggregathypothese ohne weitere Zusatzannahmen nicht garantiert, dass es auch nur ein einziges Individuum gibt, welches der interessierenden psychologischen Hypothese folgt. Lautet etwa die psychologische Hypothese, dass die Regression einer abhängigen Variablen (AV) Y auf eine unabhängige Variable (UV) X für jedes Individuum linear ist, und ist die Regression de facto für jedes Individuum nichtlinear, so kann sich auf der Ebene der statistischen Hypothese – wenn man die Daten über die Individuen mittelt – unter bestimmten Umständen dennoch ein linearer Zusammenhang zwischen Y und X ergeben (ein konkretes Beispiel nennt z. B. Steyer, 1992, S. 34 f.). Ebenso können mittlere Lernkurven, die einen Lernindikator (AV) als Funktion der

Lernzeit (UV) darstellen, einen kontinuierlichen Lernprozess suggerieren, auch wenn die Lernprozesse jedes einzelnen Individuums durch unstetige Lernkurven mit qualitativen Leistungssprüngen gekennzeichnet sind.

Die Selbstverständlichkeit, mit der in der Psychologie psychologische über statistische Hypothesen geprüft werden, muss in Anbetracht dieses SH-PH-Überbrückungsproblems überraschen. Wie Erdfelder und Bredenkamp (1994) gezeigt haben, lässt sich das übliche Vorgehen jedoch unter bestimmten Voraussetzungen rechtfertigen. Ihr Ansatz greift das bei Gadenne (1976) angedeutete und von Hager und Westermann näher ausgeführte Konzept einer »strengen« (Hager u. Westermann, 1983a, 1983b) beziehungsweise »validen« (Westermann u. Hager, 1986; Hager, 1987, 1992, in diesem Band; Westermann, 1987) statistischen Prüfung von psychologischen Hypothesen auf. Während Hager und Westermann darin allerdings eher eine regulative Idee und weniger ein quantitativ präzisierbares Konstrukt sahen, haben Erdfelder und Bredenkamp (1994) aufgezeigt, wie sich das Validitätskonzept Gewinn bringend präzisieren lässt, wenn man bereit ist, zwei Aspekte voneinander zu trennen: (a) die statistische Explikation umgangssprachlich formulierter psychologischer Hypothesen einerseits und (b) die deduktive Überprüfung einer statistisch explizierten psychologischen Hypothese andererseits (vgl. auch Iseler, 1999). Da der erste, »vorstatistische« Schritt die psychologische Hypothese überhaupt erst eindeutig macht, sodass klar wird, was genau zu überprüfen ist, kann sich eine deduktivistische Methodologie der Hypothesenprüfung sinnvoll nur auf den zweiten Aspekt, das heißt die statistische Prüfung bereits explizierter psychologischer Hypothesen beziehen.

Das Validitätspostulat als oberste methodologische Regel

Hager und Westermann (1983a) folgend kann man zwei Arten von Fehlerwahrscheinlichkeiten bei der statistischen Prüfung einer psychologischen Hypothese unterscheiden. Mit e wird die Wahrscheinlichkeit eines fälschlichen Bewährungsurteils bezeichnet, mit f entsprechend die Wahrscheinlichkeit eines fälschlichen Nichtbewährungsurteils:

$e = p$(Bewährungsurteil, gegeben, dass die PH nicht gilt)

$f = p$(Nichtbewährungsurteil, gegeben, dass die PH gilt).

Eine wissenschaftliche Untersuchung kann als umso strenger bezeichnet werden, je kleiner *e* ist. Das Komplement 1–*e* dieser Fehlerwahrscheinlichkeit eignet sich folglich als Maß der *Strenge* einer Untersuchung. Analog kann eine wissenschaftliche Untersuchung als umso fairer bezeichnet werden, je kleiner *f* ist. Das Komplement 1–*f* kann folglich als Maß der Fairness einer Untersuchung dienen.

Um Missverständnisse zu vermeiden, müssen zwei Dinge klargestellt werden. Zum einen sind *e* und *f* nicht mit den Fehlerwahrscheinlichkeiten α und β beim Testen statistischer Hypothesen zu verwechseln (vgl. dazu Erdfelder et al. in diesem Band). Eine direkte Korrespondenz zwischen α und β auf der einen Seite und *e* und *f* auf der anderen Seite besteht nur dann, wenn die psychologische Hypothese mit der getesteten statistischen Hypothese äquivalent ist, was – wie im letzten Abschnitt ausgeführt – zumindest in der Psychologie sehr selten der Fall ist. Zum anderen muss die psychologische Hypothese – obgleich nicht mit der getesteten statistischen Hypothese äquivalent – dennoch selbst eine statistisch formulierte Hypothese sein, da nur dann sinnvoll von *Wahrscheinlichkeiten* bei der Prüfung der psychologischen Hypothese die Rede sein kann. Formal gesehen handelt es sich bei der statistischen Prüfung psychologischer Hypothesen also um das Problem, eine eigentlich interessierende statistische Hypothese – die PH – auf dem »Umweg« über eine andere statistische Hypothese – die SH – zu prüfen.

Ist dies eine realistische Prämisse für die psychologische Forschung? Anders als Hager (1987, 1992, in diesem Band) bin ich der Meinung, dass diese Frage zu bejahen ist, denn der wesentliche Unterschied zwischen der zu prüfenden psychologischen Hypothese und der als Prüfinstanz fungierenden statistischen Hypothese besteht nicht in der Sprache, in der die Hypothesen formuliert sind. Das Hauptproblem besteht vielmehr darin, dass eine geeignete stochastische Explikation der psychologischen Hypothese zwar im Regelfall problemlos möglich ist, aber eben nicht zu der statistischen Hypothese führt, die letztlich getestet wird. So gesehen wird das PH-SH-Überbrückungsproblem durch eine stochastische Formulierung der psychologischen Hypothese also keineswegs gelöst, sondern – im Gegenteil – es wird dadurch erst richtig deutlich.

Der Einfachheit halber soll dies im Folgenden exemplarisch für die große Klasse der psychologischen Hypothesen illustriert werden, die behaupten, dass für alle Individuen *s* einer Population von Personen ein quantitatives Verhaltensmerkmal *Y*, die so genannte abhängige Variable (AV), unter einer Bedingung $X = 1$ (z. B. Experimentalbedingung, EB) stärker ausgeprägt ist als unter einer Bedingung $X = 0$ (z. B. Kontrollbedingung, KB). Die beiden

Ausprägungen von X bilden die Stufen der unabhängigen Variablen (UV). Hypothesen dieser Art finden sich in allen Teildisziplinen der Psychologie (»Frustration verursacht Aggression«, »Visualisierung verbessert die Gedächtnisleistung«, »Konfrontationstherapie reduziert phobische Angst« etc.). Eine angemessene statistische Formulierung wäre:

(1) Für alle s: $E(Y|X, S = s) = a_s + b_s \cdot X$, mit $b_s > 0$.

In Gleichung (1) wie auch im Folgenden bezeichnen kursive Großbuchstaben (z. B. S) Zufallsvariablen (ZVn) und die korrespondierenden kursiven Kleinbuchstaben (z. B. s) konkrete Werte beziehungsweise Realisationen dieser Zufallsvariablen. Die Werte der bedingten Erwartung $E(Y|X, S = s)$ sind entsprechend die Erwartungswerte von Y in Bedingung $X = 1$ und $X = 0$ für das Individuum s. Durch Formulierung der psychologischen Hypothese in Termini von individuumsspezifischen Erwartungswerten soll dem Sachverhalt Rechnung getragen werden, dass Variablen in der Psychologie typischerweise keine perfekte Reliabilität besitzen. Messfehlereinflüssen wird dadurch Rechnung getragen, dass lediglich von der *durchschnittlichen* Ausprägung von Y (und nicht etwa von jeder konkreten Realisation von Y) ein größerer Wert unter $X = 1$ verglichen mit $X = 0$ gefordert wird, nämlich $E(Y|X = 1, S = s) = a_s + b_s$ für die Experimentalbedingung und $E(Y|X = 0, S = s) = a_s$ für die Kontrollbedingung. Der Parameter b_s – der »individuelle kausale Effekt« der Person s (vgl. Rubin, 1974; Steyer, in diesem Band, Steyer et al., 2000a, 2000b) – gibt die Differenz der durchschnittlichen AV-Werte der Person s unter der Experimental- und der Kontrollbedingung an. Die psychologische Hypothese besagt also nichts anderes, als dass alle individuellen kausalen Effekte bezüglich der abhängigen Variablen Y größer als null sind.

Lassen sich allgemeinpsychologische psychologische Hypothesen auch dann noch problemlos statistisch formulieren, wenn die abhängige Variable Y, auf die sich die psychologische Hypothese bezieht, eine nicht direkt beobachtbare latente Variable ist, die – wie in der Psychologie häufig der Fall – ein theoretisches Konstrukt oder eine intervenierende Variable repräsentiert (McCorquodale u. Meehl, 1948)? Man verwendet dann typischerweise anstelle von Y eine empirische Variable Y' als abhängige Variable, die häufig als »Operationalisierung« von Y bezeichnet wird. Gemeint ist damit allerdings keine Operationalisierung im klassischen, deterministischen Sinne (Bridgman, 1938; Carnap, 1936, 1937), sondern eine fehlerbelastete Indikatorvariable, deren bedingte Erwartung zur interessierenden latenten Variablen Y in einer linearen Beziehung steht:

(2) $E(Y'|Y) = c + d \cdot Y$, mit $d > 0$.

Kann man ferner voraussetzen, dass der Messfehler $F := Y' - E(Y'|Y)$ weder durch die unabhängige Variable X noch durch die Person S beeinflusst wird (d. h. $E(F|X, S) = 0$), lässt sich aus der allgemeinpsychologischen Hypothese (1) und dem linearen Messmodell (2) leicht herleiten:

(3) Für alle s: $E(Y'|X, S = s) = E(c + d \cdot Y|X, S = s)$ (nach Gleichung 2)
$= c + d \cdot E(Y|X, S = s)$
$= c + d \cdot (a_s + b_s \cdot X)$ (nach Gleichung 1)
$= (c + d \cdot a_s) + d \cdot b_s \cdot X$, mit $d > 0$, $b_s > 0$.

Die individuellen kausalen Effekte haben sich durch den erforderlichen Übergang von Y zu Y' zwar von b_s zu $d \cdot b_s$ verändert, müssen jedoch nach wie vor sämtlich positiv sein, da beide Parameter positiv sind.

Als Zwischenfazit kann festgehalten werden, dass sich typische allgemeinpsychologische Hypothesen relativ problemlos als statistische Hypothese über individuelle kausale Effekte formulieren lassen, und zwar unabhängig davon, ob sich die psychologische Hypothese nun auf eine empirisch beobachtbare Verhaltensvariable bezieht oder stattdessen auf eine latente Variable, die ein hypothetisches Konstrukt wie »Vertrautheit« oder »Depressivität« repräsentiert. Im letzteren Fall ist lediglich zusätzlich (mindestens) eine empirische Indikatorvariable zu fordern, für die ein lineares Messmodell mit additivem, von der unabhängigen Variablen unabhängigem Fehler plausibel ist.

Bestünde die Möglichkeit, statistische Hypothesen über individuelle kausale Effekte zu testen, wäre ein Übergangsproblem zwischen zu prüfenden psychologischen Hypothesen und als Prüfinstanz fungierenden statistischen Hypothesen also nur insofern gegeben, als immer nur eine Auswahl von statistischen Hypothesen konkret geprüft werden kann, während psychologische Hypothesen sich auf eine offene Menge von statistischen Hypothesen (für alle potenziellen Versuchspersonen, die derzeit leben, in der Vergangenheit gelebt haben oder in Zukunft leben werden) beziehen. Leider kommt aber als weiteres Problem noch hinzu, dass sich Hypothesen über individuelle kausale Effekte sehr häufig nicht an einzelnen Individuen prüfen lassen, weil eine Versuchsperson entweder nur unter der Kontrollbedingung oder nur unter der Experimentalbedingung sinnvoll untersucht werden kann, nicht aber unter beiden Bedingungen nacheinander. Der Grund hierfür ist, dass ein Treatment eine Versuchsperson so verändern kann, dass die anschließende Beobachtung unter der zweiten Treatmentbe-

dingung nicht allein deren Effekt erfasst, sondern zusätzlich die Wechselwirkung mit dem zuvor applizierten Treatment (Bredenkamp, 1980; Erdfelder u. Bredenkamp, 1994). Man entschließt sich aus diesem Grunde häufig zu einer so genannten interindividuellen Bedingungsvariation, bei der Versuchspersonen unter genau einer Bedingung beobachtet werden. Folglich können dann keine statistische Hypothesen über individuelle kausale Effekte getestet werden, sondern ersatzweise nur statistische Hypothesen über die Mittelwertsdifferenz zwischen den beiden Versuchsgruppen:

(4) H_1: $E(Y|X = 1) - E(Y|X = 0) > 0$

Das zentrale methodologische Problem der statistischen Prüfung psychologischer Hypothesen besteht darin, die Prüfung von allgemeinpsychologischen Hypothesen der Form (1) oder (3) über statistische Aggregathypothesen der Form (4) zu rechtfertigen (Bredenkamp, 1972, 1980). Ganz offensichtlich folgt die Gültigkeit der PH (1) nicht aus der Gültigkeit der SH (4), und umgekehrt folgt die Gültigkeit der SH (4) auch nicht ohne weiteres aus der Gültigkeit der PH (1) oder (3). Wie muss eine Untersuchung konzipiert sein, damit dieses Vorgehen dennoch rational gerechtfertigt werden kann?

Zur Beantwortung dieser Frage haben Erdfelder und Bredenkamp (1994) ein komparatives Validitätskonzept für empirische Untersuchungen eingeführt, das auf die genannten Fehlerwahrscheinlichkeiten e und f Bezug nimmt:

> *Definition (Validität):*
> Seien U1 und U2 zwei mögliche Untersuchungen zur Überprüfung einer psychologischen Hypothese, die sich in mindestens einem methodischen Aspekt unterscheiden. Dann ist U1 valider als U2, wenn
> (a) die Strenge $1-e_1$ von U1 größer ist als die Strenge $1-e_2$ von U2 und die Fairness $1-f_1$ von U1 nicht geringer ist als die Fairness $1-f_2$ von U2, oder wenn
> (b) die Fairness $1-f_1$ von U1 größer ist als die Fairness $1-f_2$ von U2 und die Strenge $1-e_1$ von U1 nicht geringer ist als die Strenge $1-e_2$ von U2.

Eine Untersuchung ist also valider als eine potenzielle Alternativuntersuchung, wenn Erstere weder weniger streng noch weniger fair als Letztere ist, jedoch in mindestens einem der beiden Aspekte der Alternativuntersuchung überlegen ist.

Wir können nun die oberste methodologische Regel formulieren (vgl. Erdfelder u. Bredenkamp, 1994):

Oberste methodologische Regel (Validitätspostulat):
Meide die weniger valide Untersuchung und realisiere die validere Untersuchung. Stehen mehr als zwei potenziell realisierbare Untersuchungen zur Auswahl, dann ist die Untersuchung durchzuführen, die valider als alle realisierbaren Alternativen ist. Gibt es keine derartige Untersuchung, sind auf jeden Fall die Untersuchungen zu meiden, zu denen validere realisierbare Untersuchungen existieren. Sind einige infrage kommende Untersuchungen hinsichtlich Strenge, andere dagegen hinsichtlich Fairness überlegen, so wird keine verbindliche Wahl vorgeschrieben. Man kann hier also nach persönlicher Präferenz die strengere oder die fairere Untersuchung auswählen.

Anatomie von Strenge und Fairness einer Untersuchung

Das Validitätspostulat ist ein geeigneter Ausgangspunkt für eine angewandte Methodologie, da es in vielen Fällen Entscheidungen zwischen mehreren potenziellen Untersuchungsmethoden normiert. Voraussetzung für eine fruchtbare Anwendung dieses Postulats ist allerdings die Klärung der Determinanten von Strenge und Fairness. Wie oben ausgeführt, kann die psychologische Hypothese zwar statistisch formuliert, aber nicht direkt statistisch geprüft werden, da (a) die psychologische Hypothese sich auf eine unbegrenzte Menge von Individuen bezieht und (b) bei interindividueller Bedingungsvariation Hypothesen zu individuellen kausalen Effekten nicht für einzelne Versuchspersonen geprüft werden können. Deshalb erfolgt die Prüfung in der Regel über eine statistische Hypothese, die eine bestimmte Rangordnung von Lokationsparametern postuliert. Der häufigste Fall ist die Alternativhypothese eines *t*-Tests für zwei unabhängige Stichproben (H_1: $E(Y|X = 1) > E(Y|X = 0)$). Er dient uns hier als Beispiel.

Mit h sei im Folgenden die Wahrscheinlichkeit bezeichnet, dass die als Prüfinstanz fungierende statistische Hypothese unzutreffend ist (und somit H_0 gilt), falls die zu prüfende psychologische Hypothese zutrifft, das heißt alle Individuen der relevanten Grundgesamtheit positive individuelle kausale Effekte aufweisen. Der kritische Wert der Teststatistik wird wie üblich so gewählt, dass mit einer Wahrscheinlichkeit von höchstens α zugunsten von H_1 entschieden wird, wenn H_0 gilt. Analog sei g die Wahrscheinlichkeit, dass die als Prüfinstanz fungierende statistische Hypothese (hier also H_1) gültig ist, wenn die zu prüfende psychologische Hypothese nicht gilt, es also mindestens ein Individuum gibt, dessen individueller kausaler Effekt null

oder negativ ist. Wenn dies der Fall ist – die psychologische Hypothese also ungültig ist und die H_1 mit Wahrscheinlichkeit g gilt –, so sei ferner d die Wahrscheinlichkeit, dass die Effektgröße einen kritischen Wert erreicht oder überschreitet, der eine Teststärke (*power*) des statistischen Tests von mindestens $(1-\beta)$ garantiert. Entsprechend sei $1-d$ die komplementäre Wahrscheinlichkeit, dass die H_1 zwar zutrifft, die kritische Effektgröße aber nicht erreicht wird, sodass der Effekt in die so genannte Indifferenzzone des Tests fällt (Bredenkamp, 1972). In diesem Fall nimmt die Teststärke einen Wert von lediglich $1-\beta'$ an, der größer als α und kleiner als $1-\beta$ ist.

Alle drei Wahrscheinlichkeiten h, g und d werden natürlich einerseits durch Vorzeichen und Stärke der individuellen kausalen Effekte beeinflusst, zusätzlich aber auch durch viele methodische Aspekte der Studie, etwa durch die Anzahl der untersuchten Versuchspersonen, die Selektionswahrscheinlichkeiten $p(S = s)$ der einzelnen Individuen für die Untersuchungsstichprobe, das Zuteilungsverfahren der Versuchspersonen zu den Versuchsbedingungen, die als Prüfinstanz gewählte statistische Hypothese (z. B. Erwartungswerthypothese, Medianhypothese, vgl. dazu Iseler, 1996; Townsend, 1990) und vieles andere mehr. Sie sind folglich zunächst einmal unbekannt. Dennoch macht es Sinn, diese Wahrscheinlichkeiten einzuführen. Sie lassen sich durchaus als asymptotische relative Häufigkeiten frequentistisch interpretieren. Nehmen wir an, es lägen V verschiedene Vorschläge unterschiedlicher psychologischer Laboratorien vor, wie die statistische Prüfung einer psychologischen Hypothese der Form (1) genau erfolgen soll, das heißt wie groß die Stichprobe zu sein hat, welche Selektionswahrscheinlichkeiten für die Untersuchungsstichprobe gelten, wie die zu untersuchenden Probanden auf Experimentalbedingung und Kontrollbedingung aufgeteilt werden, welche statistische Hypothese als Überprüfungsinstanz dient et cetera. Eine Methodologie der Hypothesenprüfung kann dann dadurch repräsentiert werden, dass man den V verschiedenen methodischen Vorschlägen jeweils Realisationswahrscheinlichkeiten p_v, $v = 1, \ldots V$, zuordnet, wobei sich die Wahrscheinlichkeiten zu 1 aufsummieren müssen.[3] Je nach zufällig ausgewähltem Überprüfungsvorschlag kann dann die als Prüfinstanz gewählte statistische Hypothese gültig oder auch ungültig sein, auch wenn die psychologische Hypothese in immer gleicher Weise (d. h. mit gleichen individuellen kausalen Effekten) zutrifft. Die Wahrscheinlichkeit h entspräche dann dem relativen Anteil der Überprüfungs-

3 Der Extremfall einer »anarchistischen Methodologie« würde beispielsweise jedem Vorschlag v die gleiche Realisierungswahrscheinlichkeit $p_v = 1/V$ zuweisen, das heißt, wir geben jeder Methode die gleiche Chance, realisiert zu werden.

fälle, in dem die als Prüfinstanz gewählte H_1 nicht gilt. Analoges gilt, wenn die zu prüfende psychologische Hypothese in einer bestimmten Art und Weise falsch ist, weil beispielsweise eine bestimmte Teilpopulation von Individuen die behaupteten individuellen kausalen Effekte nicht zeigt. Auch hier können sich je nach ausgewählter Überprüfungsmethodik Fälle mit gültiger statistischer Hypothese und Fälle mit ungültiger statistischer Hypothese ergeben, weil sich zum Beispiel die Selektionswahrscheinlichkeiten für die kritische Teilpopulation der »hypothesenwidrigen« Individuen zwischen den V Vorschlägen unterscheiden können (Iseler, 1997). Die Wahrscheinlichkeit g entspräche dann dem Anteil der Überprüfungsfälle, in dem die als Prüfinstanz gewählte statistische Hypothese (H_1) gilt, obwohl die psychologische Hypothese nicht gilt.

Wir können nun die interessierenden Fehlerwahrscheinlichkeiten e und f, welche über die Validität einer Untersuchung entscheiden, als Funktionen von h, g und d sowie den statistischen Irrtumswahrscheinlichkeiten α, β und β' darstellen. Betrachten wir zunächst die Strenge $1-e$. Ist die als Prüfinstanz gewählte statistische Hypothese wie im Beispielfall eine statistische Alternativhypothese H_1, so wird die vorgeordnete psychologische Hypothese in einer Untersuchung für bewährt erklärt, wenn zugunsten der H_1 entschieden wird. Ist die vorgeordnete psychologische Hypothese faktisch ungültig, so kann dies auf drei Wegen geschehen:

1) Die als Prüfinstanz gewählte H_1 ist mit Wahrscheinlichkeit g dennoch gültig, die kritische Effektgröße wird mit der Wahrscheinlichkeit d erreicht oder überschritten und es wird mit der Wahrscheinlichkeit von mindestens $1-\beta$ zugunsten der H_1 entschieden;
2) die als Prüfinstanz gewählte H_1 ist mit Wahrscheinlichkeit g gültig, die kritische Effektgröße wird mit Wahrscheinlichkeit $1-d$ jedoch nicht erreicht und es wird mit der Wahrscheinlichkeit von lediglich $1-\beta'$ zugunsten der H_1 entschieden ($\alpha < 1-\beta' < 1-\beta$);
3) Die als Prüfinstanz gewählte H_1 ist mit Wahrscheinlichkeit $1-g$ unzutreffend – es gilt also H_0 – und es wird aufgrund eines α-Fehlers dennoch zugunsten von H_1 entschieden.

Addiert man alle drei Wahrscheinlichkeiten, so ergibt sich:

$$(5)\ e(H_1) = g \cdot d \cdot (1-\beta) + g \cdot (1-d) \cdot (1-\beta') + (1-g) \cdot \alpha$$
$$= \alpha + (1 - \alpha - \beta' - d \cdot (\beta - \beta')) \cdot g$$

Die Wahrscheinlichkeit fälschlicher Bewährungsurteile für eine psychologische Hypothese, die über eine H_1 geprüft wird, ist also eine linear steigen-

de Funktion von g mit α als unterer Grenze. Die Steigung dieser Funktion wächst mit dem Effektstärkeparameter d und der Teststärke. Wenn die kritische Effektgröße so klein gewählt wird, dass man die Indifferenzzone vernachlässigen kann ($d = 1$), dann ergibt sich der bei Erdfelder und Bredenkamp (1994) behandelte Spezialfall:

$$(6)\ e(H_1) = \alpha + (1 - \alpha - \beta) \cdot g$$

Auf völlig analoge Weise kann man für die Wahrscheinlichkeit f fälschlicher Nichtbewährungsurteile bei Verwendung einer H_1 als Prüfinstanz zeigen, dass

$$(7)\ f(H_1) = h \cdot (1 - \alpha) + (1-h) \cdot d \cdot \beta + (1-h) \cdot (1-d) \cdot \beta'$$
$$= \beta' + d \cdot (\beta - \beta') + (1 - \alpha - \beta' - d \cdot (\beta - \beta')) \cdot h.$$

Hierbei handelt es sich um eine zu (5) parallele linear steigende Funktion von h, jedoch mit der Wahrscheinlichkeit $\beta' + d \cdot (\beta - \beta')$ als unterer Grenze. Der bei Erdfelder und Bredenkamp diskutierte Spezialfall ergibt sich auch hier, wenn man annimmt, dass die Indifferenzzone vernachlässigt werden kann ($d = 1$):

$$(8)\ f(H_1) = \beta + (1 - \alpha - \beta) \cdot h.$$

Wird eine psychologische Hypothese nicht über eine statistische Alternativhypothese, sondern über eine Nullhypothese als Prüfinstanz getestet – beispielsweise die Hypothese eines ausschließlich linearen Trends über mehrere Versuchsgruppen hinweg –, dann entsprechen genau umgekehrt Entscheidungen für die Beibehaltung der H_0 den PH-Bewährungen, Entscheidungen gegen H_0 dagegen Entscheidungen gegen die psychologische Hypothese. Die statistischen Fehlerwahrscheinlichkeiten sind gegenüber den Gleichungen (5) bis (8) also lediglich zu vertauschen, sodass man für die beiden validitätsrelevanten Fehlerwahrscheinlichkeiten erhält:

$$(9)\ e(H_0) = \beta' + d \cdot (\beta - \beta') + (1 - \alpha - \beta' - d \cdot (\beta - \beta')) \cdot g$$

$$(10)\ f(H_0) = \alpha + (1 - \alpha - \beta' - d \cdot (\beta - \beta')) \cdot h.$$

Auch hierbei handelt es sich wieder um zu (5) und (7) parallele lineare Funktionen von g und h, jedoch nunmehr mit vertauschten Ordinatenab-

schnitten. Der bei Erdfelder und Bredenkamp (1994) behandelte Spezialfall ergibt sich erneut aus der Restriktion $d = 1$.

Deduktion methodologischer Regeln

Im Folgenden soll untersucht werden, welche methodologischen Regeln aus dem Validitätspostulat als erster methodologischer Regel auf der Grundlage der Gleichungen (5) bis (10) ableitbar sind und welche Konsequenzen dies in Bezug auf methodische Entscheidungen im Forschungsalltag nach sich zieht.

Bredenkamp und Erdfelder (1994) haben bereits betont, dass eine scheinbare methodische Selbstverständlichkeit auf dem Hintergrund des Validitätspostulats infrage zu stellen ist: Die Reduzierung die statistischen Fehlerwahrscheinlichkeiten α und β, oft als Routinemethode zur Erhöhung der statistischen Validität einer Hypothesenprüfung empfohlen (z. B. Hager u. Westermann, 1983a), folgt nicht ohne weiteres aus dem Validitätspostulat. Wie die Gleichungen (5) bis (10) zeigen, führt die Herabsetzung einer der beiden statistischen Fehlerwahrscheinlichkeit α oder $(\beta' + d \cdot (\beta - \beta'))$ für positives g und h ceteris paribus immer zur Reduktion genau einer der beiden Wahrscheinlichkeiten e oder f, während die andere sich erhöht. Validitätssteigerungen können also im Allgemeinen nicht durch isolierte Reduktion von α oder β bewirkt werden, sondern nur durch eine kombinierte Reduktion von α und β, die sicher stellt, dass weder e noch f durch diese Maßnahme erhöht werden. Wie in Anhang A gezeigt wird, ist dies aber nur möglich, wenn $g + h < 1$.

Die erste methodologische Regel, die aus dem Validitätspostulat abzuleiten ist, sollte garantieren, dass $g + h < 1$ in Anwendungen erfüllt ist, denn ohne diese Voraussetzung kann nicht einmal die Reduktion von α und β als Maßnahme zur Validitätssteigerung ausgewiesen werden. Fragen wir daher zunächst, ob methodische Maßnahmen vorstellbar sind, die eine bestimmte Obergrenze für die Wahrscheinlichkeit g einer gültigen statistischen Hypothese (als gewählter Prüfinstanz) bei ungültiger psychologischer Hypothese garantieren. Diese Frage ist definitiv zu verneinen, wenn der Standardfall vorliegt, dass die zu prüfende psychologische Hypothese eine allgemeinpsychologische Hypothese und die Prüfungsinstanz eine statistische Aggregathypothese ist. Um dies einzusehen, muss man sich lediglich klarmachen, dass es für die Falschheit einer allgemeinpsychologischen psychologischen Hypothese ja ausreicht, dass nur ein einziges Individuum der

relevanten Grundgesamtheit den in der psychologischen Hypothese postulierten Effekt nicht zeigt, während alle anderen Individuen sich hypothesenkonform verhalten. Tritt dieser Fall ein, so ist es also äußerst wahrscheinlich, dass die als Prüfinstanz dienende statistische Hypothese gilt, da mit sehr hoher Wahrscheinlichkeit nur solche Individuen in die Untersuchungsstichprobe gelangen, die sich PH-konform verhalten. Selbst wenn der unwahrscheinliche Fall eintreten sollte, dass das für die Falschheit der psychologischen Hypothese verantwortliche Individuum in die Stichprobe gelangt, ist der hypothesenkonträre individuelle kausale Effekt vermutlich nicht stark genug, um die hypothesenkonformen kausalen Effekte der anderen Versuchspersonen im Mittel aufzuwiegen. Dieses bedeutet aber nichts anderes, als dass g im ungünstigsten Fall beliebig nahe an $g = 1$ heranrücken kann, ohne dass erkennbar wäre, wie dies verhindert werden kann.

Berücksichtigt man das Ergebnis, dass Maßnahmen zur Kontrolle von g im Regelfall nicht existieren, bleibt nur eine einzige Konsequenz, die mit der geschilderten Forderung $g + h < 1$ vereinbar ist:

> *Regel 1 (Implikationsregel):*
> Wähle die als Prüfinstanz dienende SH, das Stichprobengewinnungsverfahren und das Zuteilungsverfahren zu den Versuchsbedingungen so, dass die PH die SH logisch impliziert, mit der Konsequenz, dass die Wahrscheinlichkeit h einer ungültigen SH bei gültiger PH immer den Wert $h = 0$ annehmen muss.

Man beachte, dass diese Regel nicht als Deskription der Forschungspraxis gemeint ist. Folglich ist die Frage, ob diese Regel die psychologische Forschung korrekt beschreibt oder nicht, nicht sinnvoll. Die Implikationsregel ist nichts anderes als die zwingende Konsequenz aus dem Validitätspostulat, welches der statistischen Prüfung psychologischer Hypothesen ein methodologisches Fundament gibt. Anders ausgedrückt: Ohne Maßnahmen, welche die Gültigkeit der Implikationsregel garantieren, ist nicht erkennbar, worin der Sinn der statistischen Prüfung psychologischer Hypothesen besteht. Diese für eine deduktivistische Methodologie in der Psychologie zentrale Einsicht wurde bereits von Bredenkamp (1972, S. 18) unmissverständlich formuliert: »Wir können uns nur eine einzige Möglichkeit vorstellen, die dem statistischen Test seinen Wert zur Prüfung allgemeinpsychologischer Hypothesen belässt: dass nämlich die Wahrheit einer solchen Hypothese die Richtigkeit einer statistischen Hypothese nach sich zieht (implikative Verbindung).«

Die Implikationsregel verlangt eine Reihe von Folgeregeln, die zusammengenommen deren Gültigkeit garantieren. Damit beispielsweise die psychologische Hypothese, dass alle individuellen kausalen Effekte positiv sind, die statistische H_1 eines positiven durchschnittlichen Effekts für die Erwartungswerte impliziert, muss eine von drei Nebenbedingungen erfüllt sein (vgl. Steyer, in diesem Band; Rubin, 1974):

1) Durch Anwendung der experimentellen Kontrolltechnik der Konstanthaltung (Bredenkamp, 1969) wurden alle Störvariableneinflüsse erfolgreich eliminiert, sodass alle individuellen Kontrollgruppenwerte auf der AV (a_s) gleich sind;
2) durch Anwendung der experimentellen Kontrolltechnik der Parallelisierung (Bredenkamp, 1969) werden für homogene Paare (bei mehr als zwei Versuchsbedingungen: Subgruppen) von Versuchspersonen homogene a_s-Werte garantiert, die dann jeweils unterschiedlichen Versuchsbedingungen zugeteilt werden;
3) durch Anwendung der Kontrolltechnik der Randomisierung (Bredenkamp, 1969) wird sichergestellt, dass der Erwartungswert der a_s-Werte für Versuchspersonen in der Kontroll- und in der Experimentalbedingung gleich ist.

Die ersten beiden Kontrolltechniken garantieren, wenn sie wirklich alle relevanten Störeinflüsse auf die a_s-Werte berücksichtigen, dass die Differenz der Erwartungswerte der Versuchsgruppen dem durchschnittlichen individuellen kausalen Effekt *in jeder Stichprobe* entspricht, während die Kontrolltechnik der Randomisierung sicherstellt, dass die Differenz der Erwartungswerte der Versuchsgruppen dem Erwartungswert der individuellen kausalen Effekte entspricht (Rubin, 1974). Wenn gemäß der psychologischen Hypothese alle individuellen kausalen Effekte positiv sind, muss dies bei Anwendung einer der oben genannten Kontrolltechniken folglich auch für die Erwartungswertdifferenz der Versuchsgruppen gelten. Wenn die psychologische Hypothese Nulleffekte oder negative kausale Effekte für jedes Individuum behauptet, hat das analoge Konsequenzen für die Erwartungswertdifferenz. Mit anderen Worten: Durch Anwendung von experimentellen Kontrolltechniken wie Konstanthaltung und Parallelisierung (wenn die relevanten Störvariablen bekannt sind) oder Randomisierung (wenn nicht alle relevanten Störeinflüsse konstant gehalten oder ausbalanciert werden können) kann bei geeigneter Wahl der statistischen Hypothese erreicht werden, dass die psychologische die statistische Hypothese impliziert, sodass $h = 0$.

Regel 2 (Kontrolltechnikregel):
Verwende die Kontrolltechnik der Konstanthaltung oder der Parallelisie-

rung bei bekannten Störeinflüssen beziehungsweise die Technik der Randomisierung bei teilweise unbekannten Störeinflüssen, um eine Implikationsbeziehung zwischen den Erwartungswerthypothesen PH und SH sicherzustellen.

Lässt sich eine analoge Regel auch für den Fall formulieren, dass individuelle und durchschnittliche kausale Effekte nicht mit Bezug auf Erwartungswertdifferenzen, sondern mit Bezug auf Mediandifferenzen formuliert werden? Iseler (1996) hat gezeigt, dass dies nicht der Fall ist: Ist die individuelle Mediandifferenz zwischen einer Experimental- und einer Kontrollbedingung für jedes Individuum einer Grundgesamtheit positiv, so muss dasselbe auch bei Anwendung der Randomisierungstechnik nicht zwingend für die über Individuen aggregierte Mediandifferenz zwischen beiden Versuchsbedingungen gelten. Mit anderen Worten: Rangordnungshypothesen über Verteilungsmediane sind anders als Rangordnungshypothesen über Erwartungswerte nicht zwangsläufig aggregationsstabil, selbst wenn durch experimentelle Kontrolltechniken stochastische Unabhängigkeit zwischen Störvariablen und Versuchsbedingungen hergestellt wird. Wir können festhalten:

Regel 3 (Hypothesenwahlregel):
Meide statistische Aggregathypothesen, für die auch bei Anwendung angemessener experimenteller Kontrolltechniken Aggregationsstabilität nicht nachgewiesen werden kann. Meide aus diesem Grund unter anderem Medianhypothesen (Iseler, 1996).

Wenn durch geeignete Wahl der als Prüfungsinstanz fungierenden statistischen Hypothese und der zu verwendenden experimentellen Kontrolltechniken die Implikationsregel garantiert ist, vereinfachen sich die Formeln für PH-Fehlbeurteilungswahrscheinlichkeiten bei Überprüfung über eine H_1 zu

$$(11)\ e(H_1) = \alpha + (1 - \alpha - \beta' - d \cdot (\beta - \beta')) \cdot g$$

$$(12)\ f(H_1) = \beta' + d \cdot (\beta - \beta')$$

und bei Überprüfung über eine H_0 zu

$$(13)\ e(H_0) = \beta' + d \cdot (\beta - \beta') + (1 - \alpha - \beta' - d \cdot (\beta - \beta')) \cdot g$$

$$(14)\ f(H_0) = \alpha.$$

Das Problem der Kontrolle von *f* und damit der Fairness einer Untersuchung reduziert sich unter diesen Bedingungen also auf die Kontrolle der Typ 1- und Typ 2-Fehlerwahrscheinlichkeiten beim statistischen Testen, die keine prinzipiellen Schwierigkeiten bereitet (s. dazu Erdfelder et al., in diesem Band). Die Typ 1-Fehlerwahrscheinlichkeit α steht dabei im Vordergrund, wenn eine H_0 als Prüfungsinstanz dient, die Typ 2-Fehlerwahrscheinlichkeit β, wenn eine H_1 als Prüfungsinstanz dient (s. bereits Bredenkamp, 1972):

> *Regel 4 (Fehlerwahrscheinlichkeitsregel):*
> Ist die Implikationsregel erfüllt, realisiere eine möglichst kleine Typ 2-Fehlerwahrscheinlichkeit $\beta' + d \cdot (\beta - \beta')$, wenn eine H_1 als Prüfinstanz fungiert und eine möglichst kleine Typ 1-Fehlerwahrscheinlichkeit α, wenn eine H_0 als Prüfinstanz fungiert, um die Fairness zu optimieren. Durch Reduktion des jeweils anderen Fehlerwahrscheinlichkeitstyps (α bei H_1 als Prüfinstanz und $\beta' + d \cdot (\beta - \beta')$ bei H_0 als Prüfinstanz) kann ferner die Strenge optimiert werden.

Unbefriedigend ist, dass die Strenge zwar durch Optimierung der Typ 1- oder Typ 2-Irrtumswahrscheinlichkeiten erhöht werden kann, jedoch in Anbetracht der unbekannten *g*-Wahrscheinlichkeit dennoch nur eine sehr hohe Obergrenze von $(1 - \beta' - d \cdot (\beta - \beta'))$ (Gleichung 11) beziehungsweise $(1-\alpha)$ (Gleichung 13) für die Wahrscheinlichkeit *e* fälschlicher PH-Bewährungsurteile garantiert werden kann. Diese Obergrenze wird im Extremfall $g = 1$ erreicht.

Jede Maßnahme, die nachweislich *g* reduziert, trägt folglich dazu bei, dass α- und β-Reduktionen nicht nur die Fairness, sondern auch die Strenge beeinflussen. Stehen mehrere statistische Hypothesen als potenzielle Prüfinstanzen zur Verfügung, die untereinander in Implikationsbeziehungen zueinander stehen, so hilft folgende Regel weiter:

> *Regel 5 (Strengeregel):*
> Stehen zwei Hypothesen SH_1 und SH_2 als Prüfinstanzen zur Auswahl und gilt $SH_1 \Rightarrow SH_2$ während die Umkehrung nicht gilt, so wähle die strengere Hypothese SH_1 als Prüfinstanz, welche die Konkurrenzhypothese impliziert.

Die Rechtfertigung für die Strengeregel ergibt sich aus dem in Anhang B belegten Satz, dass $SH_1 \Rightarrow SH_2$ die Relation $g_1 < g_2$ impliziert. Damit ist die Wahl der strengeren Hypothese ein Beitrag zur Minimierung von *e* und somit auch zur Maximierung der Validität einer Untersuchung. Auch diese

Regel wurde bereits zuvor von Bredenkamp als zentrales Element einer deduktivistischen Methodologie herausgestellt, sehr deutlich zum Beispiel in seiner Kritik (Bredenkamp, 1984) an Hager und Westermann (1983a).

Mit der Darstellung von fünf wichtigen Konsequenzen des Validitätspostulats konnte der methodologische Gehalt dieses Prinzips in dieser Arbeit nur angedeutet werden. Zahlreiche weitere methodologische Regeln lassen sich daraus gewinnen, von denen ein kleiner Teil bei Erdfelder und Bredenkamp (1994) dargestellt ist. Deutlich geworden sein sollte, dass sich wichtige Elemente einer deduktivistischen Forschungsmethodologie aus dem Validitätspostulat stringent herleiten lassen, die sich nahtlos in die Prinzipien des methodologischen Falsifikationismus (Popper, 1934/1982; Lakatos, 1970) einfügen, wenn man sie um die statistische Entscheidungstheorie von Neyman und Pearson ergänzt.

Logische Probleme? Eine Antwort auf Westmeyer (1998)

Westmeyer (1998) hat die Arbeit von Erdfelder und Bredenkamp (1994), auf der das vorliegende Kapitel aufbaut, kritisiert. Seine Kritik betrifft vier Punkte, die im Folgenden der Reihe nach behandelt werden.

Sind Popper, Lakatos sowie Neyman und Pearson vereinbar?

Zunächst sieht Westmeyer (1998) in der von Erdfelder und Bredenkamp (1994) vertretenen Methodologie »ein Amalgam aus wenigstens drei Ansätzen« (Neyman-Pearson-Theorie, Poppers Falsifikationstheorie und der Methodologie wissenschaftlicher Forschungsprogramme nach Lakatos), deren Kompatibilität »zumindest bezweifelt werden kann« (Westmeyer, 1998, S. 258). Diese Kritik steht zunächst einmal im Widerspruch zur Position von Lakatos, der in dem »Amalgam« eine zwingende Konsequenz des methodologischen Falsifikationismus sieht: »Probabilistic theories may qualify now as ›scientific‹: although they are not falsifiable they can be easily made ›falsifiable‹ ... by specifying certain rejection rules which may make statistically interpreted evidence ›inconsistent‹ with the probabilistic theory.... Indeed, this methodological falsificationism is the philosophical basis of some of the most interesting developments in modern statistics. The Neyman-Pearson approach rests completely on methodological falsificationism« (Lakatos, 1970, S. 109).

Die behauptete Inkompatibilität zwischen Lakatos sowie Neyman und Pearson ist auf dem Hintergrund dieses Zitats schwer nachzuvollziehen. Auch Popper (1934/1982, Abschnitt 68) hat – allen Polemiken zwischen ihm und Lakatos zum Trotz – Ideen geäußert, die als Vorwegnahme der oben zitierten Position von Lakatos zu deuten sind: »Wir schließen deshalb die unbeschränkte Verwendung der Wahrscheinlichkeitsaussagen aus – durch den methodologischen Beschluß, Effekte . . . niemals auf gehäufte Zufälle zurückzuführen« (Popper, 1934/1982, S. 153). Diese und weitere Textpassagen, die man nennen könnte, machen deutlich, dass die Verknüpfung einiger grundlegender Ideen von Popper, Lakatos, Neyman und Pearson – soweit sie sich auf die Verwendung von statistischen Hypothesen und die Entscheidung über statistische Hypothesen beziehen – nicht nur möglich ist, sondern auch nahe liegend erscheint. Dass man sicher auch Aussagen finden kann, die auf Meinungsverschiedenheiten speziell zwischen Popper und Lakatos schließen lassen, ändert an der prinzipiellen Vereinbarkeit der genannten Ideen nichts.

Ist die Implikationsregel realistisch?

Westmeyer (1998, S. 261 f.) bezweifelt zweitens die Herstellbarkeit von Implikationsbeziehungen zwischen psychologischen Hypothesen und statistischen Hypothesen. Es ist ausreichend, auf diesen Kritikpunkt mit einer Existenzbehauptung zu reagieren: Oben wurde bereits eine Klasse von psychologischen Hypothesen und zugehörigen statistischen Hypothesen vorgestellt, für welche die angeblich nicht herstellbare Beziehung PH $\Rightarrow$ SH immer dann zutrifft, wenn geeignete experimentelle Kontrolltechniken realisiert werden. Dieses Beispiel – neben vielen anderen, die man nennen könnte – widerlegt die prinzipielle Unmöglichkeit einer logischen Brücke zwischen psychologischer Hypothese und statistischer Hypothese. Richtig ist zweifellos, dass eine logische Verknüpfung nicht in jedem Fall und vor allem nicht immer problemlos gelingt. Das schmälert jedoch nicht den Wert der Methodologie für die große Klasse von psychologischen Hypothesen, bei denen der skizzierte Weg beschritten werden kann.

Sind statistische Aggregathypothesen prinzipiell sinnlos?

Ein dritter Kritikpunkt Westmeyers (1998) betrifft die Ebene der Hypothesenprüfung. Überprüfungen von allgemeinpsychologischen Hypothesen anhand von statistischen Hypothesen, die für Aggregate formuliert werden,

hält er für »sinnlos« (Westmeyer, 1998, S. 261). Zu entgegnen ist, dass die hier vertretene Methodologie eine Hypothesenprüfung auf Aggregatebene keineswegs zwingend fordert. Im Gegenteil: Wo Überprüfungen allgemeinpsychologischer Hypothesen an einzelnen Individuen praktikabel sind und wo sie die Validität der Untersuchung und der statistischen Analyse nicht gefährden, sollten sie unbedingt durchgeführt werden (vgl. u. a. Bredenkamp, 1972, 1980, 1990; Bredenkamp et al., 1988). Wenn die individuelle Prüfung dagegen unter Praktikabilitäts- oder Validitätsgesichtspunkten nicht möglich ist, muss wohl oder übel auf die Aggregatebene übergegangen werden (s. auch Iseler, 1997). Dieser Übergang ist – wie gezeigt – dann nicht sinnlos, wenn die allgemeinpsychologische Hypothese eine Aggregathypothese logisch impliziert: Die Verwerfung der Letzteren impliziert dann nämlich die Verwerfung der Ersteren. Insofern genügt der Ansatz der grundlegenden Forderung, dass eine empirische Kritik psychologischer Hypothesen prinzipiell möglich sein muss (Herrmann, 1979). Dass ein analoger Schluss vom Aggregat auf das einzelne Individuum bei Hypothesenbewährungen nicht gilt, ist unbestritten. Das Verifikationsproblem ist bekanntlich ein fundamentales Problem bei der Überprüfung universeller Aussagen und für die hier vertretene Methodologie keineswegs spezifisch.

Erzeugt der Verzicht auf einen Falsifikationsautomatismus logische Probleme?

Der explizite Verzicht auf einen Falsifikationsautomatismus bei Ergebnissen, die gegen die aus einer psychologischen Hypothese abgeleitete statistische Hypothese sprechen, bildet den vierten Kritikpunkt Westmeyers (1998). Erdfelder und Bredenkamp (1994, S. 609) haben diesbezüglich richtigerweise festgestellt, dass ein Ergebnis, welches für die Verneinung der aus der psychologischen Hypothese deduzierten statistischen Hypothese spricht, »notwendig aber nicht hinreichend« für die Verwerfung der vorgeordneten psychologischen Hypothese ist. Westmeyer (1998) argumentiert, dass diese Position auf dem Hintergrund einer Implikationsbeziehung zwischen vorgeordneter psychologischer Hypothese und geprüfter statistischer Hypothese logisch unhaltbar ist:[4] »Plötzlich soll auch die Verneinung von SH nur eine notwendige Bedingung für die Verneinung (Falsifikation) von PH sein ... Diese Bescheidenheit hat aber Konsequenzen. Zunächst einmal

4 Dieses Argument wurde auch von Hager (1999) geteilt. Meine Antwort auf Westmeyer (1998) betrifft also auch Hager (1999).

steht sie im Widerspruch zu ›Wenn PH, dann SH‹. Durch Kontraposition folgt daraus nämlich ›Wenn nicht SH, dann nicht PH‹, d. h. die Verneinung von SH ist eine hinreichende Bedingung für die Verneinung von PH« (Westmeyer, 1998, S. 259).

Dieser Analyse und weiteren Schlussfolgerungen, die Westmeyer daraus ableitet, liegt ein Missverständnis zugrunde. Natürlich ist »ein Ergebnis, das für die Verneinung der implizierten SH spricht« (Erdfelder u. Bredenkamp, 1994, S. 609) – anders als Westmeyer (1998) und auch Hager (1999) offenbar annehmen – mit der Negation der statistischen Hypothese nicht logisch äquivalent. Wäre dies der Fall, dann hätte Westmeyer Recht: Legt man den Wahrheitswert »falsch« für die statistische Hypothese fest, erzwingt dies via modus tollens die Falsifikation der vorgeordneten psychologischen Hypothese. Tatsächlich ist jedoch »ein Ergebnis, das für die Verneinung der implizierten SH spricht« immer ein bestimmter Wert einer Teststatistik, bei Gruppenvergleichen zum Beispiel eine *t*- oder *F*-Statistik. Wie auch immer der Wert dieser Teststatistik ausfällt: Eine Festlegung des Wahrheitswerts der geprüften statistischen Hypothese ist in keinem Fall logisch impliziert. Weder die Anerkennung noch die Negation der statistischen Hypothese kann durch Daten erzwungen werden, schon gar nicht durch Daten einer einzelnen empirischen Untersuchung. Ob man in Anbetracht eines Prüfstatistik-Werts, der in den Rejektionsbereich einer statistischen Hypothese fällt, die geprüfte statistische Hypothese für falsch hält, ist daher vernünftigerweise von Randbedingungen abhängig zu machen, beispielsweise von der Stabilität des entsprechenden Befunds in Replikationsstudien, von etwaigen weiteren »belastenden« Befunden zur gleichen psychologischen Hypothese, von der Verfügbarkeit alternativer psychologischer Hypothesen, welche die gleichen Befunde besser erklären können und so weiter (vgl. Bredenkamp, 1980). Aus diesem Grunde ist »ein Ergebnis, das für die Verneinung der implizierten SH spricht«, im Rahmen eines methodologischen Falsifikationismus selbstverständlich nicht hinreichend für die Falsifikation der vorgeordneten psychologischen Hypothese. Notwendig für eine Falsifikation ist ein solcher Befund allerdings, weil wir Hypothesen niemals »grundlos« verwerfen wollen, das heißt niemals ohne mindestens ein belastendes empirisches Resultat. Dadurch wird vernünftigerweise ausgeschlossen, dass etwa die vermeintliche Unplausibilität eines theoretischen Ansatzes ohne ein einziges belastendes empirisches Resultat als Grund für die Zurückweisung einer Theorie ausreicht.

Man mag einwenden, dass die Unmöglichkeit, Hypothesenfalsifikationen allein durch Daten zu erzwingen, der Preis ist, den man für die Einführung statistischer Hypothesen in den Hypothesenprüfungsprozess zu bezahlen

hat. Dieses Argument beruht jedoch auf einer unkritischen Sicht des Falsifikationsmechanismus bei rein deterministischen Theorien. Auch dort setzt die Anerkennung falsifizierender »Protokollsätze« oder »Basissätze« konventionalistische Elemente voraus. Niemals ist die Übertragung von Beobachtungen in »Protokollsätze« oder »Basissätze« ein Automatismus, der ohne methodologische Konventionen auskommt. Schon Popper (1934/1982, S. 71) sagte ganz klar: »Die Basissätze werden durch Beschluss, durch Konvention anerkannt, sie sind Festsetzungen.« Lakatos (1970, S. 99) stellte noch allgemeiner fest: »No factual proposition can ever be proved from an experiment. Propositions can only be derived from propositions, they cannot be derived from facts.« Folglich ergeben sich in einer deduktivistischen Methodologie der statistischen Prüfung psychologischer Hypothesen keine grundsätzlich neuen Probleme, die durch deterministische Theorien, welche anhand von Basissätzen sensu Popper (1934/1982) geprüft werden, vermeidbar wären.

Fazit

Die vorstehenden Abschnitte haben gezeigt, dass die Forderung nach »strengen« beziehungsweise »validen« statistischen Prüfungen wissenschaftlicher Hypothesen im Sinne von Hager und Westermann (u. a. 1983a) eine Reihe sinnvoller methodologischer Regeln impliziert und damit eine stringente Begründung für methodische Maßnahmen liefert, die häufig nur intuitiv für plausibel gehalten werden. Der »methodologische Gehalt« des Validitätspostulats wurde in der vorliegenden Arbeit in Form von fünf Regeln nur angedeutet; weitergehende Ableitungen sind möglich, die unter anderem die Begründung der Wahl einer Effektstärke beim Signifikanztest, das Problem »Skalenniveau und Wahl der Prüfstatistik«, lineare Modelle mit fixierten Faktoren und Zufallsfaktoren, die Entscheidung zwischen »strengen« und »schwachen« Prüfstrategien bei multiplen Signifikanztests, die fragliche Bedeutung der Stichprobeneffektstärke, das Problem von Variablentransformationen zur Sicherstellung von Verteilungsvoraussetzungen und Indikationen sowie Kontraindikationen für nonparametrische statistische Verfahren betreffen. Insgesamt erscheint die Einschätzung gerechtfertigt, dass sich die Forderung nach kleinen Fehlerwahrscheinlichkeiten e und f als »oberste methodologische Regel« beziehungsweise als »Regel von höherem Typus« (Popper, 1934/1982, S. 26) im Rahmen einer deduktivistischen Methodologie der statistischen Prüfung wissenschaftlicher Hypothesen eignet.

Literatur

Bredenkamp, J. (1969). Experiment und Feldexperiment. In C.F. Graumann (Hg.), *Handbuch der Psychologie, Band 7: Sozialpsychologie. 1. Halbband: Theorien und Methoden* (S. 332–374). Göttingen: Hogrefe.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1984). Anmerkungen und Korrekturen zu Hager & Westermann: Entscheidung über statistische und wissenschaftliche Hypothesen: Probleme bei mehrfachen Signifikanztests zur Prüfung einer wissenschaftlichen Hypothese. *Zeitschrift für Sozialpsychologie, 15,* 224–229.

Bredenkamp, J., Klein, K.-M., Hayn, S. v., Vaterrodt, B. (1988). Gedächtnispsychologische Untersuchungen eines Rechenkünstlers. *Sprache & Kognition, 7,* 69–83.

Bredenkamp, J. (1990). Kognitionspsychologische Untersuchungen eines Rechenkünstlers. In H. Feger (Hg.), *Wissenschaft und Verantwortung. Festschrift für Karl Josef Klauer* (S. 47–70). Göttingen: Hogrefe.

Bridgeman, P.W. (1938). Operational analysis. *Philosophy of Science, 5,* 114–131.

Carnap, R. (1936). Testability and meaning, Part 1. *Philosophy of Science, 3,* 418–471.

Carnap, R. (1937). Testability and meaning, Part 2. *Philosophy of Science, 4,* 2–40.

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann und W.H. Tack (Hg.), *Methodologische Grundlagen der Psychologie.* (Enzyklopädie der Psychologie, Themenbereich B, Serie I, Band 1, S. 604–648). Göttingen: Hogrefe.

Gadenne, V. (1976). *Die Gültigkeit psychologischer Untersuchungen.* Stuttgart: Kohlhammer.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung.* Tübingen: Mohr.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer u. M.S. Morgan (Hg.), *The probabilistic revolution. Vol. II: Ideas in the sciences* (S. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren u. C. Lewis (Hg.), *A handbook for data analysis in the behavioral sciences* (S. 311–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Murray, D.J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen der Psychologie. In G. Lüer (Hg.), *Allgemeine Experimentelle Psychologie* (S. 43–264). Stuttgart: Fischer.

Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen.* Göttingen: Hogrefe.

Hager, W. (1999). About some misconceptions and the discontent with statistical tests in psychology. *Methods of Psychological Research – Online, 4,* 1–22

Hager, W., Westermann, R. (1983a). Entscheidung über statistische und wissenschaftliche Hypothesen: Probleme bei mehrfachen Signifikanztests zur Prüfung einer wissenschaftlichen Hypothese. *Zeitschrift für Sozialpsychologie, 14,* 106–117.

Hager, W., Westermann, R. (1983b). Zur Wahl und Prüfung statistischer Hypothesen in psychologischen Untersuchungen. *Zeitschrift für experimentelle und angewandte Psychologie, 30,* 67–94.

Herrmann, T. (1976). *Die Psychologie und ihre Forschungsprogramme.* Göttingen: Hogrefe.

Herrmann, T. (1979). *Psychologie als Problem. Herausforderungen der psychologischen Wissenschaft.* Stuttgart: Klett-Cotta.

Iseler, A. (1996). A paradoxical property of aggregate hypotheses referring to the order of medians. *Methods of Psychological Research – Online, 1,* 1–17.

Iseler, A. (1997). Populationsverteilungen von Merkmalen und Geltungsbereiche individuenbezogener Aussagen als Gegenstand der Inferenzstatistik in psychologischen Untersuchungen. In H. Mandl (Hg.), *Bericht über den 40. Kongress der Deutschen Gesellschaft für Psychologie in München 1996* (S. 699–708). Göttingen: Hogrefe.

Iseler, A. (1999). *Explikative und deduktive Komponenten der »Herleitung« statistischer Vorhersagen aus psychologischen Hypothesen.* Referat auf der 4. Tagung der Fachgruppe Methoden der Deutschen Gesellschaft für Psychologie in Leipzig, 25.9.1999.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos u. A. Musgrave (Hg.), *Criticism and the growth of knowledge* (S. 91–195). London: Cambridge University Press.

McCorquodale, K., Meehl, P.E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95–107.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Popper, K.R. (1982). *Logik der Forschung* (7. Aufl.). Tübingen: Mohr (Originalausgabe: 1934).

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle.* Stuttgart: Fischer.

Steyer, R., Gabler, S., von Davier, A.A., Nachtigall, C., Buhl, T. (2000a). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research – Online, 5,* 39–71.

Steyer, R., Gabler, S., von Davier, A.A., Nachtigall, C. (2000b). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research – Online, 5,* 55–86.

Townsend, J.T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin, 108,* 551–567.

Westermann, R. (1987). *Strukturalistische Theorienkonzeption und empirische Forschung in der Psychologie. Eine Fallstudie.* Berlin: Springer-Verlag.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik.* Göttingen: Hogrefe.

Westermann, R., Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11,* 117–146.

Westmeyer, H. (Hg.) (1989). *Psychological theories from a structuralist point of view.* New York: Springer-Verlag.

Westmeyer, H. (1992). *The structuralist program in psychology: Foundations and applications.* Toronto: Hogrefe & Huber.

Westmeyer, H. (1998). Psychologische Methoden und soziale Konventionen. In K.C. Klauer und H. Westmeyer (Hg.), *Psychologische Methoden und soziale Prozesse* (S. 250–266). Lengerich: Pabst.

Anhang A

Satz:
Validitätserhöhungen können durch Reduktion von α und β nur dann erreicht werden, wenn $g + h < 1$.

Beweis:
Betrachten wir den durch die Gleichungen (9) und (10) beschriebenen Fall, dass eine psychologische Hypothese über eine H_0 geprüft werden soll. Wenn wir α um den Betrag $\Delta(\alpha)$ verringern ($0 < \Delta(\alpha) < \alpha$), um f zu senken, können wir aus Gleichung (9) errechnen, um welchen Betrag $\Delta(\beta)$ die Irrtumswahrscheinlichkeit $\beta' + d \cdot (\beta - \beta')$ reduziert werden muss, damit e nicht steigt. Dies ist der Fall, wenn

(A1) $\beta' + d \cdot (\beta - \beta') + (1 - \alpha - (\beta' + d \cdot (\beta - \beta'))) \cdot g = (\beta' + d \cdot (\beta - \beta') - \Delta(\beta)) + (1 - (\alpha - \Delta(\alpha)) - (\beta' + d \cdot (\beta - \beta') - \Delta(\beta))) \cdot g$.

Elementare Termumformung liefert die Lösung

(A2) $\Delta(\beta) = g / (1-g) \cdot \Delta(\alpha)$

Die Reduktion von β darf die durch α-Senkung erreichte Verringerung von f aber nicht wieder zunichte macht. Das ist genau dann der Fall, wenn

(A3) $(\alpha - \Delta(\alpha)) + (1 - (\alpha - \Delta(\alpha)) - ((\beta' + d \cdot (\beta - \beta')) - \Delta(\beta))) \cdot h$
$< \alpha + (1 - \alpha - (\beta' + d \cdot (\beta - \beta'))) \cdot h$

$\Leftrightarrow 1 > h + g/(1-g) \cdot h$
$\Leftrightarrow h + g < 1$, q.e.d.

Der Beweis lässt sich analog führen, wenn man durch Reduktion von α und $\beta' + d \cdot (\beta - \beta')$ die Wahrscheinlichkeit e senken will, ohne f zu erhöhen.

Anhang B

Satz:
Wenn $SH_1 \Rightarrow SH_2$, dann $g_1 < g_2$

Beweis:
Nach Voraussetzung gilt
Prämisse 1: $p(SH_2|SH_1 \text{ und non-PH}) = 1$
Prämisse 2: $p(SH_1|SH_2 \text{ und non-PH}) = c; 0 < c < 1$

Behauptung:
$g_1 := p(SH_1|\text{non-PH}) < g_2 := p(SH_2|\text{non-PH})$

Aus der Definition von g_1 und g_2 folgt:
$g_1 = p(SH_1 \text{ und } SH_2|\text{non-PH}) + p(SH_1 \text{ und non-}SH_2|\text{non-PH})$
$g_2 = p(SH_1 \text{ und } SH_2|\text{non-PH}) + p(\text{non-}SH_1 \text{ und } SH_2|\text{non-PH})$

Folglich gilt für die Differenz:
$g_2 - g_1 = p(\text{non-}SH_1 \text{ und } SH_2|\text{non-PH}) - p(SH_1 \text{ und non-}SH_2|\text{non-PH})$
$= p(\text{non-}SH_1 \text{ und } SH_2|\text{non-PH}) - 0$ (nach Prämisse 1)
$= p(\text{non-}SH_1 \text{ und } SH_2 \text{ und non-PH}) / p(\text{non-PH})$
$= p(\text{non-SH1}|\text{SH2 und non-PH}) \cdot p(\text{SH2 und non-PH}) / p(\text{non-PH})$
$= (1-c) \cdot g_2$
$\Leftrightarrow g_2 \cdot (1-(1-c)) = g_1$
$\Leftrightarrow g_2 \cdot c = g_1$
$\Leftrightarrow g_2 > g_1$, q.e.d.

Teil II:
Methoden

Rolf Steyer

Was wollen und was können wir durch empirische Kausalforschung erfahren?[1]

Zusammenfassung

Anhand zweier Beispiele wird zunächst gezeigt, dass und warum die Analyse von Mittelwerten und ihren Unterschieden zwischen Gruppen zu kausalen Fehlschlüssen führen kann. Danach werden die Grundbegriffe der Theorie individueller und durchschnittlicher kausaler Effekte eingeführt. Anhand eines Beispiels zur nonorthogonalen Varianzanalyse wird dann demonstriert, dass schon diese elementaren Begriffe genügen, um ein altes, bisher nicht gelöstes Problem der Angewandten Statistik zu lösen. Schließlich wird die Theorie um den Begriff der Unkonfundiertheit ergänzt, der insbesondere – aber nicht nur – für die kausale Analyse von nichtexperimentellen Daten relevant ist.

Einleitung

Empirische Kausalforschung ist in der Psychologie und ihren Nachbardisziplinen allgegenwärtig. Psychologen wollen zum Beispiel wissen, ob und wie ein spezielles Therapieverfahren einem anderen bei einer bestimmten Störungsform überlegen ist oder ob eine bestimmte Unterrichtsform gegenüber einer anderen zu besseren Resultaten hinsichtlich eines bestimmten Lernkriteriums führt. Politiker interessiert, ob eine sozialpolitische Maßnahme zur gewünschten Reduktion bestimmter Gewalterscheinungen beiträgt, ob Einsparungen im Haushalt des Bundes oder ob eine bestimmte Steuerreform zur Verminderung des Defizits beitragen. Sportwissenschaftler fragen, ob ein bestimmtes Training in einer bestimmten Sportart zu höheren Leistungen führt als ein anderes Training. Umweltwissenschaftler untersuchen, ob von Menschen hervorgerufener Kohlendioxidausstoß zur Erderwärmung beiträgt. Mediziner wollen wissen, ob Rauchen zu Lungenkrebs, fettreiche Nah-

1 Überarbeitete Fassung meiner Antrittsvorlesung an der Friedrich-Schiller-Universität Jena am 19. November 1997.

rung und mangelnde Bewegung zu Herzinfarkt führen etc. Ohne kausale Theorien über die Wirkungen unseres Handelns wären wir handlungsunfähig. Handeln – im Gegensatz zu Verhalten – setzt Absicht und diese zumindest eine implizite Theorie oder Erwartung über den Effekt des Handelns voraus. Auch im Bereich des menschlichen Lernens gewinnt die Sichtweise des *kausalen Lernens* gegenüber der des puren assoziativen Lernens zunehmend an Bedeutung (siehe z. B. Waldmann, 2000, 2001).

Empirische Kausalforschung ist überall dort angesagt, wo wir unsicher über die Wirkung bestimmter Handlungsalternativen sind. Die Beispiele zeigen: Korrekte kausale Theorien sind lebens-, wenn nicht sogar überlebenswichtig für einen Staat, für soziale Gruppen und für einzelne Individuen, im privaten wie im professionellen Handeln.[2]

Entsprechend seiner Bedeutsamkeit hat das Thema »Kausalität« viele Wissenschaftsphilosophen, Statistiker, Methodiker und Einzelwissenschaftler angezogen. Auch wenn es an dieser Stelle nicht möglich ist, einen Überblick über die verschiedenen Kausalitätstheorien zu geben, der diesen Namen verdienen würde (siehe dazu eher Steyer, 1992, Kap. 2), sollen doch einige Autoren und die zugehörigen Stichworte in Erinnerung gerufen werden.

Wie es kaum anders sein könnte, hat Aristoteles erste wichtige Unterscheidungen eingeführt. Auch Humes (1739/1888) Charakterisierung einer kausalen Beziehung durch »Kontiguität«, »zeitliche Aufeinanderfolge« und »konstante Verbindung« ist als Markstein in der Ideengeschichte des Kausalitätsbegriffs zu nennen. Noch einschlägiger sind wohl die Millschen (1843/1862) Methoden der wissenschaftlichen Forschung, insbesondere die *Unterschiedsmethode*: die Variation der möglichen Ursache bei Konstanthaltung aller anderen Einflussvariablen und Feststellung der Kovariation der Ursache mit der zu erklärenden Variablen. Hier sind wir schon ganz dicht an der heute üblichen experimentellen Vorgehensweise.

Die moderne nichtexperimentelle Kausalforschung basiert im Wesentlichen auf dem Prinzip der *statistischen Konstanthaltung* der Störvariablen. Dieses Prinzip wurde in einer Vielzahl von Arbeiten im Rahmen der Literatur zur Pfadanalyse (z. B. des Genetikers Wright, 1918, 1921, 1923, 1934, 1960a, 1960b) und zu Strukturgleichungsmodellen (Bollen, 1989; Jöreskog, 1993) ausgearbeitet und verfeinert. Darüber hinaus wurde dort die Unterscheidung zwischen direkten, indirekten und totalen Effekten eingeführt

2 Wenn ich hier nicht zögere, von »überlebenswichtig« zu sprechen, denke ich an die großen politischen Katastrophen, die wir Deutsche und Europäer im vergangenen Jahrhundert erlebt haben.

(siehe dazu weitere Angaben in Steyer, 1992, Kap. 2). Seit einigen Jahren wurden auch wichtige Fortschritte im Bereich kausaler Modelle mit qualitativen Variablen erzielt (siehe z. B. Pearl, 1993, 1995, 2000).

Nicht unerwähnt bleiben darf in diesem Kontext die in der Erziehungswissenschaft und der Psychologie populäre *Theorie interner Validität* von Campbell und Stanley (1963) sowie Cook und Campbell (1979, 1986), die im deutschsprachigen Raum nicht zuletzt durch Bredenkamp (1969, 1980) und später von Hager (1987) kritisiert und weiterentwickelt wurde. Die Kernfrage ist hier die nach alternativen Erklärungen für eine beobachtete Abhängigkeit. Viele fruchtbare und wichtige methodologische Diskussionen zu Fragen des Designs von Untersuchungen und der Datenanalyse basieren auf diesem Begriff.

Unter empirisch arbeitenden Wissenschaftlern weniger einflussreich ist die Vielzahl stochastischer Kausalitätstheorien. Hervorheben möchte ich hier die auf Suppes (1970) zurückgehenden Arbeiten von Stegmüller (1983) und Spohn (1980, 1983, 1990, 1991).

Auch die Arbeiten und Ansätze in der Statistik haben sich bisher noch nicht sehr weit in der empirischen Kausalforschung durchgesetzt. Zu nennen sind hier die Arbeiten von Rubin (1974, 1977, 1985), Rubin und Thayer (1978), Holland (1986, 1988), Rosenbaum (1984a–c), Sobel (1994), Little und Yau (1998) sowie Jo (2002), die auf Neyman (1923/1990) zurückgehen. Gleiches gilt auch für die Arbeiten von Pearl (1993, 1995), die auf der Literatur zu grafischen Modellen (z. B. Whittaker, 1990) aufbauen.

Auch wenn ich hier nur auf einen kleinen Teil der Autoren hinweisen konnte, die sich dem Thema »Kausalität« gewidmet haben, wird doch deutlich, dass dieses Thema auf viele Wissenschaftler eine große Faszination ausübt. Wie kommt diese Faszination zustande? Ziel dieses Beitrags ist, dies ein Stück weit verständlich zu machen und dabei einige elementare Konzepte und Sätze der Kausalitätstheorie zu vermitteln.

Der Vergleich von Mittelwerten und relativen Häufigkeiten

Was tun wir üblicherweise, wenn wir wissen wollen, ob eine Variable X einen Effekt auf eine Variable Y hat? Das Naheliegende ist zu prüfen, ob die beiden Variablen miteinander *kovariieren*. Dies kann man etwa mit der Berechnung der Korrelation zwischen diesen beiden Variablen tun. Ist X eine Variable mit nur wenigen Ausprägungen, dann kann man auch die Mittelwerte von Y zwischen den verschiedenen Ausprägungen von X miteinander

vergleichen. Kann auch Y nur zwei Werte annehmen, so werden die relativen Häufigkeiten, mit denen Y den einen der beiden Werte annimmt, miteinander verglichen. Betrachten wir ein einfaches Beispiel!

Tabelle 1: Evaluation einer Behandlung

	Behandlung		
Erfolg	ja	nein	Gesamt
ja	500	600	1100
nein	500	400	900
	1000	1000	2000

Anmerkung: Fiktive Zahlen nach Novick (1980).

In Tabelle 1 seien die Ergebnisse einer Datensammlung zur Evaluation einer Behandlung zusammengestellt. Das Kriterium sei die Variable »Erfolg« (z. B. nicht rückfällig zu werden nach einer Alkoholismustherapie) mit den Ausprägungen »Ja« und »Nein«. Wie würde man über die Fortführung der Behandlung entscheiden, wenn man mit diesen Daten konfrontiert wäre?

Standard ist es, den Anteil der Erfolgreichen bei den Behandelten mit dem entsprechenden Anteil bei den nicht Behandelten zu vergleichen. Bei den Behandelten beträgt der Anteil der Erfolgreichen 50 Prozent, bei den nicht Behandelten dagegen 60 Prozent (siehe Abb. 1). Bei diesem Vergleich kommen wir also zu dem für die Behandlung negativen Urteil, dass sie ihr Geld nicht wert ist.

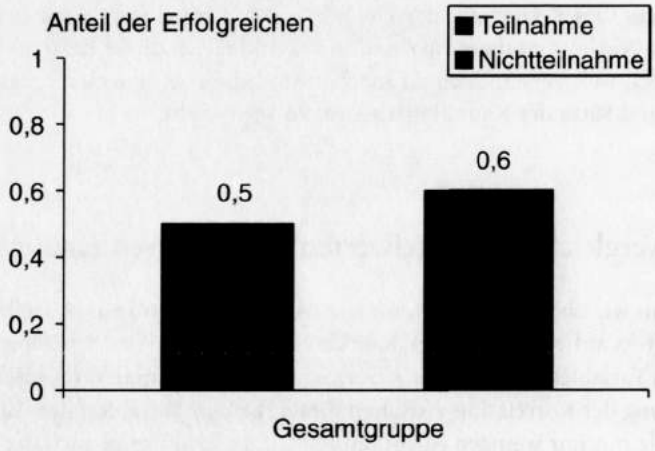


Abbildung 1: Histogramm der relativen Häufigkeiten für Erfolg in Tabelle 1.

Das Simpson-Paradox

Ganz anders allerdings sieht die Antwort aus, wenn man innerhalb der Gruppe der Männer dieselbe Betrachtung anstellt. Hier sind nun 40 Prozent erfolgreich bei den Behandelten und nur 30 Prozent bei den nicht Behandelten (siehe Tab. 2a). Anders als in der Gesamtpopulation scheint die Behandlung bei den Männern also erfolgreich zu sein.

Tabelle 2: Evaluation einer Behandlung getrennt nach Geschlecht

(a) Männer

	Behandlung		
Erfolg	ja	nein	Gesamt
ja	300	75	375
nein	450	175	625
	750	250	1000

(b) Frauen

	Behandlung		
Erfolg	ja	nein	Gesamt
ja	200	525	725
nein	50	225	275
	250	750	1000

Anmerkung: Fiktive Zahlen nach Novick (1980).

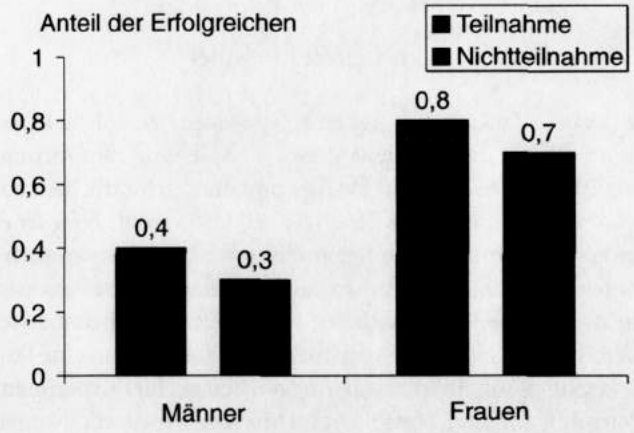


Abbildung 2: Histogramm der relativen Häufigkeiten für Erfolg in Tabelle 2, getrennt nach Männern und Frauen.

Wie sieht es nun bei der Gruppe der Frauen aus? Unserer Erwartung zufolge müssten eigentlich die nicht Behandelten in starkem Maße erfolgreicher sein als die Behandelten, um die Ergebnisse in der Gesamtgruppe erklären zu können. Ein Blick in die Tabelle 2b zeigt jedoch, dass uns unser gesunder Menschenverstand hier trügt. Auch bei den Frauen ist der Anteil der Erfolgreichen bei den Behandelten größer als bei den nicht Behandelten.

In Abbildung 2 sind die betreffenden Anteile dargestellt. Wir sehen: In beiden Geschlechtsgruppen ist offenbar die Behandlung erfolgreich, in der Gesamtgruppe, bestehend aus eben diesen Männern und Frauen, jedoch nicht. In anderen Worten: Wir können sowohl Männer als auch Frauen erfolgreich behandeln, nicht jedoch Menschen. So könnte das paradox erscheinende Fazit der Vergleiche der relativen Häufigkeiten in den Abbildungen 1 und 2 lauten.

Wie kann das sein? In jeder der beiden Subpopulationen ein positiver Effekt, und in der Gesamtpopulation, die sich genau aus diesen beiden Subpopulationen zusammensetzt, ein negativer Effekt! Müsste nicht in der Gesamtpopulation der Durchschnitt der Effekte in den Subpopulationen zu finden sein? Wieso läuft unser intuitives Denken hier in die Irre? Im Wesentlichen liegt es am gleichzeitigen Auftreten zweier Bedingungen: erstens, die ungleichen Erfolgswahrscheinlichkeiten bei Männern und Frauen (siehe die rechte Spalte der Tab. 2; 375/1000 vs. 725/1000) und zweitens, die ungleiche Behandlungswahrscheinlichkeit (750/1000 bei den Männern, 250/1000 bei den Frauen). Kommt beides zusammen, entstehen derartige, paradox erscheinende Phänomene.

Ein weiteres Beispiel

Ein weiteres Beispiel kann uns helfen zu verstehen, wie solche Paradoxa zustande kommen. Daneben dient dieses Beispiel zur Einführung einiger Grundbegriffe der Theorie individueller und durchschnittlicher kausaler Effekte. Bemerkenswert ist dieses Beispiel auch insofern, als *bei jeder einzelnen* Person ein *positiver* Effekt einer Behandlung vorliegt, aber dennoch eine *negative* Mittelwertsdifferenz auftreten kann, wenn man diese Personen in zwei Gruppen aufteilt, die der Behandelten und die der nicht Behandelten.

Machen wir ein Gedankenexperiment! In Tabelle 3a sei eine Population von acht Personen mit ihren Erwartungswerten in einer Experimental- und einer Kontrollbedingung, sowie deren Differenz dargestellt. Natürlich sind in jedem normalen Anwendungsfall diese Werte unbekannt, aber wir können uns vorstellen, dass die angegebenen individuellen Erwartungswerte

Tabelle 3: Zahlenbeispiel, in dem für jede einzelne Person ein positiver Effekt der Experimentalgruppe besteht und bei dem eine nicht ausgewogene Auswahl der Personen für die Experimental- und Kontrollgruppe (siehe die fett markierten Zahlen) zu verfälschten Schätzungen des durchschnittlichen Effekts führt

(a)				(b)		
Person	$\mu_{E,u}$	$\mu_{K,u}$	$\mu_{E,u}-\mu_{K,u}$	$\mu_{E,u}$	$\mu_{K,u}$	$\mu_{E,u}-\mu_{K,u}$
u_1	82	68	14	**82**	68	14
u_2	89	81	8	**89**	81	8
u_3	101	89	12	**101**	89	12
u_4	108	102	6	**108**	102	6
u_5	118	112	6	118	**112**	6
u_6	131	119	12	131	**119**	12
u_7	139	131	8	139	**131**	8
u_8	152	138	14	152	**138**	14
Mittel	115	105	10	95	125	–30
(c)				(d)		
Person	$\mu_{E,u}$	$\mu_{K,u}$	$\mu_{E,u}-\mu_{K,u}$	$\mu_{E,u}$	$\mu_{K,u}$	$\mu_{E,u}-\mu_{K,u}$
u_1	**82**	68	14	82	**68**	14
u_2	89	**81**	8	**89**	81	8
u_3	**101**	89	12	**101**	89	12
u_4	**108**	102	6	108	**102**	6
u_5	118	**112**	6	**118**	112	6
u_6	**131**	119	12	131	**119**	12
u_7	139	**131**	8	139	**131**	8
u_8	152	**138**	14	**152**	138	14
Mittel	105.5	115.5	–10	115	105	10

Anmerkung: **Fett** markierte Werte in den Tabellen (b) bis (d) sind die ausgewählten Werte.

das empirische Phänomen genauso steuern, wie die Wahrscheinlichkeit die möglichen Ergebnisse bei einem Würfelwurf. Würden wir diese individuellen Erwartungswerte kennen, so würden wir den kausalen Effekt der Behandlung *E* gegenüber der Behandlung *K* für jede einzelne Person, das heißt die *individuellen kausalen Effekte* kennen – der Traum jedes empirischen Kausalforschers und jedes angewandten Psychologen: Tue ich *E*, erwarte ich bei der Person u_1 den Wert 82 auf der betrachteten Lebenszufriedenheitsskala, tue ich *K* erwarte ich dagegen nur den Wert 68. Also tue ich besser *E*. Bei der Person u_6 erwarte ich bei *E* den Wert 131, bei *K* nur den Wert 119. Eine idealere Situation können wir uns kaum vorstellen!

Leider kann eine Person in der Regel entweder nur der Experimental- oder nur der Kontrollbedingung unterworfen werden, nicht aber beiden Bedingungen zugleich. Entweder ich therapiere nach Verfahren *A* oder nach Verfahren *B*, entweder ich unterrichte nach Lehrmethode *A* oder nach

Lehrmethode *B*. Paul Holland (1986) hat dies das *fundamentale Problem der kausalen Inferenz* genannt. Es hat zur Folge, dass wir in der Regel die individuellen kausalen Effekte auch nicht schätzen können.

Allerdings kann man unter bestimmten Bedingungen etwas über den Durchschnitt der individuellen kausalen Effekte, den *durchschnittlichen kausalen Effekt* (siehe die letzte Zahl rechts unten in der Tab. 3a) in einer Population und in Subpopulationen erfahren. Wären wir in der Lage, die Spaltenmittelwerte der ersten beiden Spalten in Tabelle 3a zu schätzen, indem wir *einige Personen unter der einen* (*E*) und *andere Personen unter der anderen experimentellen Bedingung* (*K*) *beobachten*, so hätten wir mit der Differenz dieser Spaltenmittelwerte eine Schätzung für den durchschnittlichen kausalen Effekt (siehe die letzte Zeile von Tab. 3a).

Die Spaltenmittelwerte der ersten beiden Spalten der Tabelle 3a zu schätzen, indem wir einige Personen unter der einen (*E*) und andere Personen unter der anderen experimentellen Bedingung (*K*) beobachten, kann man auf unterschiedliche, mehr oder weniger kluge Weisen versuchen. In Tabelle 3b ist eine weniger kluge Weise dargestellt, deren Betrachtung aber dennoch lehrreich ist. Dabei wurden die ersten vier Personen der Experimentalbedingung zugeordnet, die letzten vier der Kontrollbedingung (siehe die fett markierten Zahlen). In der Experimentalbedingung ergibt sich dabei ein Mittelwert der vier ausgewählten Werte von 95, in der Kontrollbedingung von 125. Die Differenz beträgt –30, die miserabelste Schätzung für den durchschnittlichen kausalen Effekt (+10), die in diesem Beispiel auftreten kann.

Was ist hier passiert? Offenbar ist die Zuweisung der Personen zu den beiden Bedingungen, *E* und *K*, nicht unabhängig von den Erwartungswerten der Personen in den beiden experimentellen Bedingungen (siehe die zweite Zahlenspalte in Tab. 3a).[3] Im Gegenteil, die Personen mit den niedrigsten Erwartungswerten wurden der Experimental-, die mit den höchsten Erwartungswerten der Kontrollbedingung zugeordnet. Die Konsequenz ist für die kausale Interpretierbarkeit der resultierenden Mittelwertsdifferenz verheerend.

Etwas besser, aber immer noch gründlich daneben liegt die Berechnung des durchschnittlichen kausalen Effekts durch die Mittelwertsdifferenz bei der in Tabelle 3c dargestellten Zuordnung der Personen zu den beiden Bedingungen *E* und *K*. Statt + 10 erhält man die Mittelwertsdifferenz –10.

---

3 Diese beiden Spalten der individuellen Erwartungswerte, die wir im Folgenden als *erwartete Outcome-Variable* bezeichnen, treten in der hier vorgestellten Version der Theorie der individuellen und durchschnittlichen kausalen Effekte an die Stelle der Rubinschen *potenziellen Outcome-Variablen* Y_0 und Y_1 (s. auch Holland, 1986).

Erst bei der in Tabelle 3d dargestellten Zuordnung kommt man zum richtigen Ergebnis, das heißt, erst bei einer Auswahl, in der eine ausgewogene Berücksichtigung von Personen mit hohen und niedrigen Erwartungswerten in beiden Bedingungen realisiert wird, ist die resultierende Mittelwertsdifferenz unverfälscht in dem Sinn, dass sie gleich dem durchschnittlichen kausalen Effekt ist.

Wie lassen sich diese Überlegungen auf reale Experimente übertragen, in denen die Erwartungswerte der Personen nicht bekannt sind? Die klassische und wohlbekannte Antwort ist, die Personen zufällig auf die experimentellen Bedingungen aufzuteilen. Man spricht in diesem Kontext von *Randomisierung* und vom *randomisierten Experiment.*

Randomisierung garantiert, dass die Erwartungswertdifferenz $\mu_E - \mu_K$ (im Beispiel der Tab. 4 mit der vorletzten Spalte: $115 - 105 = 10$)[4] gleich dem durchschnittlichen kausalen Effekt ist. In Stichproben weichen dann die entsprechenden Mittelwertsdifferenzen nur zufällig vom durchschnittlichen kausalen Effekt ab. Neben der Vergleichbarkeit der erwarteten Outcome-Variablen zwischen den experimentellen Bedingungen stellt man mit diesem Vorgehen sogar eine Vergleichbarkeit der experimentellen Bedin-

Tabelle 4: Zahlenbeispiel, in dem für jede einzelne Person ein positiver Effekt der Experimentalgruppe gegenüber der Kontrollgruppe besteht. Die Wahrscheinlichkeit, der Experimentalbedingung zugewiesen zu werden, ist entweder unabhängig (vorletzte Spalte) oder abhängig (letzte Spalte) von der Person

Person	$\mu_{E,u}$	$\mu_{K,u}$	$\mu_{E,u} - \mu_{K,u}$	$P(E\|u)$	$P(E\|u)$
u_1	82	68	14	1/2	8/9
u_2	89	81	8	1/2	7/9
u_3	101	89	12	1/2	6/9
u_4	108	102	6	1/2	5/9
u_5	118	112	6	1/2	4/9
u_6	131	119	12	1/2	3/9
u_7	139	131	8	1/2	2/9
u_8	152	138	14	1/2	1/9
			10		

Anmerkung: Die Wahrscheinlichkeit, eine der acht Personen zu ziehen, ist 1/8 und die (unbedingte) Wahrscheinlichkeit, der Experimentalbedingung zugewiesen zu werden, ist in beiden Fällen (charakterisiert durch die beiden letzten Spalten) 1/2.

4 Der Erwartungswert μ_E lässt sich nach der allgemein gültigen Formel $\mu_E = \Sigma_i \mu_{E,ui} P(u_i|E)$ berechnen. Die dazu notwendigen Zahlen findet man in Tabelle 4, wobei allerdings die dort angegebenen individuellen Behandlungs-Wahrscheinlichkeiten $P(E|u_i)$ in die bedingten Wahrscheinlichkeiten $P(u_i|E)$ umgerechnet werden müssen. Das Entsprechende gilt natürlich für μ_K.

gungen hinsichtlich aller möglichen Eigenschaften der Personen her, die diesen Bedingungen nach dem Zufallsprinzip zugewiesen werden.

Hängt die Wahrscheinlichkeit der Zuweisung der Personen dagegen von den erwarteten Outcome-Variablen ab (siehe Tab. 4 mit der letzten Spalte; die Zahlen sind ansonsten willkürlich gewählt), erhält man eine Erwartungswertdifferenz $\mu_E - \mu_K$ (im Beispiel: 103,3 – 116,6 = –13,3), die deutlich vom durchschnittlichen kausalen Effekt abweicht.

Man kann sich leicht überlegen, dass die zufällige Zuweisung der Personen zu den experimentellen Bedingungen überflüssig ist, wenn die Population homogen ist, das heißt, wenn die erwarteten Outcome-Variablen innerhalb jeder experimentellen Bedingung konstant ist, also alle Personen völlig gleich hinsichtlich der betrachteten Variablen funktionieren. In Tabelle 5 ist veranschaulicht, was damit gemeint ist. Funktionieren alle Personen hinsichtlich der betrachteten Variablen X und Y völlig gleich, spielt es keine Rolle, wie man sie auf die experimentellen Bedingungen aufteilt. Natürlich ist klar, dass diese Homogenität in unseren empirischen Disziplinen nur selten vorkommt. Man kann allerdings versuchen, sich ihr durch die Auswahl von Subpopulationen anzunähern. Dies entspricht der bereits erwähnten Millschen Idee der Konstanthaltung aller anderen Einflussgrößen.

Tabelle 5: Veranschaulichung einer homogenen Population

Person	$\mu_{E,\,u}$	$\mu_{K,\,u}$	$\mu_{E,\,u} - \mu_{K,\,u}$
u_1	110	100	10
u_2	110	100	10
u_3	110	100	10
u_4	110	100	10
u_5	110	100	10
u_6	110	100	10
u_7	110	100	10
u_8	110	100	10
Mittel	110	100	10

Die Theorie individueller und durchschnittlicher kausaler Effekte

Fassen wir die bisher eingeführten Begriffe und die damit gewonnenen Erkenntnisse zusammen (siehe Abb. 3)! Eine Erwartungswertdifferenz $\mu_E - \mu_K$ bezüglich einer Kriteriumsvariablen Y zwischen zwei Bedingungen E und K, der *Prima-facie-Effekt* (siehe Abb. 3), ist *nicht per se* das, woran wir in der empirischen Kausalforschung interessiert sind.

$$PFE_{EK} := \mu_E - \mu_K$$

Individueller kausaler Effekt

$$ICE_{EK}(u) := \mu_{E,u} - \mu_{K,u}$$

Durchschnittlicher kausaler Effekt

$$ACE_{EK} := \frac{1}{N} \sum_{i=1}^{N} (\mu_{E,u_i} - \mu_{K,u_i}).$$

Kausale Unverfälschtheit des Prima-facie-Effekts

$$PFE_{EK} = ACE_{EK}.$$

Abbildung 3: Zusammenfassung der Grundbegriffe der Theorie individueller und durchschnittlicher kausaler Effekte.

Ein *individueller kausaler Effekt* einer Bedingung E bezüglich einer Bedingung K auf eine Kriteriumsvariable Y für eine Person u ist die Differenz zwischen dem Erwartungswert von Y unter der Bedingung E und dem Erwartungswert von Y unter der Bedingung K bei der gleichen Person u.

Der *durchschnittliche kausale Effekt* einer Bedingung E bezüglich einer Bedingung K auf eine Kriteriumsvariable Y ist der Durchschnitt der individuellen Erwartungswerte über die betrachtete Population von Personen.

Eine Erwartungswertdifferenz $\mu_E - \mu_K$ zwischen zwei experimentellen Bedingungen kann unter bestimmten, günstigen Voraussetzungen (siehe unten) kausal unverfälscht sein. *Kausale Unverfälschtheit* eines Prima-facie-Effekts $\mu_E - \mu_K$ heißt: Die Differenz $\mu_E - \mu_K$ ist gleich dem durchschnittlichen kausalen Effekt.

Dies sind die wichtigsten Grundbegriffe. Zwei wichtige Theoreme, die diese Grundbegriffe miteinander in Verbindung bringen, werden in Abbildung 4 illustriert. Diese Theoreme sind letztlich die Grundlage für die Techniken des Experimentierens.

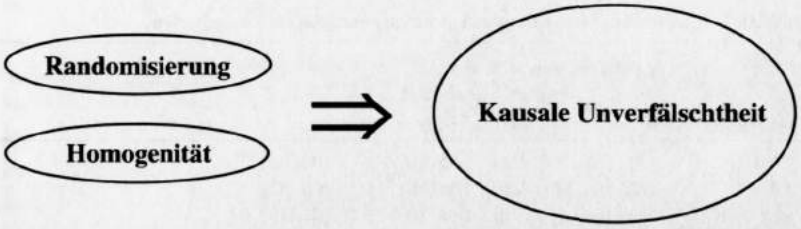


Abbildung 4: Hinreichende Bedingungen für die kausale Unverfälschtheit des Prima-facie-Effekts. Beide Bedingungen, Randomisierung und Homogenität, sind für sich alleine hinreichend für die kausale Unverfälschtheit.

Eine erste hinreichende Bedingung für die kausale Unverfälschtheit des Prima-facie-Effekts $\mu_E - \mu_K$ ist eine Zuweisung der Personen zu den experimentellen Bedingungen, die *unabhängig* von der Person (und ihren Eigenschaften) ist. Eine zweite hinreichende Bedingung für die kausale Unverfälschtheit des Prima-facie-Effekts ist die Homogenität der Population.

Diese Ergebnisse erscheinen zu Recht nicht besonders überraschend. Tatsächlich ist bisher lediglich eine Präzisierung des Begriffs der kausalen Unverfälschtheit des Prima-facie-Effekts $\mu_E - \mu_K$ zu vermerken. Alle anderen Ergebnisse entsprechen ganz unserem intuitiven Vorwissen. Ob sich die Mühe der Präzisierung gelohnt hat, wird sich erst dann zeigen, wenn es gelingt, bisher ungelöste Probleme einer Lösung zu zuführen. Einige solcher Probleme gibt es im Bereich der nonorthogonalen Varianzanalyse und des Designs und der Analyse nicht randomisierter Untersuchungen. Beginnen wir zunächst mit dem Bekannten, der Analyse orthogonaler zweifaktorieller Versuchspläne!

Design und Analyse orthogonaler Versuchspläne

Die oben zusammengefassten Ergebnisse lassen sich direkt auf Fragen des Designs und der Analyse orthogonaler varianzanalytischer Versuchspläne anwenden. In Tabelle 6 ist ein Beispiel für einen solchen orthogonalen varianzanalytischen Versuchsplan dargestellt, bei dem der Faktor A drei experimentelle Bedingungen (hier drei Therapieverfahren) und der Faktor B eine Eigenschaft der Versuchspersonen vor Beginn der Behandlung darstellt (hier die »Behandlungsbedürftigkeit« in drei Abstufungen, stark, mittel und schwach behandlungsbedürftig).

Tabelle 6: Beispiel für einen orthogonalen varianzanalytischen Versuchsplan

	Behandlungsbedürftigkeit			
	stark	mittel	schwach	Gesamt
Therapie	$Z = z_1$	$Z = z_2$	$Z = z_3$	
1 $X = x_1$	120 (10)	110 (20)	60 (10)	100 (40)
2 $X = x_2$	100 (10)	100 (20)	100 (10)	100 (40)
3 $X = x_3$	80 (10)	90 (20)	140 (10)	100 (40)
Gesamt	100 (30)	100 (60)	100 (30)	100 (120)

Anmerkung: Zellenfrequenzen (in Klammern) und wahre Zellenmittelwerte.

Dabei sei man wie folgt vorgegangen: Es wurden zunächst $n = 120$ Personen ausgewählt, die am Experiment mitwirken sollten. Nach einer Anamnese wurde festgestellt, wie viele und welche dieser Probanden stark, mittel oder schwach behandlungsbedürftig sind. (Im Beispiel seien dies 30, 60 oder 30.) Innerhalb jeder dieser drei Bedürftigkeitsgruppen wurden *nach dem Zufallsprinzip* drei gleich große Gruppen gebildet, die jeweils nach einem der drei Therapieverfahren behandelt wurden. Nach einer angemessenen Zeit wurde bei allen Probanden die Kriteriums- oder Outcome-Variable erhoben.

Geht man nach diesem Schema vor, so kann man die Mittelwertsunterschiede zwischen den drei experimentellen Bedingungen, das heißt die Differenzen zwischen den Zeilenmittelwerten, als unverfälschte Schätzungen der durchschnittlichen kausalen Effekte interpretieren. Dies ist die gängige und wohl begründete Praxis der orthogonalen Varianzanalyse.

Auch innerhalb jeder der drei Bedürftigkeitsgruppen lassen sich die Mittelwertsunterschiede zwischen den drei experimentellen Bedingungen als Schätzungen der durchschnittlichen kausalen Effekte interpretieren, allerdings nicht bezogen auf die Gesamtpopulation, sondern bezüglich der jeweiligen Subpopulation.

Nonorthogonale Designs

Wie sieht es jedoch aus, wenn wir Vorwissen über die Wirkungsweise der drei Therapieformen haben und daher beispielsweise aus ethischen Gründen nicht bereit sind, bei jeder der drei Ausprägungen der Behandlungsbedürftigkeit jede der drei Therapieformen *mit gleicher Wahrscheinlichkeit* anzuwenden? Nehmen wir an, wir könnten uns zwar zu einer zufälligen Zuweisung der Probanden zu den drei Therapieverfahren entschließen, *aber mit unterschiedlichen Wahrscheinlichkeiten, je nach Ausprägung der Behandlungsbedürftigkeit.* Dieses Verfahren nennen wir *bedingte Randomisierung.* Die Bedürftigen könnten dann eine größere Behandlungswahrscheinlichkeit haben als die Nichtbedürftigen. Die Zuordnungswahrscheinlichkeiten seien dabei proportional zu den in Tabelle 7 angegebenen Anzahlen der Beobachtungen in den Zellen, den *Zellenfrequenzen.* Für eine stark bedürftige Person sei also die Wahrscheinlichkeit, mit der Therapie x_1 behandelt zu werden, 20/30 und für eine mittel bedürftige Person sei die Wahrscheinlichkeit, mit der Therapie x_2 behandelt zu werden, 40/60. Zerstören diese ungleichen Wahrscheinlichkeiten, den Therapiebedingungen zugeordnet zu werden, die kausale Unverfälschtheit?

Tabelle 7: Beispiel für einen nonorthogonalen varianzanalytischen Versuchsplan

	Behandlungsbedürftigkeit			
	stark	mittel	schwach	Gesamt
Therapie	$Z = z_1$	$Z = z_2$	$Z = z_3$	
1 $X = x_1$	120 (20)	110 (10)	60 (3)	(33)
2 $X = x_2$	100 (7)	100 (40)	100 (7)	(54)
3 $X = x_3$	80 (3)	90 (10)	140 (20)	(33)
Gesamt	(30)	(60)	(30)	(120)

Anmerkung: Zellenfrequenzen (in Klammern) und wahre Zellenmittelwerte.

Die Antwort heißt »Jein«. »Ja«, wenn man die normalen Zeilenmittelwerte betrachtet und »Nein«, wenn man die Daten richtig auszuwerten weiß. Kausal unverfälscht – im Sinne des durchschnittlichen kausalen Effekts in der entsprechenden Subpopulation – bleiben weiterhin die Mittelwertsunterschiede zwischen den Therapiebedingungen *innerhalb* jeder der drei Ausprägungen des Faktors »Behandlungsbedürftigkeit«. Darüber hinaus kann man hier auch die durchschnittlichen kausalen Effekte in der Gesamtpopulation berechnen, die sich allerdings nicht mehr als Differenz der (normalen) Zeilenmittelwerte ergeben. Stattdessen muss man *aus den Zellenmittelwerten durch Gewichtung mit den relativen Spaltenhäufigkeiten adjustierte Zeilenmittelwerte* berechnen. Letztere sind in diesem Beispiel alle drei gleich, nämlich gleich 100. Demnach sind in diesem Beispiel auch die durchschnittlichen kausalen Effekte in der Gesamtpopulation gleich null, ein Ergebnis, das keines der gängigen Verfahren der nonorthogonalen Varianzanalyse liefert, wie sie in SAS, SPSS, SYSTAT oder ähnlichen Softwarepaketen angeboten werden. Hier dient die Kausalitätstheorie also der Ableitung neuer Analyseverfahren, die alles andere als selbstverständlich und altbekannt sind.

In Abbildung 5 sind die für nonorthogonale Designs relevanten Ergebnisse zusammengefasst. Dabei repräsentiert die Variable Z mit ihren Ausprägungen z den Faktor »Behandlungsbedürftigkeit«. Details findet man in Steyer (2003), Steyer et al. (2002) und Wüthrich-Martone (2001).

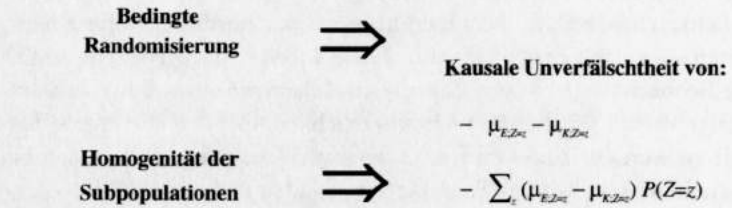


Abbildung 5: Hinreichende Bedingungen für die kausale Unverfälschtheit der Prima-facie-Effekte in Subpopulationen und der adjustierten Effekte in der Gesamtpopulation.

Analyse nichtexperimenteller Beobachtungsstudien

Auch aus reinen Beobachtungsstudien, in denen wir die natürlich auftretende Korrelation verschiedener Variablen erfassen, wollen wir oft auf kausale Abhängigkeiten schließen. In epidemiologischen Untersuchungen, in denen es beispielsweise um den Zusammenhang zwischen Rauchen und Lungenkrebs, fettreiche Ernährung und Herzinfarkt, Strahlenbelastung in der Nähe von Kernkraftwerken und Leukämie, Stress und Krankheitsanfälligkeit geht, sind kausale Interpretationen von Abhängigkeiten und Mittelwertsunterschieden von zentraler Bedeutung.

Bleiben wir bei den einfachen Fällen, in denen sich Abhängigkeiten als Mittelwertsunterschiede darstellen lassen! Voraussetzungsgemäß liegt bei den in diesem Abschnitt betrachteten Fällen weder ein randomisiertes, noch ein bedingt randomisiertes Experiment vor. Auch Homogenität, die ja ebenfalls eine kausale Unverfälschtheit einer Mittelwertsdifferenz garantieren würde, kann nicht ohne weiteres angenommen werden (siehe Abb. 4). Nachdem also in solchen Studien diese bisher behandelten hinreichenden Bedingungen für die kausale Unverfälschtheit von Mittelwertsunterschieden ausfallen, gibt es dann wenigstens notwendige Bedingungen, die aus der Behauptung folgen, dass eine Mittelwertsdifferenz kausal unverfälscht ist?

Die Antwort heißt leider »Nein«. Aus der Aussage der kausalen Unverfälschtheit einer Mittelwertsdifferenz lässt sich nichts logisch ableiten, was sich *in der betreffenden Studie* empirisch überprüfen ließe. Das liegt ganz einfach daran, dass die individuellen kausalen Effekte und damit auch deren Durchschnitt über eine Menge von Personen (die Population) nicht schätzbar sind. Damit ist die Behauptung, eine konkrete Mittelwertsdifferenz sei kausal unverfälscht, das heißt, $PFE_{EK} = ACE_{EK}$ (siehe Abb. 3), *in der betreffenden Studie* nicht falsifizierbar.

Bevor wir die Konsequenzen dieser Feststellung weiter erörtern, soll sie an einem Beispiel illustriert werden. Angenommen, jemand behauptet, dass der Unterschied zwischen Rauchern und Nichtrauchern in der Wahrscheinlichkeit, an Lungenkrebs zu sterben, im oben definierten Sinn (d. h. $PFE_{EK} = ACE_{EK}$) kausal unverfälscht ist.[5] Die individuellen kausalen Effekte sind hier nicht schätzbar, weil jede Person entweder Raucher oder aber Nichtraucher ist. Damit ist auch deren Durchschnitt über eine Menge von

5 Dabei sei angemerkt, dass ein solcher Wahrscheinlichkeitsunterschied auch als Mittelwertsdifferenz aufgefasst werden kann, nämlich als Mittelwertsdifferenz einer Variablen mit den Werten 0 und 1. Dabei steht in diesem Beispiel der Wert 1 für das Ereignis, an Lungenkrebs zu sterben, und der Wert 0 für das Komplementärereignis.

Personen, die Population, nicht schätzbar. Wie sollte man dann also prüfen, ob ein festgestellter Unterschied zwischen Rauchern und Nichtrauchern in der Wahrscheinlichkeit, an Lungenkrebs zu sterben, gleich dem Durchschnitt der individuellen kausalen Effekte ist?

Die Feststellung, dass eine Aussage, eine konkrete Mittelwertsdifferenz sei kausal unverfälscht, in der betreffenden Studie nicht empirisch falsifizierbar ist, hat wichtige Konsequenzen. Sie sei daher noch einmal wohl bedacht! Im Prinzip könnten wir eine solche Aussage doch in einem *weiteren*, randomisierten Experiment überprüfen. Daher habe ich oben betont: *»in der betreffenden Studie«*. In der Praxis stellen sich diesem Vorgehen allerdings oft Probleme in den Weg. Dort, wo wir uns auf die Analyse von reinen Beobachtungsdaten stützen müssen, lässt sich ja gerade kein randomisiertes Experiment durchführen. Selbst im Dienst der Wissenschaft können wir keine Zufallsstichprobe aus der Population ziehen, die betreffenden Personen zwingen, zu Rauchern zu werden, und den Anteil, der an Lungenkrebs stirbt, mit dem entsprechenden Anteil bei einer Zufallsauswahl von Personen vergleichen, die wir zwingen, Nichtraucher zu bleiben.

Was man jedoch tun kann, ist zu prüfen, ob wir gute Gründe haben anzunehmen, dass eine Mittelwertsdifferenz oder eine Differenz in den Wahrscheinlichkeiten kausal unverfälscht ist. Ein guter Grund wäre, dass wir bereit sind anzunehmen, dass eine *hinreichende Bedingung* für die kausale Unverfälschtheit der betreffenden Mittelwertsdifferenz erfüllt ist. Da es neben den bereits erwähnten Bedingungen der Randomisierung und der Homogenität noch weitere hinreichende Bedingungen für die kausale Unverfälschtheit gibt (siehe dazu Steyer et al., 2000a), ist es hier wichtig, dass man die *schwächste* hinreichende Bedingung für die kausale Unverfälschtheit prüft, die noch einer empirischen Überprüfung zugänglich ist.

Es würde an dieser Stelle zu weit führen, hier diese Bedingung, die *Unkonfundiertheit*, allgemein zu definieren. Dazu sei auf Steyer et al. (1996) sowie auf Steyer et al. (2000a) verwiesen. Im Wesentlichen ist der Begriff der Unkonfundiertheit dadurch definiert, dass die Erwartungswerte, deren Differenz kausal unverfälscht sein soll, hier also μ_E und μ_K, selbst immer Erwartungswerte der subpopulationsspezifischen Erwartungswerte sind. Wenn wir zwei experimentelle Bedingungen, E und K betrachten, und wenn Z eine Variable ist, welche die Population in Subpopulationen partitioniert (z. B. das Geschlecht, der Familienstand), dann müssen demnach gelten:

$$\mu_E = \Sigma_z \, \mu_{E, Z = z} \, P(Z = z),$$

$$\mu_K = \Sigma_z \, \mu_{K, Z = z} \, P(Z = z),$$

wobei hier über alle Werte z der Variablen Z summiert werden soll. Gilt dies für jede solche Variable Z, dann definiert dies die *Unkonfundiertheit* der bedingten Erwartungswerte μ_E und μ_K.[6] Entscheidend in diesem Kontext ist, dass die Unkonfundiertheit sowohl bei Randomisierung als auch bei Homogenität immer erfüllt ist und ihrerseits die kausale Unverfälschtheit impliziert (siehe Abb. 5). Es sei auch angemerkt, dass erst die Unkonfundiertheit garantiert, dass kein Simpson-Paradox mehr auftreten kann. Mit anderen Worten: Trotz Gültigkeit der kausalen Unverfälschtheit eines Prima-facie-Effekts kann es passieren, dass in der Gesamtpopulation ein negativer Prima-facie-Effekt vorliegt, obwohl diese in Subpopulationen mit jeweils positiven Prima-facie-Effekten zerlegt werden kann (siehe Steyer et al., 2000b). In diesem Fall würde also nicht die Analyse auf Gesamtpopulations-, sondern die auf Subpopulationsebene verfälschte Effekte liefern. Erst die obige Unkonfundiertheitsbedingung garantiert, dass solche Phänomene nicht mehr auftreten können. (Zur Verallgemeinerung dieser Unkonfundiertheitsbedingung für kontinuierliche Variable X und Z siehe Steyer, 1992).

Zusammenfassung und Ausblick

Was wollen und was können wir also in der empirischen Kausalforschung erfahren? Was wir erfahren *wollen*, ist etwas über die *individuellen kausalen Effekte*[7]. Worüber wir unter günstigen Umständen etwas erfahren *können*, ist deren *Durchschnitt* in einer Gesamtpopulation oder in verschiedenen Subpopulationen. Die günstigen Umstände kann man in einem randomisierten Experiment herstellen, in dem man die Personen den experimentellen Bedingungen nach dem Zufallsprinzip zuordnet. Woran wir im Kontext empirischer Kausalforschung nicht *per se* interessiert sind, sind Mittelwerte und deren Differenzen.

Insbesondere die Erkenntnis über das randomisierte Experiment ist nicht unbedingt neu. Neu ist auch nicht, dass man aus der Theorie der individuellen und durchschnittlichen kausalen Effekte ableiten kann, wie

6 Für eine ausführlichere Darstellung der Theorie, verschiedener Theoreme und Beispiele sei auf Steyer et al., (2000a) verwiesen.

7 Nach Drucklegung dieses Artikels haben wir auch Designs und Modelle entwickelt, mit denen wir die individuellen Effekte sowie deren Varianzen und Kovarianzen identifizieren können.

die Daten einer orthogonalen Varianzanalyse analysiert werden sollten. Für die *nonorthogonale Varianzanalyse* haben wir allerdings ein altes Problem lösen können. Ähnliches ist übrigens auch für die *Kovarianzanalyse* zu erwarten, wenn die Kovariate und die experimentelle Variable nicht unabhängig sind.

In solchen Untersuchungen, in denen die Personen den Behandlungsbedingungen *nicht* zufällig zugewiesen werden können, bleibt die kausale Modellierung schwierig. Die Unkonfundiertheitsbedingung ist nicht verifizierbar, aber doch falsifizierbar. Sie ist nicht verifizierbar, weil sie immer nur für *einige* von unendlich vielen potenziellen Störvariablen überprüft werden kann. Sie ist falsifizierbar, weil der Nachweis der Konfundiertheit bezüglich einer einzigen potenziellen Störvariablen Z genügt, um die Unkonfundiertheit für alle potenziellen Störvariablen zu widerlegen. Die Unkonfundiertheit ist meines Wissens die schwächste empirisch falsifizierbare hinreichende Bedingung für die kausale Unverfälschtheit eines Prima-facie-Effekts. Scheitern in einer konkreten Anwendung Versuche zu ihrer Falsifikation, so wird die kausale Unverfälschtheit eines Prima-facie-Effekts plausibler. Gelingt die Falsifikation der Unkonfundiertheit für eine spezielle Störvariable, so bleibt die kausale Interpretation der Prima-facie-Effekte *in den Subpopulationen*, die durch die betreffende Störvariable erzeugt werden. Die betreffende Störvariable wäre dann also konstant zu halten.

Es gilt, sowohl Designs der Datenerhebung und Methoden der Datenanalyse neu zu entwickeln, als auch bekannte Designs daraufhin zu untersuchen, ob sie sich für kausale Modellierungen eignen. Die theoretischen Grundlagen dazu sind geschaffen und warten auf ihre Weiterentwicklung und kreative Anwendung. Klar ist, dass inhaltliche Forschung, die nicht auf dem gesicherten Fundament einer wohlverstandenen Kausalitätstheorie beruht, Gefahr läuft, in die Irre zu führen und auf die Dauer nicht verantwortbar ist. Hierin liegt eine große Herausforderung und Chance für die Methodenlehre der Sozial- und Verhaltenswissenschaften.

Literatur

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bredenkamp, J. (1969). Experiment und Feldexperiment. In C.F. Graumann (Hg.), *Handbuch der Psychologie: Sozialpsychologie. 1. Halbband: Theorien und Methoden* (S. 332–373). Göttingen: Hogrefe.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Campbell, D.T., Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Hg.), *Handbook on research on teaching* (S. 171–246). Chicago, IL: Rand McNally.

Cook, T.D.C. (1986). The causal assumptions of quasi-experimental practice. *Synthese, 68,* 141–180.

Cook, T.D.C., Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Hg.), *Allgemeine experimentelle Psychologie* (S. 43–264). Stuttgart: G. Fischer.

Holland, P. (1986). Statistics and causal inference (with comments). *Journal of the American Statistical Association, 81,* 945–970.

Holland, P.W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology, 18,* 449–484.

Hume, D. (1888). *A treatise on human nature* (Original work published 1739/40). Oxford: Clarendon Press.

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods, 7,* 178–193.

Jöreskog, K. (1993). Testing structural equation models. In K.A. Bollen u. J.S. Long (Hg.), *Testing structural equation models* (S. 294–316). Newbury Park: Sage.

Little, R.J., Yau, L.H.Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3,* 147–159.

Mill, J.S. (1862). *System der deductiven und inductiven Logik. Erster Theil.* (Übersetzung von J. Schiel. 2. deutsche, nach der fünften des Originals erweiterte Auflage. Original erschienen 1843). Braunschweig: Vieweg.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (reprint 1990). *Statistical Science, 5,* 465–472.

Novick, M.R. (1980). Statistics as psychometrics. *Psychometrika, 45,* 411–424.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science, 8,* 266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82,* 669–710.

Pearl, J. (2000). *Causality – Models, reasoning, and inference.* Cambridge, MA: Cambridge University Press.

Rosenbaum, P.R. (1984a). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association, 79,* 565–574.

Rosenbaum, P.R. (1984b). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P.R. (1984c). The consequences of adjustment for a concomitant variable

that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, 147,* 656–666.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D.B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics, 2,* 463–472.

Rubin, D.B., Thayer, D. (1978). Relating tests given to different samples. *Psychometrika, 43,* 3–10.

Sobel, M.E. (1994). Causal inference in latent variable models. In A. von Eye u. C.C. Clogg (Hg.), *Latent variables analysis* (S. 3–35). Thousand Oaks, CA: Sage.

Spohn, W. (1980). Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic, 9,* 73–99.

Spohn, W. (1983). *Deterministic and probabilistic reasons and causes* (Rep. No. 19). Dordrecht u. Boston: D. Reidel.

Spohn, W. (1990). *Direct and indirect causes* (Rep. No. 9). Dordrecht: Kluwer.

Spohn, W. (1991). A reason for explanation: Explanations provide stable reasons. In W. Spohn, B.C. van Fraassen u. B. Skyrmes (Hg.), *Existence and explanation* (S. 165–196). Dordrecht: Kluwer.

Stegmüller, W. (1983). *Erklärung, Begründung, Kausalität (Probleme und Resultate der Wissenschaftstheorie and Analytischen Philosophie, Band I.).* Berlin: Springer-Verlag.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle.* Stuttgart: G. Fischer.

Steyer, R. (2003). *Wahrscheinlichkeit und Regression.* Berlin: Springer-Verlag.

Steyer, R., Gabler, S., Rucai, A.A. (1996). Individual causal effects, average causal effects, and unconfoundedness in regression models. In F. Faulbaum u. W. Bandilla (Hg.), *Softstat '95. Advances in statistical Software 5* (S. 203–210). Stuttgart: Lucius & Lucius.

Steyer, R., Gabler, S., von Davier, A.A., Nachtigall, C. (2000a). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research – Online, 5,* 55–86.

Steyer, R., Gabler, S., von Davier, A.A., Nachtigall, C., Buhl, T. (2000b). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research – Online, 5,* 39–71.

Steyer, R., Nachtigall, C., Wüthrich-Martone, O., Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research – Online, 7,* 41–68.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North-Holland.

Waldmann, M.R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 53–76.

Waldmann, M.R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychometric Bulletin & Review, 8,* 600–608.

Whittaker, J. (1990). *Graphical models in applied multivariable statistics.* Chichester: Wiley.

Wright, S. (1918). On the nature of size factors. *Genetics, 3,* 367–374.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–585.

Wright, S. (1923). The theory of path coefficients – A reply to Niles' criticism. *Genetics, 8,* 239–255.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5,* 161–215.

Wright, S. (1960a). Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics, 16,* 189–202.

Wright, S. (1960b). The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics, 16,* 423–445.

Wüthrich-Martone, O. (2001). *Causal modeling in psychology with qualitative independent variables.* Aachen: Shaker.

Edgar Erdfelder, Axel Buchner, Franz Faul und Martin Brandt

GPOWER: Teststärkeanalysen leicht gemacht

Zusammenfassung

Statistische Tests können mehr Schaden anrichten als Nutzen bringen, wenn die Stärke (*Power*) eines Tests in Anwendungen unberücksichtigt bleibt. GPOWER[1] ist ein Computerprogramm, das die Durchführung von Teststärkeanalysen für ein sehr breites Spektrum gebräuchlicher statistischer Tests erlaubt. Allerdings ist das Programm in der bisherigen Form nur bei Standardtests leicht anwendbar; Tests, die nicht zu dieser Klasse zählen, erfordern Zusatzkenntnisse, die in diesem Beitrag für eine Reihe von *z*-Tests und *F*-Tests vermittelt werden.

Die Notwendigkeit der Teststärkekontrolle

Bei statistischen Entscheidungen zwischen einer Nullhypothese (H_0) und einer Alternativhypothese (H_1) können zwei Arten von Fehler auftreten: Zum einen kann H_0 gelten, aber es wird fälschlich zugunsten von H_1 entschieden (Fehler 1. Art oder α-Fehler), zum anderen kann H_1 gelten, aber es wird fälschlich zugunsten von H_0 entschieden (Fehler 2. Art oder β-Fehler). Lange Zeit hat in der Psychologie der auf Sir Ronald Aylmer Fisher zurückgehende Nullhypothesen-Signifikanztest dominiert, der ausschließlich die α-Fehlerwahrscheinlichkeit kontrolliert und β sowie dessen Komplement $1 - \beta$, die so genannte Stärke oder *Power* des statistischen Tests, unberücksichtigt lässt. Dies hat zur Konsequenz, dass Nullhypothesen zwar mit kontrollierten Fehlerraten verworfen, aber nie angenommen werden können. Bredenkamp (1969, 1972) hat gezeigt, dass dieser Sachverhalt insbesondere dann fatale Konsequenzen hat, wenn wissenschaftliche Hypo-

1 GPOWER ist ein Freeware-Computerprogramm. Es darf nicht zu kommerziellen Zwecken weiterverbreitet werden. Die jeweils aktuellen Versionen kann man kostenlos für verschiedene Plattformen beziehen über die URL http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

thesen mittels statistischer Tests geprüft werden. Dies ist in der psychologischen Forschung nach wie vor der Regelfall. Meistens implizieren psychologische Hypothesen statistische Alternativhypothesen, etwa in Form einer Erwartungswertdifferenz zwischen einer Experimental- und einer Kontrollgruppe (H_1: $\mu_1 > \mu_2$). Ist dies der Fall, so können wir die Alternativhypothese zwar bei signifikantem Testresultat als bewährt betrachten, nicht aber zurückweisen. Dies ist im Rahmen einer deduktivistischen Methodologie, die den Erkenntnisfortschritt an die Elimination falscher Hypothese koppelt (Popper, 1934/1982), völlig unakzeptabel. Keineswegs unproblematischer ist der seltenere Fall, dass die psychologische Hypothese eine Nullhypothese impliziert (z. B. H_0: $\mu_1 = \mu_2$). Eine solche Hypothese kann im Rahmen des Fisherschen Signifikanztests keine fundierten Bewährungsurteile erfahren, sodass unklar bleibt, woran man ihren Erfolg eigentlich messen soll.

Der Verzicht auf Teststärkekontrollen hat äußerst bedenkliche Konsequenzen für die Publikationspraxis psychologischer Fachzeitschriften: Signifikante Befunde, die eine H_1 stützen, lassen sich vergleichsweise leicht publizieren, nicht signifikante Ergebnisse dagegen nur schwer oder gar nicht. Die Folge ist eine systematische Verzerrung der Befundlage zugunsten von H_1-konformen Hypothesen und gegen H_0-konforme Hypothesen (Bredenkamp, 1972, 1980). Folglich ist zu erwarten, dass sich unter den publizierten Befunden ein erheblicher Anteil von α-Fehlern verbirgt, die auch nicht aufgedeckt werden, weil gescheiterte Replikationen mit nicht signifikanten Befunden nur minimale Chancen haben, publiziert zu werden. Eine weitere, eng damit zusammenhängende Konsequenz ist, dass gelegentlich vermeintliche Effektunterschiede für verschiedene abhängige Variable publiziert und psychologisch interpretiert werden, hinter denen möglicherweise gar keine realen Effektunterschiede stecken. Buchner und Wippich (2000) sowie Buchner und Brandt (2003) haben beispielsweise gezeigt, dass die in der Forschung zum impliziten Gedächtnis gerne inhaltlich interpretierten einfachen Dissoziationen zwischen expliziten und impliziten Gedächtnistests auch dadurch zustande kommen können, dass statistische Tests auf Unterschiede zwischen Bedingungen auf der Basis von typischen expliziten Gedächtnistests deutlich teststärker sind als Unterschiedstests auf der Basis von impliziten Gedächtnistests. Wenn also eine bestimmte experimentelle Manipulation Effekte auf Maße aus expliziten, nicht aber auf Maße aus impliziten Gedächtnistests zeigt, dann muss das nicht daran liegen, dass die interessierende experimentelle Manipulation nur jene Strukturen oder Prozesse betrifft, die einem expliziten Gedächtnis zugeordnet werden. Der Grund kann auch sein, dass der Unterschiedstest

auf der Basis der Daten aus dem impliziten Gedächtnistest wegen dessen geringerer Reliabilität viel zu wenig Teststärke besaß.

Zusammenfassend bleibt festzuhalten, dass Signifikanztests, welche Teststärkegesichtspunkte ignorieren, Gefahr laufen, »... worse than nothing« zu sein (J. Neyman, zitiert nach Gigerenzer u. Murray, 1987).

Wie lassen sich Teststärkekontrollen realisieren?

Zwei Einwände haben der routinemäßigen Anwendung von Teststärkeanalysen bei Signifikanztests in der Vergangenheit im Weg gestanden: der Einwand, dass Teststärkekontrollen oft unmöglich seien (z. B. Wottawa, 1981, S. 157) sowie der häufig unter der Hand geäußerte Einwand, dass Teststärkekontrollen zu aufwändig und damit letztlich unpraktikabel seien. Der erste Einwand stützt sich auf das Argument, dass oft unklar sei, welche Punkthypothese eigentlich als H_1 fungiere beziehungsweise welche Effektgröße für die zugrunde liegende Population zu veranschlagen sei. Die Festlegung einer geeigneten Effektgröße kann jedoch im Regelfall unter Erwägung der praktischen Bedeutung, die eine entsprechende Abweichung von der H_0 hat, auf rein apriorischer Basis erfolgen, gegebenenfalls unter Berücksichtigung der Effektstärkekonventionen von Cohen (1988) sowie der Ergebnisse verwandter Studien, die zuvor durchgeführt wurden. Die Effektgrößenspezifierung verlangt lediglich eine Entscheidung darüber, ab welcher Effektgröße die H_0 bei kontrollierter Fehlerrate β zurückgewiesen werden soll oder – anders ausgedrückt – bis zu welcher Effektgröße man eine Abweichung von der H_0 für noch vernachlässigbar hält.

Gewarnt werden muss vor der so genannten retrospektiven Teststärkeanalyse, bei der die Effektstärke aus denselben Stichprobendaten geschätzt wird, für die auch die Teststärkeanalyse durchgeführt werden soll. Fehlerhaft ist dieses Vorgehen, das beispielsweise im Programmsystem SPSS als »Observed Power« (beobachtete Teststärke) bezeichnet wird, in erster Linie deshalb, weil die Effektstärke in der Stichprobe je nach Stichprobenumfang erheblich von der wahren Effektstärke in der zugrunde liegenden Population abweichen kann und im Regelfall noch nicht einmal ein erwartungstreuer Schätzer der Populationseffektstärke ist. Besonders bei kleineren Stichprobenumfängen können die Ergebnisse also sehr irreführend sein. Aber selbst wenn man von dieser Problematik absieht, stellt sich die Frage, warum ausschließlich die vorliegenden Daten darüber entscheiden sollten, wie groß die zu entdeckende Effektstärke sein muss. Die so bestimmte Ef-

fektstärke und die daraus errechnete Teststärke wären dann ja völlig losgelöst von den Kosten- und Nutzenaspekten, die mit statistischen Entscheidungen verbunden sind. Derartige Teststärkekontrollen verlieren unseres Erachtens ihren Sinn.

Ist der zweite Einwand gegen Teststärkekontrollen berechtigt? Ob der für Teststärkeanalysen notwendige Aufwand im Regelfall wesentlich über dem Aufwand bei konventionellen Fisherschen Signifikanztests liegt, kann auch für die Vergangenheit mit Fug und Recht bezweifelt werden. Bredenkamp (1969) hat bereits vor 35 Jahren eine leicht anwendbare approximative Formel zur Berechnung des minimalen Stichprobenumfangs bei Mittelwertsvergleichen publiziert, mit dem eine gewünschte Teststärke bei vorgegebener Effektstärke garantiert werden kann. Bredenkamps (1969) Approximationsformel funktioniert trotz ihrer einfachen Anwendbarkeit »amazingly well even for very small sample sizes«, wie ihr Kupper und Hafner (1989, S. 101) im *American Statistician* in einem vergleichenden Übersichtsreferat zu verschiedenen Stichprobenumfangsformeln bescheinigt haben,[2] und sie deckt außerdem den dominierenden Anwendungsfall statistischer Tests in der Psychologie ab (Zwei-Gruppen-*t*-Test). Trotzdem sind Anwendungen dieser Formel wie auch später entwickelter Techniken der Teststärkekontrolle (z. B. Cohen, 1988) in der psychologischen Forschung rar geblieben.

Dass wir uns vor etwa zehn Jahren trotz dieser wenig motivierenden Sachlage entschlossen haben, mit dem Programm GPOWER ein weiteres Werkzeug der Teststärkeanalyse zu entwickeln (Erdfelder et al., 1996), liegt in der Einsicht begründet, dass sich viele Anwendungsprobleme von Teststärkeanalysen mit einem PC-basierten Computerprogramm optimal lösen lassen. Vor allem drei Gründe sprechen für GPOWER. Erstens kann sich die Notwendigkeit sehr genauer Powerberechnungen ergeben, insbesondere wenn α oder β bei multiplen Tests zum gleichen Datensatz adjustiert werden müssen (Hager, 1992; Westermann u. Hager, 1986). GPOWER kann dies problemlos leisten, während Approximationsformeln oder entsprechende Tabellenwerke (z. B. Cohen, 1988) zu ungenau sind. Zweitens lässt sich das Praktikabilitätsargument nur mit einem sehr allgemeinen und leicht handhabbaren Instrument überzeugend entkräften, das – wie GPOWER – Teststärkeanalysen für eine breite Palette gebräuchlicher statistischer Tests auf den verbreiteten Computerplattformen realisieren kann.

2 Die Vollständigkeit gebietet es, an dieser Stelle hinzuzufügen, dass Kupper und Hafner (1989) die Formel von Bredenkamp (1969) zwar darstellen und kommentieren, ihn aber nicht als Urheber zitieren, sondern stattdessen vier englischsprachige Arbeiten, die in den Jahren 1983 bis 1988 publiziert wurden.

Drittens ist der Standardfall bisheriger Teststärkeanalysen, die so genannte A-priori-Poweranalyse (Cohen, 1988), die den erforderlichen Stichprobenumfang zur Realisierung einer gewünschten Teststärke 1-β bei vorgegebenem α-Niveau vor der Durchführung einer Untersuchung berechnet, zwar der Idealtypus der Teststärkekontrolle, keineswegs aber die einzig sinnvolle Möglichkeit, Teststärkeüberlegungen in statistische Entscheidungen oder in die Bewertung statistischer Entscheidungen anderer Forschergruppen einfließen zu lassen. Je nach Sachlage können noch vier weitere Formen der Teststärkeanalyse wertvolle Dienste leisten, die sich nur mithilfe von Computerprogrammen sinnvoll realisieren und kombinieren lassen: Post-hoc-Analysen (Cohen, 1988), Kompromissanalysen (Erdfelder et al., 1996), Sensitivitätsanalysen (z. B. Welsch et al., 2000, Kap. 5.4) und Kriteriumsanalysen. A-priori-, Post-hoc- und Kompromissanalysen sind bereits als Auswahloptionen in GPOWER 2.0 implementiert, während die zuletzt genannten Analysen bislang nur durch iterative Anwendung von Post-hoc-Analysen realisiert werden können. Für Version 3.0 von GPOWER sind Sensitivitäts- und Kriteriumsanalysen aber ebenfalls als Auswahloptionen vorgesehen.

Formen der Teststärkeanalyse

Die fünf Formen der Teststärkeanalyse decken unterschiedliche Anwendungsfälle ab. *A-priori-Analysen* ermöglichen die Planung von Studien und Experimenten, die vorgegebene Güteeigenschaften der statistischen Entscheidungen erfüllen, das heißt α, β und die erwartete Stärke des Effekts in der Population stehen fest und gesucht wird die dafür notwendige Stichprobengröße *N*. *Post-hoc-Analysen* liefern als Ergebnis die Teststärke 1 – β beziehungsweise die Wahrscheinlichkeit eines β-Fehlers zu einem gegebenen *N*, dem gewählten α und einer bestimmten Effektstärke. Meist werden diese Analysen eingesetzt, um zu prüfen, mit welcher Fehlerwahrscheinlichkeit eine Entscheidung zugunsten einer H_0 gefallen ist.

Kompromiss-Analysen können dann sinnvoll sein, wenn das verfügbare *N* gemessen am idealen *N* zu klein ist, etwa bei klinischen Studien zu einem seltenen Krankheitsbild. Um der Alternativhypothese eine faire Chance zu geben, wird ein Kompromiss zwischen α und der Teststärke 1 – β angestrebt. Dazu wird zunächst entschieden, wie wichtig β im Vergleich zu α sein soll. Anschließend bestimmt man auf der Basis dieser Gewichtung und des verfügbaren *N* nicht nur β, sondern simultan auch α und den damit

verbundenen Wert der Teststatistik. Ein weiterer interessanter Anwendungsfall für Kompromiss-Analysen ist bei Modellanpassungstests gegeben, wo das verfügbare N bisweilen »zu groß« sein kann, sodass bereits triviale Abweichungen der vom Modell implizierten von der empirischen Datenstruktur zu einer Zurückweisung des Modells führen würden. In diesem Fall legt man über die Effektgröße fest, welche Abweichung man tolerieren möchte. Auf dieser Basis und aufgrund des vorliegenden N und der gewünschten Gewichtung von β relativ zu α liefert die Analyse dann die Werte für β sowie für α und den damit verbundenen Wert der Teststatistik, bei dessen Überschreiten das Modell zurückgewiesen wird.

Sensitivitäts-Analysen liefern eine Antwort auf die Frage, wie groß die Stärke eines Effekts sein muss, damit man diesen mit einem bestimmten begrenzten N und mit bestimmten idealen Fehlerwahrscheinlichkeiten α und β entdecken kann. Das Ergebnis einer solchen Analyse kann durchaus sein, dass man einen bestimmten interessierenden Effekt mit den gegebenen Ressourcen und den vorliegenden Ansprüchen an die Fehlerwahrscheinlichkeiten gar nicht entdecken kann, weil man annehmen muss, dass der interessierende Effekt deutlich kleiner ist als der unter den gegebenen Voraussetzungen entdeckbare Effekt. Wie beurteilt man, ob ein interessierender Effekt »zu klein« ist? Wie üblich kommen hierfür entweder Effektstärkekonventionen oder die Ergebnisse verwandter Studien infrage. Beispielsweise müsste in einer Zwei-Gruppen-t-Test-Situation bei $\alpha = \beta = .05$ und einem maximal verfügbaren $N = 30$ ($n_1 = n_2 = 15$) der entdeckbare Gruppenunterschied eine Effektstärke von $d = 1.23$ besitzen. Wenn wir in Anlehnung an Cohen (1988) für unsere Fragestellung höchstens »mittlere« Effekte, das heißt Effekte der Größe $d = 0.5$ in Bezug auf den Gruppenunterschied erwarten, ist klar, dass unsere begrenzten Ressourcen und unsere Ansprüche an die Fehlerwahrscheinlichkeiten zu einer Testsituation führen würden, die nicht sensitiv genug ist, um einen solchen Effekt zu entdecken. In diesem Fall kann man sich zum einen entscheiden, eine solche Untersuchung nicht durchzuführen. Das spart immerhin Zeit und Geld. Alternativ kann man mithilfe einer Kompromiss-Analyse prüfen, ob sich unter den gegebenen Ressourcenbegrenzungen statt der idealen immer noch akzeptable Wahrscheinlichkeiten für α- und β-Fehler finden lassen.

Kriteriumsanalysen schließlich liefern den kritischen Wert der jeweiligen Teststatistik (und damit auch das α-Fehlerniveau) zu vorgegebenem Stichprobenumfang sowie gegebener Effektstärke und Teststärke. Diese Art der Analyse ist insbesondere bei der Durchführung so genannter Äquivalenztests hilfreich. Äquivalenztests sind vor allem in der biometrischen Pharmaforschung gebräuchlich, in der man häufig vor dem Problem steht, die

Effektivitätsäquivalenz eines neuen, kostengünstigen Medikaments mit einem etablierten, teuren Medikament nachzuweisen. Ganz ähnliche Probleme können sich aber auch in der Psychologie ergeben, wenn man etwa mit einem Experimental-Kontrollgruppenvergleich zeigen will, dass eine kostengünstige Kurzformpsychotherapie genauso effektiv wie eine teure etablierte Langzeittherapie wirkt (Rogers et al., 1993). Der Grundgedanke hierbei ist, die Null- und die Alternativhypothese eines gewöhnlichen Zwei-Gruppen-*t*-Tests gewissermaßen gegeneinander auszutauschen: H_0 besagt, dass die beiden Gruppen nicht äquivalent sind, sodass die mittlere Effektivität der etablierten Behandlung (μ_e) die mittlere Effektivität der neuen Behandlung (μ_n) mindestens um einen klein gewählten Grenzwert Δ überschreitet (H_0: $\mu_e - \mu_n \geq \Delta$). Die Alternativhypothese der Äquivalenz behauptet hingegen, dass die Effektivität der etablierten Behandlung gar nicht oder allenfalls nur minimal über die der neuen Behandlung dominiert (H_1: $\mu_e - \mu_n < \Delta$). Diese Hypothesen lassen sich mit einem modifizierten einseitigen *t*-Test für zwei unabhängige Gruppen gegeneinander testen. Wenn die Untersuchung bereits durchgeführt wurde und die Stichprobenumfänge in den beiden Gruppen bereits feststehen, stellt sich die Frage nach dem kritischen Wert und der Entscheidungsregel, mit der zwischen H_0 und H_1 bei kontrollierten Fehlerwahrscheinlichkeiten zu entscheiden ist. Kriteriumsanalysen helfen hier weiter. Die Verteilung unter H_0 ist in diesem Fall eine nonzentrale *t*-Verteilung zur Effektgröße $d = \Delta/\sigma$, wobei σ die (ggf. erwartungstreu zu schätzende) Populationsstreuung in den beiden Gruppen ist. Will man den Äquivalenztest bei $\alpha = .05$ durchführen, ist in GPOWER als geforderte »power« $1 - \alpha = .95$ vorzugeben, womit wunschgemäß sichergestellt ist, dass 95 Prozent der nonzentralen Verteilung zur Punkthypothese $\mu_e - \mu_n = \Delta$ oberhalb vom kritischen Wert und 5 Prozent unterhalb vom kritischen Wert liegen. Berechnet man nun eine Kriteriumsanalyse, so erhält man mit α den Anteil der zentralen *t*-Verteilung rechts vom Entscheidungskriterium. Die Teststärke des Äquivalenztests, also die Wahrscheinlichkeit, bei faktisch vorliegender Äquivalenz ($\mu_e - \mu_n = 0$) auch korrekt zugunsten der Äquivalenzhypothese zu entscheiden, ist folglich gerade das Komplement des in der Kriteriumsanalyse errechneten α-Werts.

Kriteriumsanalysen ermöglichen die Durchführung von Äquivalenztests für prinzipiell alle Testfamilien, nicht nur für Zwei-Gruppen-*t*-Tests. Zu beachten ist lediglich, dass die Größe, die in GPOWER als *Power* bezeichnet wird, bei Äquivalenztests gerade $1 - \alpha$ entspricht, während umgekehrt die Größe, die GPOWER als *alpha* bezeichnet, bei Äquivalenztests gerade der β-Fehlerwahrscheinlichkeit entspricht.

Spezielle Teststärkeanalysen mit GPOWER

Das Programm GPOWER ist als allgemeines Programm für Teststärkeanalysen konzipiert worden (Erdfelder et al., 1996). Tatsächlich kann das Programm die im letzten Abschnitt beschriebenen Formen der Teststärkeanalyse für alle statistischen Tests durchführen, deren Prüfgrößen unter der Null- und der Alternativhypothese einer Normalverteilung, einer (nonzentralen) χ^2-Verteilung, einer (nonzentralen) t-Verteilung oder einer (nonzentralen) F-Verteilung folgen. Fast alle in den Sozial- und Verhaltenswissenschaften geläufigen Tests erfüllen diese Bedingung.[3] Die konkrete Handhabung des Programms ist allerdings nur für einige häufig vorkommende Standardtests einfach. Diese Fälle sind von Erdfelder et al. (1996) sowie Buchner et al. (1996) an konkreten Beispielen erläutert worden. Im Folgenden soll deshalb ausschließlich auf einige spezielle Testprobleme eingegangen werden, deren Lösung mit GPOWER Zusatzwissen verlangt, das bislang in der Literatur nicht dargestellt wurde.

Spezielle Tests mit normalverteilten Teststatistiken

In diesem Abschnitt werden Teststärkeanalysen für Proportions- und Korrelationstests dargestellt. Der Vorzeichentest *(sign test)* ist als Spezialfall eines Proportionstests ebenfalls eingeschlossen.

Tests zu einer Proportion

Nehmen wir an, in einer Stichprobe des Umfangs N sei die Proportion eines bestimmten Merkmals erhoben worden, das heißt die relative Häufigkeit x/N, mit der eine dichotome Variable (z. B. Erfolg vs. Misserfolg einer Psychotherapie) eine bestimmte Ausprägung (z. B. Erfolg) annimmt. Zugrunde gelegt wird ein Binomialverteilungsmodell, das heißt, die N Beobachtungen seien unabhängig aus derselben Grundgesamtheit gezogen worden. Getestet werden soll beispielsweise die einseitige Nullhypothese, dass die Proportion höchstens den Wert p_0 annimmt. Unter der Alternativhypothese wird entsprechend ein größerer Wert p_1 für die Proportion angenom-

3 Für einige Tests gilt dies allerdings nur asymptotisch, sodass die Powerberechnungen bei kleinem Stichprobenumfang als Approximationen zu werten sind.

men. So kann etwa von Interesse sein, ob die Erfolgsquote einer neuen Psychotherapieform den Wert $p_0 = .60$ einer etablierten Therapie nicht überschreitet (H_0: $p \leq p_0$) oder aber größer ist (H_1: $p > p_0$). Als zu entdeckende substanzielle Verbesserung werde z. B. $p_1 = .80$ festgelegt.

Unter H_0 ist dann bekanntlich die Prüfstatistik

$$z = \frac{x - N \cdot p_0}{\sqrt{N \cdot p_0 \cdot (1-p_0)}}$$

asymptotisch standardnormal verteilt (z. B. Hays, 1972, S. 305). Gilt H_1, so ist die gleiche Statistik ebenfalls normal verteilt, allerdings mit der Streuung

$$\sigma_1 = \frac{\sqrt{N \cdot p_1 \cdot (1-p_1)}}{\sqrt{N \cdot p_0 \cdot (1-p_0)}}$$

und dem Erwartungswert

$$\mu_1 = \frac{N \cdot (p_1 - p_0)}{\sqrt{N \cdot p_0 \cdot (1-p_0)}} = \sqrt{N} \cdot \frac{p_1 - p_0}{\sqrt{p_0 \cdot (1-p_0)}}.$$

Post-hoc-Teststärkeanalysen für diesen Fall lassen sich mit GPOWER mittels der Prozedur »Other *t*-Tests« durchführen. Da die nonzentrale *t*-Verteilung mit Nonzentralitätsparameter δ für $df \to \infty$ in eine Normalverteilung mit Mittelwert δ und Streuung 1 übergeht, setzt man zunächst $df = 32000$, das heißt den maximal von GPOWER zugelassenen Eingabewert, und wählt einen einseitigen Test (»one-tailed«). Der Nonzentralitätsparameter δ entspricht in der Prozedur »Other *t*-Tests« dem Produkt der Parameter $\sqrt{N}$ und f (= Effektstärke). Man wählt also N und f in GPOWER so, dass das Produkt von $\sqrt{N}$ und f gerade μ_1 entspricht, dem Erwartungswert der Teststatistik unter H_1. Für N ist folglich der Stichprobenumfang einzusetzen und für die Effektstärke

$$f = \frac{p_1 - p_0}{\sqrt{p_0 \cdot (1-p_0)}}.$$

Ein letztes zu lösendes Problem betrifft die Tatsache, das GPOWER bei »Other *t*-tests« mit $df = 32000$ von einer Streuung $\sigma = 1$ unter H_1 ausgeht, während im gegebenen Fall die tatsächliche Streuung – wie oben ausgeführt – σ_1 beträgt. Dieses Problem löst man dadurch, dass man als Signifikanzniveau α nicht das faktisch gewünschte Niveau eingibt, sondern ein adjustiertes Niveau α', welches so gewählt wird, dass der kritische Wert der *z*-Statistik zum Niveau α', abgekürzt $z(\alpha')$, zum entsprechenden kritischen Wert für das nominelle α-Niveau, $z(\alpha)$, wie folgt in Beziehung steht:

$z(\alpha') = z(\alpha)/\sigma_1$.

Diese Adjustierung garantiert, dass die korrekte Teststärke berechnet wird. Da es sich um einen asymptotischen Test handelt, ist natürlich vorauszusetzen, dass der Stichprobenumfang nicht zu klein ist. Die Übertragung auf den Fall zweiseitiger Tests liegt auf der Hand: Hier sind zwei Ablehnungsbereiche für H_0 zu betrachten statt eines Bereichs, was durch Anwählen der Option »two-tailed« geschieht.

Man beachte, dass die geschilderte Prozedur auch asymptotische Poweranalysen für den Vorzeichentest (*sign test*) korrekt durchführt, weil dieser nichts anderes ist als ein Proportionstest mit der Nullhypothese H_0: $p = .50$ (vgl. Hays, 1972, S. 193–197).

Betrachten wir hierfür ein konkretes Rechenbeispiel. Es seien $N = 100$ Fälle beobachtet worden. Die Erfolgsquote unter H_0 betrage $p_0 = .50$, unter H_1 dagegen $p_1 = .70$. Wie groß ist die Power des einseitigen Vorzeichentests bei $\alpha = .05$? Wie gesagt ist für *df* die Obergrenze 32000 und für *N* der Wert 100 im Menüpunkt »Other *t*-Tests«, »Post hoc Analysis«, festzulegen. Setzt man die genannten p_0- und p_1-Werte in die oben genannte *f*-Formel ein, ergibt sich der für die Effektgröße einzugebende Wert $f = .33806$. Für alpha ist in GPOWER 2.0 aus den genannten Gründen aber nicht .05 zu wählen. Der zu $\alpha = .05$ (einseitig) gehörige *z*-Wert von 1.649 ist zunächst durch σ_1 zu dividieren, was den adjustierten *z*-Wert 1.7947 ergibt. Der zugehörige adjustierte α-Wert (einseitig) ist etwa .0364. Setzt man ihn für α in GPOWER ein und klickt auf »Calculate«, so erhält man die erfreuliche Powerschätzung von $1 - \beta = .9437$.

Tests zu einer Korrelation

Werden in einer Stichprobe der Größe *N* zwei quantitative Variablen *X* und *Y* (z. B. die Konzentrationsfähigkeit und der IQ der Probanden) gemessen, die in der Population bivariat normal verteilt sind, können Hypothesen über die Korrelation der beiden Variablen getestet werden. Häufig steht man vor der Situation, dass das Ausmaß des linearen Zusammenhangs getestet werden soll. So könnte aufgrund einer Theorie behauptet werden, dass die Konzentrationsfähigkeit und die Intelligenz in der Population mit .50 korrelieren. In diesem Fall soll die H_0: $\rho_{XY} = \rho_0 = .50$ getestet werden. Die Differenz zwischen der gemäß der Fisherschen *r*-nach-*z*-Formel transformierten Korrelation r_{XY} in der Stichprobe und der *z*-transformierten H_0-Korrelation in der zugrunde liegenden Population,

$$0.5 \cdot \ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right) - 0.5 \cdot \ln\left(\frac{1+\rho_0}{1-\rho_0}\right),$$

ist unter der H_0 approximativ normal verteilt mit dem Erwartungswert Null und Varianz $1/(N-3)$ (Hays, 1972, S. 662–663). Unter einer bestimmten Alternativhypothese (H_1: $\rho_{XY} = \rho_1$) ist diese Differenz ebenfalls normal verteilt mit der gleichen Varianz, allerdings mit dem Erwartungswert

$$\mu_1 = 0.5 \cdot \ln\left(\frac{1+\rho_1}{1-\rho_1}\right) - 0.5 \cdot \ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = q_2',$$

welcher genau dem Effektstärkemaß q_2' nach Cohen entspricht (Cohen, 1988). Multipliziert man die z-Differenzen mit dem Kehrwert der Streuung $1/\sqrt{1/(N-3)}$, erhält man standardnormal verteilte Variablen. Unter der H_0 ist der Erwartungswert weiterhin Null, während er unter der H_1 nun $q2'/\sqrt{1/(N-3)} = q2' \cdot \sqrt{N-3}$ ist. Die Teststärkeberechnung für diesen Fall ist mit GPOWER nun recht einfach. Zunächst sind in der Prozedur »Other t-Tests« wieder 32000 Freiheitsgrade zu wählen, sodass die t-Verteilung eine Normalverteilung approximiert. Da der Nonzentralitätsparameter δ in dieser Prozedur als Produkt von Effektstärke f und $\sqrt{N}$ errechnet wird, erhält man den gewünschten Wert $\delta = q_2' \cdot \sqrt{N-3}$, wenn man für die Effektstärke den Wert q_2' wie oben definiert und als Stichprobengröße den Wert $(N-3)$ eingibt. Man wählt zusätzlich noch das gewünschte Signifikanzniveau α und – je nach Fragestellung – einen einseitigen (»one-tailed«) oder zweiseitigen (»two-tailed«) Test, um die korrekte Power des Tests zu erhalten. Wie schon im Fall des im letzten Abschnitt geschilderten Proportionstests ist allerdings zu beachten, dass es sich um einen asymptotischen Test handelt, der nur bei nicht zu kleinem N hinreichend genau ist.

Nehmen wir an, die Nullhypothese einer Korrelation von $\rho_0 = .5$ soll gegen die Alternativhypothese einer Korrelation von $\rho_1 = .6$ bei $\alpha = .05$ einseitig getestet werden. Es wurden $N = 103$ Messwertpaare untersucht, sodass die Varianz der z-transformierten Korrelation 1/100 und die Streuung somit 0.10 beträgt. Das Effektmaß q_2' nach Cohen ist, wie man einer Fisher r-nach-z-Tabelle entnehmen kann (z. B. Hays, 1972), in diesem Fall $q_2' = .6931 - .5493 = .1430$. Man errechnet folglich mit GPOWER 2.0 die Power, indem man »Other t-Tests«, »Post-hoc analysis« mit $df = 32000$ auswählt, das gewünschte $\alpha = .05$ und »one-tailed« spezifiziert, für die Effektstärke den Wert $f = q_2' = .1430$ und für N das Inverse der Varianz (100) angibt. Man erhält dann durch Klicken auf »Calculate« die ernüchternde Powerbestimmung von nur $1 - \beta = .4149$.

Vergleiche von zwei unabhängigen Korrelationen

Nehmen wir an, dass in zwei unabhängigen Stichproben jeweils N Werte zweier quantitativer Variablen X und Y, die in beiden zugrunde liegenden Population jeweils bivariat normal verteilt sind, gemessen wurden. In diesem Fall kann die Hypothese geprüft werden, dass die beiden Populationskorrelationen identisch sind (H_0: $\rho_{1(XY)} = \rho_{2(XY)}$). Beispielsweise könnte man daran interessiert sein, ob ein Persönlichkeitsmerkmal in zwei klinischen Populationen gleich mit einer Kriteriumsvariablen korreliert. Da die z-Transformation einer Korrelation nach Fisher (s. Hays, 1972, S. 662–663) normal verteilt ist mit der Varianz $1/(N-3)$, ist die Teststatistik

$$z_1 - z_2 = 0.5 \cdot \ln\left(\frac{1+r_{1(XY)}}{1-r_{1(XY)}}\right) - 0.5 \cdot \ln\left(\frac{1+r_{2(XY)}}{1-r_{2(XY)}}\right)$$

unter der H_0 ebenfalls normal verteilt mit dem Erwartungswert Null und der Varianz $2/(N-3)$. Unter der H_1: $\rho_{1(XY)} \neq \rho_{2(XY)}$ ist die Statistik ebenfalls normal verteilt, allerdings mit dem Erwartungswert

$$\mu_1 = 0.5 \cdot \ln\left(\frac{1+\rho_{1(XY)}}{1-\rho_{1(XY)}}\right) - 0.5 \cdot \ln\left(\frac{1+\rho_{2(XY)}}{1-\rho_{2(XY)}}\right) = q$$

und der Varianz $2/(N-3)$. Man beachte, dass μ_1 gerade dem Effektstärkemaß q von Cohen (1988) entspricht.

Multipliziert man die z-Differenzen mit $1/\sqrt{2/(N-3)}$, erhält man standardnormal verteilte Variablen. Unter der H_0 ist der Erwartungswert weiterhin Null, während er unter der H_1 nun $q/\sqrt{2/(N-3)} = q \cdot \sqrt{(N-3)/2}$ ist. Die Teststärkeberechnung für diesen Fall ist mit GPOWER nun problemlos möglich. Zunächst sind in der Prozedur »Other t-Tests« erneut 32000 Freiheitsgerade zu wählen, sodass die t-Verteilung eine Normalverteilung approximiert. Der Nonzentralitätsparameter δ entspricht – wie bereits aus den letzten beiden Abschnitten bekannt – dem Produkt aus der Effektstärke f und $\sqrt{N}$. Um $\delta = q \cdot \sqrt{(N-3)/2}$ zu erhalten, wählt man also für die Effektstärke f den Wert q wie oben definiert und als Stichprobengröße den Wert $(N-3)/2$. Ansonsten verfährt man analog zum Test für eine Korrelation.

Für den Fall, dass die untersuchten Korrelationen auf ungleich großen Stichproben beruhen ($N_1 \neq N_2$) kann N über

$$N' = \frac{2(N_1-3)(N_2-3)}{N_1+N_2} + 3$$

bestimmt werden (s. Cohen, 1988).

Auch hier soll ein konkretes Beispiel das Vorgehen verdeutlichen. Bei $\alpha = .01$ soll die Nullhypothese gleicher Korrelationen in zwei Populationen, aus denen jeweils 103 Versuchspersonen zufällig entnommen wurden, gegen die Alternativhypothese getestet werden, dass die Korrelation in der ersten Population .70, in der zweiten Population dagegen .90 ist. Fisher *r*-nach-*z*-Tabellen (Hays, 1972) kann man entnehmen, dass hierfür $q = 1.472 - .8673 = .6047$ das Effektmaß bildet. Man wählt also wie in den vorherigen Beispielen in GPOWER 2.0 »Other *t*-Tests«, $df = 32000$, »Post-hoc analysis«, nunmehr aber $\alpha = .01$ und »two-tailed«. Für die Effektgröße ist $f = .6047$ anzugeben und für $N = (103 - 3)/2 = 50$. Durch Klicken auf Calculate erhält man dann als Powerberechnung den erfreulich hohen Wert von $1 - \beta = .9554$.

Spezielle Tests mit *F*-verteilten Teststatistiken

Multivariate Varianzanalysen

Bredenkamp und Erdfelder (1985) empfehlen als Teststatistik bei multivariaten Varianzanalysen das Pillai-Bartlett-*V* als multivariates Testkriterium. Bei Gültigkeit der H_0 gilt, dass

$$F = \frac{\frac{V(h)/s(h)}{df_1}}{\frac{1-V(h)/s(h)}{df_2}}$$

approximativ $F(df_1, df_2)$-verteilt ist. Hierbei repräsentiert $V(h)$ das Pillai-Bartlett-*V* für die zu testende Hypothese und $s(h)$ das Minimum aus $(p, n(h))$, wobei p die Anzahl der abhängigen Variablen und $n(h)$ die Anzahl der Prädiktoren für den zu testenden Effekt bezeichnet. Für die Freiheitsgrade gilt

$$df_1 = p \cdot n(h)$$

und

$$df_2 = s(h) \cdot (N - k - p + s(h)),$$

wobei N die Gesamtzahl aller Versuchspersonen in allen k Gruppen des Ver-

suchsplans repräsentiert. Der Term $V(h)/s(h)$ variiert zwischen 0 und 1 und kann als multivariates R^2 oder η^2 betrachtet werden. Ein einfaches Maß für die standardisierte multivariate Populationseffektgröße erhält man nach

$$f^2_{\text{mult}} = \frac{V(h)/s(h)}{1 - V(h)/s(h)} = \frac{V(h)}{s(h) - V(h)}.$$

Vergleiche mit Veröffentlichungen exakter Teststärketabellen (Pillai u. Jayachandran, 1967; Stevens, 1980) zeigen, dass mit GPOWER berechnete Teststärken sehr nahe an den exakten Teststärken liegen, wenn man annimmt, dass bei Gültigkeit der H_1 die Pillai-Bartlett-V-Statistik approximativ $F(df_1, df_2, \lambda)$-verteilt ist verteilt ist mit den oben genannten Freiheitsgraden und dem Nonzentralitätsparameter

$$\lambda = s(h) \cdot N \cdot f^2_{\text{mult}}.$$

Bei der Durchführung einer globalen MANOVA zur Prüfung der Frage, ob alle verwendeten Prädiktoren Varianz in den zwei oder mehr abhängigen Variablen aufklären, gilt $n(h) = k - 1$. Für spezielle MANOVAs zur Prüfung des Effekts einer Untermenge der Prädiktoren gilt $n(h)$ = Anzahl der zu testenden Prädiktoren. Für einen Haupteffekt A mit a Stufen gilt also $n(h) = a - 1$, für einen Haupteffekt B mit b Stufen gilt $n(h) = b - 1$ und für die $A \times B$-Interaktion gilt $n(h) = (a - 1) \cdot (b - 1)$.

Post-hoc-Teststärkeanalysen für diese Fälle lassen sich mit der Prozedur »Other F-Tests« in GPOWER durchführen. In dieser Prozedur wird der Nonzentralitätsparameter λ als Produkt aus Effektstärke f^2 und der Stichprobengröße N berechnet. Um das gewünschte $\lambda = s(h) \cdot N \cdot f^2_{\text{mult}}$ zu erhalten, wählt man für f^2 den Wert f^2_{mult} und als Stichprobenumfang $s(h) \cdot N$. Als Nenner- und Zählerfreiheitsgrade sind die oben angegebenen einzusetzen.

Multivariate Varianzanalysen für Messwiederholungspläne

Varianzanalysen für Messwiederholungspläne sind stets multivariate Varianzanalysen, bei denen die Stufen der messwiederholten Variablen als verschiedene abhängige Variable betrachtet werden (O'Brian u. Kaiser, 1985). Nehmen wir an, der im vorangegangenen Abschnitt erwähnte Faktor B repräsentiere keinen Gruppenfaktor, sondern eine messwiederholte Variable,

Faktor A jedoch weiterhin verschiedene Gruppen. Dann ergeben sich $p = b - 1$ abhängige Variablen, weil die b Stufen der Variablen B mit geeigneten orthogonalen Kontrastvariablen in $b - 1$ abhängige Variablen rekodiert werden.

Für den Test des Haupteffekts A gilt

$$df_1 = a - 1$$

und

$$df_2 = N - a,$$

wobei N die Anzahl der Versuchspersonen repräsentiert. Der Nonzentralitätsparameter ist definiert als

$$\lambda = N \cdot \left(\frac{m}{1+(m-1) \cdot \rho} \right) \cdot f^2,$$

wobei N die Gesamtzahl aller Versuchspersonen, m die Stufen der messwiederholten Variablen, f^2 den Effektgrößenindex für den univariaten ANOVA-Fall (Effektvarianz/Fehlervarianz, vgl. Cohen, 1988) und ρ die Populationskorrelation zwischen den Stufen der messwiederholten Variablen repräsentiert. Letzteres mag überraschen angesichts der Tatsache, dass es bei m Stufen einer messwiederholten Variablen $(m^2 - m)/2$ Korrelationen zwischen diesen gibt. Nach Stevens (1996) ist es jedoch vertretbar, für diesen Fall eine Art »durchschnittlicher« Korrelation zwischen den Stufen der messwiederholten Variable anzunehmen. Für $m = 1$ (also ohne Messwiederholung) reduziert sich die Berechnung des Nonzentralitätsparameters auf $\lambda = N \cdot f^2$, den Nonzentralitätsparameter für den univariaten ANOVA-Fall.

Für den F-Test für die messwiederholte Variable B ergeben sich die Freiheitsgrade nach

$$df_1 = b - 1$$

und

$$df_2 = s(h) \cdot (N - k - p + s(h)),$$

wobei N die Anzahl der Versuchspersonen und k die Anzahl der Gruppen des Versuchsplans – hier also die Stufen der Variablen A – repräsentiert.

Der Nonzentralitätsparameter ergibt sich approximativ nach

$$\lambda = N \cdot m \cdot \frac{f^2}{1-\rho}.$$

Der F-Test für die $A \times B$-Interaktion zwischen der Gruppiervariablen A und der messwiederholten Variablen B schließlich hat folgende Freiheitsgrade:

$$df_1 = (a-1) \cdot (b-1),$$

$$df_2 = (N-a) \cdot (b-1),$$

Der Nonzentralitätsparameter ist wie beim Haupteffekt der messwiederholten Variablen B definiert als

$$\lambda = N \cdot m \cdot \frac{f^2}{1-\rho}.$$

Die Vorgehensweise bei der Durchführung einer Post-hoc-Teststärkeanalyse mithilfe von GPOWER ist analog zu der im vorigen Abschnitt beschriebenen.

Varianzanalysen für Messwiederholungspläne – der so genannte »univariate« Fall

Ein Spezialfall der multivariaten Varianzanalysen für Versuchspläne mit Messwiederholungen ist die so genannte »univariate Varianzanalyse für Messwiederholungen« – eigentlich eine Fehlbenennung, denn Messwiederholungs-Varianzanalysen sind grundsätzlich multivariat. Was hier jedoch gemeint ist, ist eine multivariate Varianzanalyse für Messwiederholungspläne, bei der die restriktive und selten realistische Annahme gemacht wird, dass die Varianzen aller Stufen der messwiederholten Variablen homogen und die Korrelationen zwischen allen Stufen der messwiederholten Variablen identisch sind. Dies wird oft als die Spherizitätsannahme bezeichnet. Es gibt selten einleuchtende Gründe, warum man diese Annahme machen sollte, doch in einigen speziellen Situationen kann sie vielleicht akzeptabel sein: Wenn etwa die Reihenfolge der Messwiederholungen über die Versuchspersonen hinweg randomisiert wurde oder wenn mehr Stufen der abhängigen Variablen als Versuchspersonen vorliegen. Im zuletzt genannten Fall ist das multivariate Modell nicht identifiziert. Dann profitiert man davon, dass die erwähnten Restriktionen auf den Varianzen und Kovarianzen dazu führen, dass weniger Parameter geschätzt werden müssen und dadurch zusätzliche Freiheitsgrade zur Verfügung stehen. Entsprechend erge-

durch zusätzliche Freiheitsgrade zur Verfügung stehen. Entsprechend ergeben sich für unser bereits oben verwandtes Design die Freiheitsgrade für den *F*-Test für die messwiederholte Variable *B* nach

$$df_1 = b - 1$$

und

$$df_2 = (N - a) \cdot (b - 1)$$

wobei *N* wieder die Anzahl der Versuchspersonen bezeichnet. Der Nonzentralitätsparameter ergibt sich nach

$$\lambda = N \cdot m \cdot \frac{f^2}{1-\rho}.$$

Hier müssen wir nicht den Umweg über eine gedachte »durchschnittliche« Populationskorrelation ρ gehen, denn mit der Spherizitätsannahme postulieren wir ja gerade, dass es nur eines einzigen Werts bedarf, um sämtliche Interkorrelationen zwischen den Stufen der messwiederholten Variablen zu repräsentieren. Ist diese Annahme erfüllt, so ist der so genannte univariate Zugang teststärker als die im letzten Abschnitt beschriebene multivariate Analyse von Messwiederholungsplänen.

Ist die Spherizitätsannahme verletzt, muss eine Korrekturgröße ε berücksichtigt werden, für die zwei mögliche Schätzer mit leicht unterschiedlichen Eigenschaften existieren: Greenhouse-Geisser-$\hat{\varepsilon}$ und Huynh-Feldt-$\tilde{\varepsilon}$. Zur Korrektur werden bei der Entscheidung über die statistische Signifikanz eines Testergebnisses die Zähler- und Nennerfreiheitsgrade des *F*-Tests mit dem gewählten Schätzer für ε multipliziert und dann die Signifikanz auf der Basis des *F*-Werts für die reduzierten Werte $df'_1 = df_1 \cdot \hat{\varepsilon}$ und $df'_2 = df_2 \cdot \hat{\varepsilon}$ bewertet.

Zur Approximation der Teststärke des ε-korrigierten *F*-Tests haben Muller und Barton (1989) ein analoges Verfahren vorgeschlagen. Dabei wird zusätzlich auch der Nonzentralitätsparameter λ mit dem ε-Schätzer multipliziert und die Teststärkeanalyse mit df_1, df_2 und $\lambda' = \lambda \cdot \hat{\varepsilon}$ durchgeführt.

Für den Gruppierfaktor ändert sich gegenüber der multivariaten Varianzanalyse für messwiederholte Pläne dagegen nichts, sodass sich eine Teststärkeanalyse einfach an dem oben beschriebenen Vorgehen orientieren kann.

Die konkrete Vorgehensweise bei der Durchführung einer Post-hoc-Teststärkeanalyse mithilfe von GPOWER ist analog zu der oben beschriebenen.

Schlussfolgerungen

Teststärkeüberlegungen sind bei der Bewertung statistischer Entscheidungen, besser aber schon bei der Planung von Untersuchungen unumgänglich. Wie wir zu zeigen versucht haben, bieten verschiedene Formen von Teststärkeanalysen für eine breite Palette statistischer Tests ein sehr flexibles Instrumentarium sowohl zur rationalen Planung von Untersuchungen wie auch zur inferenzstatistischen Analyse oder Neubewertung bereits durchgeführter Experimente und Studien. Mit leicht zu handhabenden Werkzeugen wie GPOWER lassen sich diese Analysen problemlos umsetzen, sodass der in der Vergangenheit häufig geäußerte Einwand, Teststärkeanalysen seien aufwändig und daher unpraktikabel, nicht mehr überzeugen kann.

Literatur

Bredenkamp, J. (1969). Über die Anwendung von Signifikanztests bei theorie-testenden Experimenten. *Psychologische Beiträge, 11,* 275–285.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J., Erdfelder, E. (1985). Multivariate Varianzanalyse nach dem V-Kriterium. *Psychologische Beiträge, 27,* 127–154.

Buchner, A., Brandt, M. (2003). Further evidence for systematic reliability differences between explicit and implicit memory tests. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 56A,* 193–209.

Buchner, A., Erdfelder, E., Faul, F. (1996). Teststärkeanalysen. In E. Erdfelder, T. Meiser, R. Mausfeld u. G. Rudinger (Hg.), *Handbuch Quantitative Methoden* (S. 123–136). Weinheim: Psychologie Verlags Union.

Buchner, A., Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology, 40,* 227–259.

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

Erdfelder, E., Faul, F., Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers, 28,* 1–11.

Gigerenzer, G., Murray, D.J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen.* Göttingen: Hogrefe.

Hays, W.L. (1972). *Statistics for the social sciences* (2. Aufl.). London: Holt, Rinehart & Winston.

Kupper, L.L., Hafner, K.B. (1989). How appropriate are popular sample size formulas? *The American Statistician, 43,* 101–105.

Muller, K.E., Barton, C.N. (1989). Approximate power for repeated measures ANOVA lacking sphericity. *Journal of the American Statistical Association, 84,* 549–555.

O'Brien, R.G., Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97,* 316–333.

Pillai, K.C.S., Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika, 54,* 195–210.

Popper, K.R. (1934/1982). *Logik der Forschung* (7. Aufl.). Tübingen: J.C.B. Mohr (im Original publiziert 1934).

Rogers, J.L., Howard, K.I., Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565.

Stevens, J.P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin, 88,* 728–737.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3. Aufl.). Mahwah, NJ: Erlbaum.

Welsch, W., Heunecke, O., Kuhlmann, H. (2000). *Auswertung geodätischer Überwachungsmessungen.* Heidelberg: Herbert Wichmann Verlag.

Westermann, R., Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11,* 117–146.

Wottawa, H. (1981). *Psychologische Methodenlehre* (2. Aufl.). München: Juventa.

Klaus Willmes

Permutations- und Randomisierungstests

The logical reason for randomization is that it is possible on the basis of a model reflecting the randomization to draw sound statistical inferences the probability basis of the model being provided not by wishful thinking but by the actual process of randomization which is part of the experiment (Scheffé, 1959)

Zusammenfassung

In diesem Beitrag wird eine Theorie über das Testen von linearen Hypothesen in linearen (ANOVA) Modellen ohne Verteilungsannahme mit Permutationstests vorgestellt. Exemplarisch werden Tests auf Wechselwirkungen in verschiedenen zweifaktoriellen Versuchsplänen mit und ohne Messwiederholungen konstruiert. Das Problem, von Haupteffekten nicht beeinflusste Permutationstests auf Wechselwirkungen zu entwickeln, ist nicht für alle Versuchspläne lösbar. Eine Verallgemeinerung des Konstruktionsprinzips auf die MANOVA ist leicht möglich. Spezielle Probleme bei der Konstruktion von multiplen Permutationstests für die Stufen eines experimentellen Faktors werden erläutert und ein generelles exaktes Permutationstestverfahren für die Komponenten mehrdimensionaler Beobachtungen präsentiert, welches die vorhandene Kovariation zwischen den Variablen vollständig ausnutzt.

Einleitung

Im Beitrag von Siemer sind die Logik des Randomisierungstests zum Testen von kausalen Hypothesen in so genannten »lokalen Populationen« und die wichtigen Eigenschaften von Randomisierungstests für eine deduktivistische Theorie des Experimentierens, wie sie Bredenkamp (1972, 1980) mit initiiert hat, dargestellt. Siemer hat zudem auf die konzeptuelle Unterscheidung von Randomisierungs- und Permutationstests hingewiesen, auch wenn beide zu gleichen Testentscheidungen führen. Hier soll detaillierter

auf die Konstruktion von Tests bezüglich linearer Hypothesen über Haupt- und vorwiegend Wechselwirkungen mit Permutations- und Randomisierungstests in mehrfaktoriellen linearen Modellen ohne spezifizierte Verteilungsannahmen eingegangen werden, auch wenn gezielte Hypothesen über (nicht-)orthogonale lineare Kontraste die per Implikation aus einer empirischen Hypothese abgeleiteten (einseitigen) statistischen (Alternativ-) Hypothesen als möglichst adäquate und erschöpfende Umsetzung einer statistischen Vorhersage darstellen (Hager, 1987, 2000, 2002). Auf Permutationstests für solche (Interaktions-)Kontraste wird aber auch kurz eingegangen.

Das Prinzip des Randomisierungs-/Permutationstests

Bei einem *Randomisierungstest* betrachtet man n Beobachtungseinheiten und die Menge aller M möglichen (zulässigen) Versuchsschemata, mit denen die n Beobachtungseinheiten einer von $k \geq 2$ experimentellen Versuchsbedingungen zugewiesen werden könnten. Alle M Schemata seien so beschaffen, dass bei jedem von ihnen jeweils dieselbe Anzahl von Einheiten einer der Bedingungen zugewiesen wird. Dann wird per Zufall ein Versuchsschema (i. Allg. haben alle die gleiche Wahrscheinlichkeit $1/M$ der Auswahl) gewählt und je Einheit ein beobachteter Wert y_i, $i = 1, \ldots, n$, registriert. Für einen Vergleich der k Bedingungen ist es unerheblich, ob die Beobachtungseinheiten zufällig aus einer Grundgesamtheit ausgewählt sind oder ob kein Bezug auf eine Grundgesamtheit genommen wird. Die übliche Nullhypothese in einem Randomisierungstest lautet dann (vgl. Willmes, 1996):

(2.1) H^R: Alle k Bedingungen haben für alle Einheiten *denselben* additiven Effekt.

Lässt man Messfehler unberücksichtigt, kann sich unter H^R in den Komponenten des Spaltenvektors $y = (y_1, \ldots, y_n)'$ Variabilität nur aufgrund von nicht im Experiment kontrollierbaren Einflüssen durch die n Beobachtungseinheiten selbst ergeben, welche in den additiven so genannten »unit values« u_i, $i = 1, \ldots, n$ erfasst wird. Die Zufallszuteilung der Beobachtungseinheiten zu den experimentellen Bedingungen, erfasst durch (diskrete) so genannte »Design-Variable« δ_{ji} mit:

(2.2) $\delta_{ji} = 1$ i-te Einheit wird der j-ten Bedingung zugewiesen mit $i = 1, \ldots, n; j = 1, \ldots, k$
$= 0$ sonst,

ist dann das einzige Zufallselement in $y_i = \mu + \alpha_i + \Sigma_l \delta_{jl}\, u_l$, $l = 1, \ldots, n$. Für die Testentscheidung wählt man als eine »geeignete« Teststatistik S etwa den Zähler der üblichen ANOVA F-Statistik, welcher große Werte annimmt, wenn unterschiedliche Effekte α_i der Versuchsbedingungen vorliegen. Die Teststatistik $S(y_1, \ldots, y_n)$ ist unter H^R aber nur noch eine Funktion der Design-Variablen und der Einheitenwerte $u = (u_1, \ldots, u_n)$ (vgl. Hooper, 1989). Zur Testentscheidung wird bei beobachtetem Wert s der Teststatistik S der p-Wert bestimmt:

(2.3) $p := p(y) = P(H^R \,\&\, u\colon S(y,\{\delta_{ji}\}) \geq s) = 1/M \cdot \text{card}\{\{\delta_{ji}\}\colon S(y,\{\delta_{ji}\}) \geq s)\}$,

wobei $\text{card}\{\delta_{ji}\}$ die Anzahl der möglichen Zufallszuteilungen (Permutationen) der Beobachtungseinheiten zu den Versuchsbedingungen bezeichnen möge. Der p-Wert in (2.3) gibt dann die relative Häufigkeit an, mit der die Teststatistik für andere Zufallszuteilungen (permutierte Beobachtungen) einen gleich großen oder größeren Wert annimmt als für die tatsächlich realisierte Zufallszuteilung (nicht permutierten Beobachtungen).

Die Testentscheidung lautet dann zu einem vorgegebenen α mit $0 < \alpha < 1$:

(2.4) H^R wird verworfen, falls $p \leq \alpha$ ist.

Bei dieser Festlegung der Testentscheidung gilt qua Konstruktion:

(2.5) $P(H^R \,\&\, u\colon p(y) \leq \alpha) \leq \alpha$.

Gleichheit in der Formel (2.5) gilt, wenn α ein erreichbares Fehlerniveau ist. Wegen des diskreten Charakters der Verteilungsfunktion der Teststatistik in Abhängigkeit von M kann α nur exakt erreicht werden, wenn es ein ganzzahliges Vielfaches von $1/M$ ist.

Die Formel (2.5) ist gültig ohne jede Verteilungsannahme an u. Wegen dieser Abhängigkeit von den Einheitenwerten ist aber prinzipiell keine Angabe von kritischen Werten der Teststatistik für ausgewählte Irrtumswahrscheinlichkeiten möglich, wie man sie etwa für normal-, χ^2-, t- oder F-verteilte Teststatistiken findet.

Im Gegensatz zu Randomisierungstests geht man bei *Permutationstests* üblicherweise davon aus, dass die Beobachtungen eine Zufallsstichprobe aus einer Grundgesamtheit darstellen. Ein Zufallsexperiment mit n Be-

obachtungen wird beschrieben durch eine n-dimensionale (stetige) Zufallsvariable (ZV) $Y = (Y_1, \ldots, Y_n)'$, die einer n-dimensionalen Verteilung $F(y)$ folgen möge (vgl. Willmes, 1996). Ergebnis eines konkreten Zufallsexperiments ist dann eine Realisation $y = (y_1, \ldots, y_n)'$ im Stichprobenraum $\boldsymbol{R}^n$. Mit einem Permutationstest werden Hypothesen H^P getestet, die sich darauf beziehen, dass die Verteilung von $F(y)$ bei Gültigkeit der H^P invariant (unverändert) bleibt unter bestimmten Mengen (genauer: Gruppen) von Vertauschungen (Permutationen) der Komponenten von $F(y)$, das heißt, $F(y)$ besitzt bestimmte Symmetrieeigenschaften. Die größtmögliche Menge von Vertauschungen ist die »symmetrische Gruppe« S_n aller $n!$ möglichen Vertauschungen der Komponenten von y. Die Nullhypothesen für Permutationstests beziehen sich im Allgemeinen auf die Invarianz von $F(y)$ unter bestimmten Teilmengen (Untergruppen) Q_n von S_n vom Umfang $M \leq n!$ (Scheffé, 1959).

Zur Testentscheidung wählt man eine »geeignete« Teststatistik $S(Y)$. Sei $S(Y)$ so gebildet, dass große Werte gegen die spezielle H^P sprechen. Unter H^P besitzen alle Werte $s_1, \ldots, s_m$ (mit $m \leq M$), die $S(y)$ für die Permutationen aus Q_n annehmen kann, dieselbe Wahrscheinlichkeit $1/M$. Für eine Testentscheidung kann man nun alle Werte der Teststatistik der Größe nach ordnen und erhält eine diskrete Verteilungsfunktion für S. Um zu einer Testentscheidung zu kommen, ist es weniger aufwändig, zu dem beobachteten y den so genannten p-Wert zu bestimmen:

$$(2.6)\ p := p(y) = P(H^P: S(Y) \geq s(y)) \text{ mit } 1/M \leq p \leq 1.$$

Im Sinne der Testtheorie von Neyman und Pearson (vgl. Willmes, 1996) werden statistische Tests durch Entscheidungsfunktionen $\varphi(y)$ dargestellt, die den empirischen Daten y Entscheidungen gegen H^P ($\varphi(y) = 1$) beziehungsweise für H^P ($\varphi(y) = 0$) zuordnen. In diesem Fall lautet sie:

$$(2.7)\ \varphi(y) = 1 \text{ falls } p^{>} + p^{=} \leq \alpha$$
$$= 0 \text{ sonst,}$$

mit

$$(2.8)\ p^{>} := 1/M \text{ card}\{\pi \in Q_n: S(\pi y) > s(y)\}$$
$$p^{=} := 1/M \text{ card}\{\pi \in Q_n: S(\pi y) = s(y)\}$$

und $\pi \in Q_n$ bezeichne die unter H^P zulässigen Permutationen der Beobachtungen.

Zur praktischen Durchführung des Permutationstests müssen nicht alle Teststatistikwerte geordnet, sondern nur abgezählt werden, wie viele im Sinne von H^P extremere Ausprägungen der Teststatistik es unter den permutierten Beobachtungen gibt.

Per Konstruktion sind Permutationstests bedingte Tests, das heißt, die Testfunktion ist abhängig von der jeweiligen Realisation von Y. Eine Vertafelung von Schwellenwerten ist also nicht möglich. Auch bei schnellen Rechnern kann M für nicht zu kleine Stichprobenumfänge so groß werden, dass sehr lange Rechenzeiten resultieren. Dann wählt man die so genannte »Monte-Carlo« Version des Permutationstests, in der nur eine große (Pseudo-)Zufallsstichprobe aller unter H^P zulässigen Permutationen gezogen wird, um mit diesen den MC p-Wert zu bestimmen, der ein konsistenter Schätzer des p-Werts unter allen $\pi \in Q_n$ ist (Jöckel, 1986). Ist eine Teststatistik sinnvoll, bei der kleine Werte am stärksten gegen H^P sprechen, werden analog zu (2.8) $p^<$ und $p^=$ bestimmt. Bei zweiseitigen Tests verwendet man:

$$(2.9)\quad p = 2 \cdot \min\{p^> + p^=, p^< + p^=, 0.5\}.$$

Dieses allgemeine Konstruktionsprinzip von Permutations- und Randomisierungstests wird nun an Beispielen verdeutlicht. In den betrachteten Fällen führen Randomisierungs- und Permutationstest trotz der konzeptuellen Unterschiede formal zu demselben Test (vgl. Willmes, 1987).

Tests für lineare Hypothesen in »verteilungsfreien« linearen Modellen

Ein Problem der Inferenzstatistik ist es, verteilungsfreie Tests für statistische Hypothesen bezüglich Haupt- und Wechselwirkungsparametern in mehrfaktoriellen Plänen zu konstruieren, denn oft sind empirische psychologische Hypothesen nur adäquat in solchen mehrfaktoriellen Versuchsplänen, unter Umständen mit wiederholten Messungen, zu prüfen. Solche Versuchspläne beinhalten neben der reinen Versuchsplan-Anlage, bezüglich welcher Faktoren eine randomisierte Zuteilung der Beobachtungseinheiten zu Faktorstufen oder bezüglich welcher Messwiederholungsfaktoren welche randomisierte Zuteilung der Untersuchungszeiten zu Faktorstufen erfolgt (Hager, 1987). Auch ohne die übliche Normalverteilung der ANOVA (vgl. Kirk, 1995) sollen weiterhin lineare Modelle mit

linearen Hypothesen von Interesse sein, denn für sie ist eine Definition von Haupt- und Wechselwirkungs-Parametern sinnvoll möglich (siehe Brunner et al., 2002, zu nichtparametrischen Hypothesen in faktoriellen Plänen). Die Abschwächung der Verteilungsannahmen führt trotz Beibehalten der Struktur der Modellparameter zu Problemen, denn bereits in einem zwei-faktoriellen Zufallsplan sind dann Tests auf Haupteffekte nicht mehr unabhängig von dem (Nicht-)Vorliegen von Wechselwirkungen. So sind für mehrfaktorielle »Rangvarianzanalysen« verschiedentlich Tests vorgeschlagen worden, die diese Unabhängigkeit nicht besitzen (Kritik bei Erdfelder u. Bredenkamp, 1984).

Pyhel (1980) hat vorgeschlagen, den allgemeinen Ansatz der Reduktion eines Testproblems durch Invarianz (vgl. Welch, 1990; Willmes, 1996, S. 199), beim Testen einer Hypothese bezüglich eines bestimmten Effekts »störende« Effekte durch spezielle Transformationen zu »entfernen«, auch für Permutationstests in mehrfaktoriellen linearen Modellen zu nutzen (vgl. Brunner u. Munzel, 2003, S. 201). Ausgangspunkt ist das lineare Modell in Matrixschreibweise (vgl. z. B. Werner, 1997):

(3.1) $Y = Xb + \varepsilon$,

mit

Y n-dim. ZV (Spaltenvektor der AV),

X $(n \times k)$-dim. Designmatrix von (vollem) Rang k mit den üblichen Reparametrisierungsbedingungen für den k-dim. Parametervektor b, $k < n$.

Die n-dimensionale Verteilungsfunktion von Y sei durch den Lageparameter $\theta = Xb$ charakterisiert:

(3.2) $F_\theta(y) = F(y - \theta)$,
mit θ Element eines linearen k-dimensionalen Unterraums U des R^n.

Im Unterschied zum ANOVA-Modell mit Normalverteilungsannahmen wird abweichend nur postuliert, dass für F, die Verteilung des Fehlervektors ε, bestimmte Symmetrieeigenschaften, das heißt Invarianzen gegenüber bestimmten Permutationen seiner Argumente gelten:

(3.3) $F(y) = F(\pi y)$ für alle $\pi \in Q^*_n$, Q^*_n Untergruppe von S_n, $y \in R^n$.

Das lineare Modell in (3.1) mit der »schwachen« Invarianzforderung (3.3) an die Fehlerverteilung wird als ein *verteilungsfreies lineares Modell* bezeichnet.

Lineare Hypothesen in dem linearen Modell lassen sich als

(3.4) H: $\theta \in U^H$ gegen A: $\theta \in U \setminus U^H$

formulieren mit $\theta \in R^n$ Element eines linearen Unterraums U^H von U der Dimension $h \leq k$.

Das Testproblem (3.4) bleibt invariant unter allen Transformationen g des Stichprobenraums g: $R^n \rightarrow R^n$ der Form:

(3.5) $g = id + \theta^*$, mit $\theta^* \in U^H$ und id identische Abbildung.

Diese Transformationen g bilden eine Gruppe G. Da U^H ein linearer Teilraum ist, gilt für die transformierten Beobachtungen Y auch: $g(Y) = \theta + \varepsilon + \theta^*$ und $\theta + \theta^* \in U^H$. Sei

(3.6) f^H: $R^n \rightarrow R^n$ orthogonale Projektion des R^n auf den linearen Unterraum U^H und;
f^{*H} die orthogonale Projektion auf den zu U^H orthogonalen Unterraum U^K des R^n,

mit

(3.7) $f^{*H} = id - f^H$, das heißt, für alle $\theta \in U^H$ gilt: $f^{*H}(\theta) = 0 \leftrightarrow \theta \in U^H$.

Die Projektion f^{*H} lässt sich durch eine (idempotente, symmetrische) $(n \times n)$-Matrix K darstellen (vgl. Brunner u. Munzel, 2003, Anhang B; Werner, 1997, Kap. 3.20), sodass gilt:

(3.8) H: $K\theta = 0$ gegen A: $K\theta \neq 0$.

Die orthogonale Projektion f^{*H} bildet zwei Elemente des R^n genau dann auf denselben Wert ab, wenn diese sich höchstens (maximal) um ein Element aus U^H unterscheiden. In U^H sind aber gerade solche θ-Werte enthalten, von denen das Testproblem (3.4) nicht abhängen soll. Es hängt also nur über die bezüglich G maximal invariante Statistik $f^{*H}(Y)$ von Y ab und ist »durch Invarianz reduziert« worden. Bei Wahl einer G-invarianten Statistik T: $R^n \rightarrow R^n$, die die n-dimensionale ZV Y in die n-dimensionale ZV $T(Y)$ überführt, gilt unter der Hypothese H: $T(Y) = T(\varepsilon)$.

Solch eine geeignete Statistik T ist die zuvor erwähnte orthogonale Projektion in den zu U^H orthogonalen Raum. Gesucht ist nun eine Permutationsgruppe Q_n, deren bei Gültigkeit von H induzierte Verteilung F^T von T permutationsinvariant ist für alle $\pi \in Q_n$, das heißt $(F^T)\pi = F^T$. Eine hinreichende Bedingung dafür ist, dass zu $\pi \in Q_n$ ein $\pi^* \in Q^*_n$ existiert mit:

$(F^T)\pi = (F\pi^*)^T = F^T$, wegen der Q^*_n-Invarianz von F.

Bei der Wahl einer orthogonalen Projektion als Statistik T gilt sogar

(3.9) $\pi T = T\pi$,

mit *derselben* Permutation π (Pyhel, 1978).

Entscheidend für die Existenz eines Invarianz-Permutationstests ist also, dass auch der so genannte Bildbereich der Statistik T eine geeignete, durch Q_n charakterisierte Permutations-Invarianzstruktur aufweist.

Letzter Schritt ist dann die Wahl einer geeigneten Teststatistik $S(T)$. Nimmt man dazu die Zähler der üblichen ANOVA F-Statistiken, gilt sogar $S(T) = S(Y)$. Man benötigt die Transformation $T(Y)$ also »nur«, um festzustellen, ob in Q_n »genügend« Permutationen für einen nicht-trivialen Permutationstest enthalten sind. Das allgemeine Konstruktionsprinzip soll nachfolgend an einem positiven Beispiel und einem Beispiel, für das kein echter Invarianz-Permutationstest existiert, erläutert werden:

Beispiel 1: *Invarianz-Permutationstest auf Wechselwirkungen in einem zwei-faktoriellen Plan mit wiederholten Messungen für einen Faktor (Split-Plot Plan; Willmes u. Pyhel, 1981)*

In diesem Versuchsplan (vgl. Hager, 1987, Kap. 3.8.9.3; Werner 1997, Kap. 7.3.3) werden $n = \Sigma\, n(i)$, $i = 1, \ldots, r$, Beobachtungseinheiten den r Stufen eines (Plot-)Faktors A zufällig zugewiesen und jede Beobachtungseinheit wird unter einer Zufallsabfolge der c Stufen des Messwiederholungsfaktors B untersucht. Es ist auch (seltener in der psychologischen Forschung) möglich, dass statt einer Beobachtungseinheit so genannte Blöcke aus ähnlichen Einheiten vorliegen, von denen jede per Zufall e0iner Stufe des Faktors B zugewiesen wird. Häufiger werden die Stufen des Faktors A auch von einer organismischen (Gruppen-)Variablen gebildet, sodass bezüglich A keine Zufallszuteilung erfolgen kann. Wenn Faktor B verschiedene Zeitpunkte darstellt, ist auch keine Zufallszuweisung möglich; man testet dann die Parallelität der Profile der abhängigen Variablen über die Zeit unter allen Stufen von A.

Im ersten Fall hat man als Modellgleichung:

(3.10) $Y_{ih(i)j} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \pi_{h(i)} + \varepsilon_{ih(i)j}$;
mit $i = 1, \ldots, r$; $j = 1, \ldots, c$; $h(i) = 1, \ldots, n(i)$; $N = n_. c$

mit den üblichen Reparametrisierungsbedingungen für eine ANOVA mit festen Effekten (vgl. Werner, 1997), damit die Design-Matrix X den vollen Spaltenrang hat. Im Unterschied zur ANOVA wird nun für den Fehlervektor $\varepsilon = (\varepsilon_{111}, \ldots, \varepsilon_{rn(r)c})'$ nur angenommen, dass er nach $F(x)$ verteilt sei mit $x \in R^N$ und der Permutationsinvarianz-Annahme:

(3.11) $F(\pi x) = F(x)$ für alle $\pi \in Q^*_N = S_c \sim S_{n_.}$;

$S_c \sim S_{n_.}$ heißt *Kranzprodukt* von S_c und $S_{n_.}$ und bezeichnet die von Beobachtungseinheit zu Beobachtungseinheit unabhängigen $c!$ Permutationen der c Stufen des Faktors B, wobei die $n_.$ Fehlervektoren $\varepsilon_{ih(i)} = (\varepsilon_{ih(i)1}, \ldots, \varepsilon_{ih(i)c})'$ selbst wieder als ganze beliebig vertauscht werden dürfen. $S_c \sim S_{n_.}$ enthält $n_.(c!)^{n_.}$ zulässige Permutationen. Diese Annahme (3.11) impliziert Gleichheit der Varianzen und Kovarianzen der Elemente von $\varepsilon_{ih(i)}$ und die Forderung, dass zwischen den $\varepsilon_{ih(i)}$ keine Variabilitätsunterschiede bestehen, da sonst die Vertauschbarkeits-Annahme über alle Stufen von A hinweg verletzt wäre. Für die Profilparallelitätshypothese kann nur die kleinere Untergruppe an $n_.!$ zulässigen Permutationen der gesamten c-dimensionalen Fehlervektoren vorausgesetzt werden. Getestet werden soll die Wechselwirkungs-Nullhypothese:

(3.12) H: $(\alpha\beta)_{11} = \ldots = (\alpha\beta)_{rc} = 0$ gegen A: $(\alpha\beta)_{ij} \neq (\alpha\beta)_{i'j'}$ für ein $(i,j) \neq (i',j')$.

Man hat damit als linearen Unterraum (ein Punkt statt Index zeige die Mittelwertbildung über diesen Index an):

$$U = \{\theta \in R^N: \theta_{ih(i)j} = \mu + \alpha_i + \beta_j + \pi_{h(i)} + (\alpha\beta)_{ij};\ i = 1, \ldots, r;\ j = 1, \ldots, c;\ h(i) = 1, \ldots, n(i)\} \text{ und}$$
$$U^H = \{\theta \in U: \theta_{ih(i)j} = \mu + \alpha_i + \beta_j + \pi_{h(i)};\ i = 1, \ldots, r;\ j = 1, \ldots, c;\ h(i) = 1, \ldots, n(i)\}.$$

Es gilt $\theta \in U^H \leftrightarrow \theta_{ih(i)j} - \theta_{ih(i).} - \theta_{.j} + \theta_{\ldots} = 0$.

In Matrixschreibweise kann man obige Bedingung mit der $(n \times n)$-Projektionsmatrix

(3.13) $K = (I_{n.}1/nJ_{n.}) \otimes (I_c^{-}1/cJ_c)$ und $K\theta = 0$ unter H

darstellen, mit I_m der $(m \times m)$-Einheitsmatrix, J_m der ganz aus $(m \times m)$ Einsen bestehenden Matrix und »⊗« dem *Kroneckerprodukt* zweier Matrizen (vgl. Werner, 1997, S. 35). Anwendung von K auf Y als Transformation $T(Y)$ ergibt für die Elemente von KY:

(3.14) $(KY)_{ih(i)j} = T_{ih(i)j} = Y_{ih(i)j} - Y_{ih(i).} - Y_{..j} + Y_{...} = (\alpha\beta)_{ij} + K\varepsilon$
(= $K\varepsilon$ unter der Wechselwirkungs-Nullhypothese).

In (3.14) sind durch die Transformation die Einflüsse der beiden Haupteffekte α_i und β_j sowie des Blockeffekts $\pi_{h(i)}$ eliminiert worden, aber auch gleichzeitig Abhängigkeiten zwischen den Fehlertermen erzeugt worden, die die Vertauschbarkeitsannahmen bezüglich ε in (3.11) einschränken. Nach den von Pyhel (1978, S. 11) angegebenen Rechenregeln kann man feststellen, dass die Untergruppe

(3.15) $Q_N = S_{n.} \otimes S_c$ von $S_c \sim S_{n.}$ die Bedingung $\pi T = T\pi$ erfüllt.

Q_N enthält nun die gegenüber Q^*_N eingeschränktere Menge von Vertauschungen aller $n.!$ c-dimensionalen Vektoren $\varepsilon_{ih(i)}$ untereinander und der c Stufen des Split-Faktors B simultan über alle $n.$ Vektoren $\varepsilon_{ih(i)}$ hinweg, die nach der »Eliminierung« (so genanntes »Alignment«) der Haupteffekt- und Blockeffekt-Parameter noch als zulässige Permutationen von Komponenten von $K\varepsilon$ (und von KY unter der Wechselwirkungs-Nullhypothese) übrig bleiben. Als Teststatistik kann man den Zähler der üblichen ANOVA F-Statistik wählen, der sich bei Anwendung von $T(Y)$ ergibt:

(3.16) $S(T(Y)) = S(KY) = \Sigma n(i)\Sigma(Y_{i.j} - Y_{i..} - Y_{..j} + Y_{...})^2$.

Der Nenner der ANOVA F-Statistik ist (generell) invariant unter den zulässigen Permutationen und kann deshalb unberücksichtigt bleiben. Da $Y_{..j}$ und $Y_{...}$ invariant bis auf Umindizierungen der Splitfaktor-Stufen unter den Permutationen aus Q_N sind, genügt auch:

(3.17) $S(Y) = \Sigma n(i)\Sigma(Y_{i.j} - Y_{i..})^2$.

Zur Ausführung des Tests braucht man wie in (2.7) nur folgende relativen Häufigkeiten zu bestimmen:

(3.18) $p^{>} = \text{card}\{\pi \in Q_N: S(\pi y) > S(y)\} / \text{card}\{Q_N\}$ und
$p^{=} = \text{card}\{\pi \in Q_N: S(\pi y) = S(y)\} / \text{card}\{Q_N\}$.

Die Wechselwirkungs-Nullhypothese (3.12) wird verworfen, falls gilt:

(3.19) $p = p^{>} + p^{=} \leq \alpha$.

Im Fall des Tests auf Profilparallelität ist $Q_N = S_n \otimes \{i_c\}$ mit i_c der identischen Permutation für c Komponenten eines Vektors und derselben Teststatistik wie oben.

Auch für den zwei-faktoriellen Blockplan kann man entsprechend mit dem Zähler der ANOVA F-Statistik einen Invarianz-Permutationstest auf Wechselwirkungen konstruieren (vgl. Willmes, 1987).

Beispiel 2: *Kein Invarianz-Permutationstest auf Wechselwirkungen im zwei-faktoriellen Zufallsplan*

Im zwei-faktoriellen Zufallsplan werden – bei gleichen Zellbesetzungen n – die insgesamt $N = rcn$ Beobachtungseinheiten den $r \cdot c$ Faktorstufenkombinationen der beiden experimentellen Faktoren zufällig zugewiesen. Das Modell lautet:

(3.20) $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ mit $i = 1, \ldots, r; j = 1, \ldots, c; k = 1, \ldots, n; N = rcn$,

mit den üblichen Reparametrisierungs-Bedingungen der ANOVA und der S_N-invarianten N-dimensionalen Verteilung aller N Fehlerkomponenten.

Getestet werden soll die Wechselwirkungs-Nullhypothese:

(3.21) $H: (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{rc} = 0$ gegen $A: (\alpha\beta)_{ij} \neq (\alpha\beta)_{i'j'}$ für ein $(i,j) \neq (i',j')$.

Man hat damit als linearen Unterraum:

$$U = \{\theta \in R^N: \theta = \eta \otimes 1_n \text{ und } \eta \in R^{rc} \text{ mit } \eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}; i = 1, \ldots, r; j = 1, \ldots, c\}$$

und

$$U^H = \{\theta \in U: \eta_{ij} = \mu + \alpha_i + \beta_j, i = 1, \ldots, r; j = 1, \ldots, c\},$$

mit 1_n dem nur aus Einsen bestehenden n-dimensionalen Spaltenvektor. Es gilt:

$$\theta \in U^H \leftrightarrow (\eta_{ij} - \eta_{i.} - \eta_{.j} + \eta_{..}) \otimes 1_n = 0 \text{ für alle } i,j.$$

In Matrixschreibweise lautet diese Bedingung mit der $(N \times N)$-Projektionsmatrix

$$(3.22)\quad K = I_{rcn} - (1/cI_r \otimes J_c + 1/rJ_r \otimes I_c - 1/rcJ_r \otimes J_c) \otimes 1/nJ_n,$$

und es gilt $K\theta = 0$ unter H.

Anwendung von K auf Y als Transformation $T(Y)$ ergibt für die Elemente von KY:

$$(3.23)\quad (KY)_{ijk}: = T_{ijk} = Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...} = (\alpha\beta)_{ij} + K\varepsilon$$
$$(= K\varepsilon \text{ unter der Wechselwirkungs-Nullhypothese (3.21)}).$$

In diesem Fall hat man als Untergruppe der nach Transformation noch zulässigen Permutationen $Q_N = S_n \sim (S_r \otimes S_c)$, welche die Bedingung $\pi T = T\pi$ erfüllen. (3.24)
Q_N mit $(n!)^{rc}r!c!$ Elementen enthält lediglich von Zelle zu Zelle unabhängige Vertauschungen innerhalb der r·c Zellen und Vertauschungen ganzer Zeilen und Spalten des Versuchsplans. Als Teststatistik kann man wieder den Zähler der üblichen ANOVA F-Statistik wählen, der sich bei Anwendung von $T(Y)$ ergibt:

$$(3.25)\quad S(T(Y)) = S(KY) = \Sigma\Sigma\, n(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2$$
$$= \Sigma\Sigma\, nT_{ij.}^2.$$

Man erkennt sofort, dass jetzt $S(Y)$ invariant ist gegenüber den unter H zulässigen Permutationen, denn Vertauschungen innerhalb der Zellen lassen alle $Y_{ij.}$ invariant; und Vertauschungen ganzer Zeilen und Spalten im Versuchsplan bewirken nur Umindizierungen der anderen Größen auf der rechten Seite von (3.25). Damit existiert kein nicht-trivialer Invarianz-Permutationstest für diese lineare Hypothese (vgl. die heuristischen Überlegungen von Edgington, 1995, Kap. 6; Welch, 1990, S. 693; Anderson u. ter Braak, 2003).

Wenn es jedoch, wie bei Hager (2002) dargestellt, darum geht, für diesen Versuchsplan spezifische, (einseitige) geplante (Interaktions-)Kontraste zwischen verschiedenen Zellen des Versuchsplans auszuführen, besteht in

dem dann sinnvollen »Zellmittelwertsmodell« der ANOVA mit $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$ keine Notwendigkeit, Haupteffektparameter zu eliminieren. Als Teststatistik kann man den Zähler der parametrischen t-verteilten linearen Kontrast-Teststatistik wählen und bestimmt gemäß (2.8) den p-Wert unter allen Vertauschungen von Beobachtungen zwischen den »Zellen« des Versuchsplans, die in den linearen Kontrast involviert sind. Im Split-Plot-Plan aus Beispiel 1 sind unter Interaktionskontrast-Nullhypothesen die Vertauschungen ganzer Blöcke zwischen den Plot-Faktorstufen, die in den Kontrast involviert sind, zulässig.

Willmes (1987) hat systematisch für die meisten gängigen ein- bis dreifaktoriellen ANOVA-Modelle analysiert, ob zu einem F-Test mit Normalverteilungsannahme ein analoger Invarianz-Permutationstest mit schwachen Verteilungsannahmen, welche die randomisierte Zuteilung zu Faktorstufen oder die randomisierte Abfolge von Faktorstufen widerspiegeln, angegeben werden kann. In Verallgemeinerung der Ergebnisse aus Beispiel 2 ist festzustellen, dass in einem mehrfaktoriellen Versuchsplan mit mindestens zwei gekreuzten Faktoren ohne Messwiederholungen ein exakter Invarianz-Permutationstest auf die höchste Wechselwirkung oder auf einfache Wechselwirkungen für diese beiden Faktoren nicht existieren kann. Eine Reihe von Vorschlägen und Simulationsstudien zur Angabe von nicht-exakten (approximativen) Permutationstests findet man aber in einer aktuellen Arbeit von Anderson und ter Braak (2003). Insbesondere untersuchen die Autoren den Fall der Eliminierung störender Effekte *ohne* Berücksichtigung der dann auch eingeschränkteren Menge zulässiger Permutationen für die resultierenden Residuen oder den Fall einer uneingeschränkten Vertauschung der Beobachtungen. Sie berichten auf Grund von Simulationsstudien je nach untersuchtem Design über mehr oder weniger starke Veränderungen des tatsächlichen Signifikanzniveaus und über verminderte Teststärke, besonders bei uneingeschränkter Vertauschung der Beobachtungen. Generell bevorzugen die Autoren ein Alignment störender Effekte bei Annahme einer *nicht* eingeschränkten Vertauschbarkeit der Residuen, ein Vorgehen, das bei dem theoretisch zufriedener stellenden hier gewählten Ansatz einer Reduktion des Testproblems durch Invarianz streng genommen nicht zu rechtfertigen ist.

Permutationstests für mehrere ein-faktorielle Pläne sind bereits früh vorgeschlagen worden (z. B. von Pitman, 1937), jedoch ohne einheitliche Theorie. Edgington (1995), der schon länger die Bedeutung von Randomisierungstests für die experimentelle Forschung betont hat, stellt ebenfalls Tests für ein-faktorielle Pläne und für Haupteffekte in mehrfaktoriellen Plänen vor.

Eine Übertragung des Permutationstestprinzips auf den Fall einer multivariaten abhängigen Variablen ist wie der Übergang von der ANOVA zum analogen MANOVA-Problem konzeptuell einfach (Mardia, 1971; vgl. ausführlich Willmes, 1987). Statt Permutationen eindimensionaler Zufallsvariablen werden dann Permutationsgruppen für p-dimensionale *Vektorpermutationen* – das heißt, alle p Dimensionen der abhängigen Variablen je Beobachtungseinheit werden jeweils komplett permutiert – betrachtet.

Ein *multivariates* Experiment mit multivariater Analysemethode sollte vorwiegend dann verwendet werden, wenn mehrere verschiedene abhängige Variable als Indikatoren für ein psychologisches Konstrukt angesehen werden und sich die experimentellen Hypothesen auf *alle* verschiedenen abhängigen Variablen beziehen. Unter dem Aspekt einer möglichst strengen Prüfung von Kausalhypothesen ist es sinnvoll, mehrere unterschiedliche Operationalisierungen der theoretischen Begriffe zu betrachten. Allerdings ist die Überprüfung eines »globalen« Haupt- oder Wechselwirkungseffekts auf alle abhängigen Variablen oft nur der erste Analyseschritt. Nach Feststellung eines signifikanten Ergebnisses bei einem MANOVA-(Permutations)test bleibt man im Allgemeinen nicht stehen, sondern versucht, mit Post-hoc-Testverfahren für die Faktorstufen(kombinationen) und für die p-dimensionalen abhängigen Variablen die Effekte genauer zu »lokalisieren«. Aber auch in einem deduktivistisch motivierten Vorgehen sind multiple Testverfahren für (geplante) Kontraste (Hager, 2002) erforderlich.

Multiple Permutationstests zum multiplen Niveau alpha

Oft befindet man sich bei der Durchführung experimenteller Untersuchungen in der Situation, mehrere statistische Entscheidungen treffen zu müssen. Werden mehrere, abhängige oder unabhängige, statistische Tests ausgeführt, muss die Wahrscheinlichkeit, *mindestens* eine der einzelnen Nullhypothesen fälschlicherweise zu verwerfen, bei einem festen Niveau a fixiert werden (Miller, 1981). In diesem Abschnitt sollen multiple Permutationstests dargestellt werden, mit denen dieses Fehlerniveau besser als mit der stets anwendbaren Bonferroni-Adjustierung ausgeschöpft und eine möglichst hohe Teststärke erzielt wird (vgl. Westfall u. Young, 1993). Folgt man der Forderung nach einer möglichst adäquaten und erschöpfenden Prüfung wissenschaftlicher Hypothesen (z. B. Bredenkamp, 1980, 1984; Hager, 2002), ergeben sich zum Teil gänzlich andere Strategien zur Kontrolle der

Fehlerwahrscheinlichkeiten erster und zweiter Art (Westermann u. Hager, 1986).

Falls in einem psychologischen Experiment eine Menge (Familie) von m Testproblemen H_i gegen A_i i = 1, . . ., m, gegeben ist, handelt es sich um ein *multiples Testproblem*. Unter einem *multiplen Test* (einer multiplen Testprozedur) versteht man einen m-dimensionalen Vektor von Entscheidungsfunktionen $\varphi = (\phi_1, \ldots, \phi_m)$ mit Komponenten ϕ_l wie in (2.7), die jeweils den Wert 0 oder 1 annehmen können. Petrondas und Gabriel (1983) haben gezeigt, dass zum Testen der einzelnen H_l nicht alle bei Gültigkeit der globalen (Durchschnitts-)Hypothese H: = nH_l zulässigen Permutationen zu berücksichtigen sind, sondern genau die Untergruppe Q_l von zulässigen Permutationen, die die Komponenten des gesamten Beobachtungsvektors betreffen, auf die sich die einzelne Teilhypothese bezieht. So sind beispielsweise im einfaktoriellen Zufallsplan (mit k Stufen) bei Gültigkeit der globalen Nullhypothese H: $\alpha_1 = \ldots = \alpha_k = 0$ alle Permutationen aus $Q_n = S_n$ zulässig. Falls die interessierende Hypothesen-Familie alle $l = 1, \ldots, k(k\text{-}1)/2$ Paarvergleiche von Faktorstufen betrifft, wird die zu H_l gehörige Permutationsuntergruppe Q_l gerade von den Vertauschungen der Elemente in y gebildet, die zu dem entsprechenden Paar $l = (i, i')$ von Faktorstufen gehören.

Nun ist für ein multiples Testproblem nicht die Kontrolle des globalen, sondern des *multiplen* α-Niveaus relevant. Für die Einhaltung des globalen α-Niveaus muss bei Gültigkeit der globalen Nullhypothese H, das heißt, *alle* $H_1, \ldots, H_m$ treffen zu, gelten:

$$(4.1)\quad P(\cup_l(\phi_l = 1)) = \alpha.$$

Westfall und Young (1993, S. 10) sprechen von »weak control of the familywise error rate«. Besser ist die starke Kontrolle (»strong control«) des α-Niveaus, das heißt, für *jede beliebige* Teilmenge $L = \{l: H_l \text{ gilt}\}$ von gültigen Nullhypothesen hat man:

$$(4.2)\quad P(\cup_{l \in L}(\phi_l = 1)) = \alpha.$$

Ein multipler Test, der das multiple α-Niveau einhält, hält auch das globale α-Niveau ein. Der einfachste multiple Test zum multiplen α-Niveau ist die Bonferroni-Adjustierung, bei der die einzelnen H_l jeweils zum Niveau α/m getestet werden (d. h. p_l mit α/m vergleichen). Falls für einen Permutationstest ein Niveau α/m – wegen unter H_l zu weniger zulässiger Permutationen – nicht erreichbar ist, wird diese Nullhypothese akzeptiert (vgl. Petrondas u. Gabriel, 1983).

Auch die gleichmäßige Verbesserung der Bonferroni-Adjustierung durch Holm (1979) hält das multiple α-Niveau ein (vgl. z. B. Horn u. Vollandt, 1995). In dieser *sequenziell verwerfenden* Prozedur werden die zu den Tests der einzelnen Hypothesen der Testfamilie bestimmten p-Werte p_l der Größe nach geordnet: $p_{(1)} \leq \ldots \leq p_{(m)}$. Dann prüft man für den kleinsten geordneten p-Wert: $m \cdot p_{(1)} \leq \alpha$? Falls das nicht zutrifft, werden alle Nullhypothesen beibehalten. Falls die Bedingung zutrifft, wird die zugehörige Nullhypothese verworfen und für den zweitkleinsten p-Wert verglichen: $(m-1) \cdot p_{(2)} \leq \alpha$?, und so weiter, bis zum ersten Mal ein Wert größer als α auftritt.

Eine weitere Verbesserung des Verfahrens ist leicht nur möglich, wenn man die *gemeinsame* Verteilung der p_l-Werte und deren Minimum bestimmen kann. Das ist für Hypothesenfamilien bezüglich der k Stufen eines Faktors zum Beispiel in einem verteilungsfreien linearen Modell wegen der Beschränkung der unter der jeweiligen Teil-Nullhypothese zulässigen Permutationen auf die von der Teil-Nullhypothese erfassten Faktorstufen prinzipiell nicht möglich. Aus Platzgründen wird auf das Verfahren von Petrondas und Gabriel (1983) nur verwiesen (vgl. Westfall u. Young, 1993, Kap. 3.1.5).

Vergleichsweise unkompliziert ist die Konstruktion eines multiplen Permutationstests für die (komplette) Hypothesenfamilie bestehend aus (Teil-) Hypothesen bezüglich der 2^p-1 Komponenten der p-dimensionalen abhängigen Variablen in einem MANOVA-Testproblem (Willmes, 1987; vgl. Westfall u. Young, 1993, Kap. 4), von denen besonders die so genannten »Elementarhypothesen« bezüglich der einzelnen p AVn interessant sind.

Sei $Y_N = (Y'_1, \ldots, Y'_N)'$ der Np-dimensionale Vektor der p-dimensionalen Beobachtungen, gegebenenfalls nach (maximal-)invarianter Transformation zur Beseitigung störender Effekte. Sei $Y^l_N = (Y^l_1, \ldots, Y^l_N)'$, $l = 1, \ldots, p$, der Beobachtungsvektor für die l-te abhängige Variable. Sei wieder die Np-dimensionale Verteilungsfunktion $F_N(y)$ von Y_N bei Gültigkeit der globalen Nullhypothese H invariant unter allen Vektorpermutationen aus $Q_N \otimes \{i_p\}$. Bei Gültigkeit der Elementarhypothesen $H_1, \ldots, H_p$ gilt dasselbe für Q_N und jede der Komponenten F^l_N von F_N. Sei $S = (S_1, \ldots, S_p)'$ der Vektor der univariaten Teststatistiken mit unter Annahme der Gültigkeit der individuellen H_l berechneten p-Werten – $p_1, \ldots, p_p$, die dann alle auf [0,1] gleich verteilt sind.

Seien $p_1 = \ldots = p_p$ die aufsteigend geordneten exakten univariaten p-Werte. Zur Vereinfachung der Notation sei angenommen, dass die Abfolge der geordneten p-Werte der Nummerierung der abhängigen Variablen entspricht. Dann ist der *multiple exakte p-Wert* $p^{[i]}$ wie folgt definiert:

$$(4.3)\quad p^{[i]}(y) = \text{card}\{\pi \in Q_N \otimes \{i_p\}: \min(i \leq l \leq p)\{p_l(S_l(\pi y^l_N))\} = p_i\} / \text{card}\{Q_N\}.$$

Es werden somit alle die Permutationen der abhängigen Variablen gezählt, die *simultan* in den Komponenten $i, \ldots, p$ zu univariaten p-Werten kleiner oder gleich dem i-ten geordneten univariaten p-Wert führen würden. Man beginnt sequenziell mit $i = 1$ und bestimmt $p^{[1]}$. Falls $p^{[1]} > \alpha$ ist, wird die globale Nullhypothese H akzeptiert; ansonsten wird die Nullhypothese H_1 verworfen und man geht zu $p^{[2]}$ über. Falls $p^{[2]} \leq \alpha$ ist, werden H_1, H_2 und $H_1 \cup H_2$ verworfen; ansonsten werden die Nullhypothesen $H_2, \ldots, H_p$ und alle aus ihnen bildbaren Teilmengen akzeptiert. Das Verfahren wird solange fortgesetzt, bis zum ersten Mal ein multipler p-Wert größer als α auftritt.

Beispiel 3: *Multiple, sequenziell verwerfende Testprozedur für die p Komponenten im verbundenen Zweistichproben-Problem*

Es wurden die Leistungen von fünf Aphasie-Patienten vor und nach einer Phase lateralisierter perzeptueller Stimulation der rechten Hirnhemisphäre in drei Untertests des Leistungsprüfsystems (3, 9, 10) und des Berufseignungstests (1, 2, 5) erhoben. Von Interesse war die einseitige (Alternativ-) Hypothese, dass nach dem Training die Leistungen für die 6-dimensionale abhängige Variable verbessert sind. Als Teststatistik für die univariaten, einseitigen verbundenen Zweistichproben-Tests eignet sich die Differenz der Mittelwerte aus Nach- und Vortest beziehungsweise äquivalent (und weniger rechenaufwändig) die Summe der Testleistungen im Nachtest (vgl. Edgington, 1995, S. 109):

(4.4) $S_l = \Sigma_j Y_{2jl}$ mit $l = 1, \ldots, p, j = 1, \ldots, n$.

Tabelle 1: Kontrolltest-Leistungen von fünf aphasischen Patienten in einer Therapie-Pilotstudie (LPS: Leistungsprüfsystem, BET: Berufseignungstest) und Wert für jede der sechs univariaten Teststatistiken (Summe der Punktwerte aus dem Nachtest)

	Vortest						Nachtest					
	LPS			BET			LPS			BET		
UT	3	9	10	1	2	5	3	9	10	1	2	5
Patient	$l=1$	2	3	4	5	6	$l=1$	2	3	4	5	6
1	14	5	9	15	8	6	19	9	14	16	5	6
2	12	6	10	9	6	13	15	11	13	14	7	7
3	20	7	15	28	13	17	23	10	19	34	12	22
4	18	13	12	26	14	21	27	24	15	20	19	20
5	13	9	7	17	13	8	20	9	12	15	14	8
Werte der Teststatistik aus (4.4)							S_l	104	63	73	99	57	63

Tabelle 2: Links: univariate Werte der Teststatistik aus (4.4) für die Daten in Tabelle 1; rechts: geordnete Werte der univariaten Teststatistiken mit univariaten exakten p-Werten

Permutations-Index	Permutations-Indikator*					Werte der univariaten Teststatistiken						Lauf-Index	Geordnete Werte der univariaten Teststatistiken						
	1	2	3	4	5	S_1	S_2	S_3	S_4	S_5	S_6		1	2	3	4	5	6	
1	0	0	0	0	0	104	63	73	99	57	63	1	77	40	53	87	50	58	
2	1	0	0	0	0	99	59	68	98	60	63	2	80	40	56	88	51	58	
3	0	1	0	0	0	101	58	70	94	56	69	3	80	43	56	89	51	58	
4	0	0	1	0	0	101	60	69	93	58	58	4	82	43	57	90	51	58	
5	0	0	0	1	0	95	52	70	105	52	64	5	83	44	58	92	52	59	
6	0	0	0	0	1	97	63	68	101	56	63	6	84	45	58	93	52	59	
7	1	1	0	0	0	96	54	65	93	59	69	7	85	47	59	93	52	59	
8	1	0	1	0	0	96	56	64	92	61	58	8	85	47	60	93	53	59	
9	1	0	0	1	0	90	48	65	104	55	64								$P6$
10	1	0	0	0	1	92	59	63	100	59	63	9	86	48	60	94	53	63	
11	0	1	1	0	0	98	55	66	88	57	64	10	87	48	61	94	54	63	
12	0	1	0	1	0	92	47	67	100	51	70	11	87	49	61	94	54	63	
13	0	1	0	0	1	94	58	65	96	55	69	12	88	49	61	95	54	63	

Permutations-Index	Permutations-Indikator*					Werte der univariaten Teststatistiken						Lauf-Index	Geordnete Werte der univariaten Teststatistiken						
	1	2	3	4	5	S_1	S_2	S_3	S_4	S_5	S_6		1	2	3	4	5	6	
14	0	0	1	1	0	92	49	66	99	53	59	13	89	51	61	95	55	64	
15	0	0	1	0	1	94	60	64	95	57	58	14	89	51	62	95	55	64	
16	0	0	0	1	1	88	52	65	107	51	64	15	89	52	62	96	55	64	
17	1	1	1	0	0	93	51	61	87	60	64	16	90	52	63	96	55	64	
18	1	1	0	1	0	87	43	62	99	54	70	17	91	54	63	98	56	64	
19	1	1	0	0	1	89	54	60	95	58	69	18	92	54	64	98	56	64	
20	0	1	1	1	0	89	54	63	94	52	65								p_4
21	0	1	1	0	1	91	55	61	90	56	64	19	92	54	64	99	56	64	
22	0	0	1	1	1	85	49	61	101	52	59	20	92	55	65	99	56	64	
23	0	1	0	1	1	85	47	62	102	50	70								p_5
24	1	0	1	1	0	87	55	61	98	56	59	21	93	55	65	99	57	65	
25	1	0	0	1	1	83	48	60	106	54	65	22	94	55	65	100	57	65	
26	1	0	1	0	1	89	56	59	94	60	58	23	94	56	65	100	57	65	
27	1	1	1	1	0	84	40	58	93	55	65	24	95	56	66	100	58	65	

Permutations-Index	Permutations-Indikator*					Werte der univariaten Teststatistiken						Lauf-Index	Geordnete Werte der univariaten Teststatistiken						
	1	2	3	4	5	S_1	S_2	S_3	S_4	S_5	S_6		1	2	3	4	5	6	
28	1	1	1	0	1	85	51	56	89	59	64	25	96	58	66	101	58	69	
29	1	1	0	1	1	80	43	57	101	53	70	26	96	58	67	101	59	69	
30	1	0	1	1	1	80	45	56	100	55	59	27	97	59	68	101	59	69	
31	0	1	1	1	1	82	44	58	96	51	65	28	98	59	68	102	59	69	
32	1	1	1	1	1	77	40	53	95	54	65	29	99	60	69	104	60	70	
											30	1	1	60	70	105	60	70	
																			p_2
											31	1	1	63	70	106	60	70	
																			p_1, p_3
											32	1	4	63	73	107	60	70	

*Angabe, ob für Patient $j = 1, \ldots, 5$ die 6-dim. AV aus Vor- bzw. Nachtest vertauscht ist (1) oder nicht (0).

Seien die p-dimensionalen Differenzvektoren $Y_j = Y_{1j} - Y_{2j}$ verteilt nach p-dimensionalen symmetrischem $F_j(x-\theta)$. Unter der (globalen) Nullhypothese H: $\theta = 0$ gegen A: $\theta \neq 0$, ist die Verteilung des gesamten Beobachtungsvektors $Y = (Y'_{11}, Y'_{21}, \ldots, Y'_{1n}, Y'_{2n})'$ invariant unter den 2^n Vertauschungen der p-dimensionalen Beobachtungsvektoren je Beobachtungseinheit zwischen Zeitpunkt 1 und Zeitpunkt 2.

In Tabelle 2 sind für alle 2^5 zulässigen Vektor-Permutationen links die Werte der univariaten Teststatistiken und rechts die geordneten Teststatistikwerte angegeben. Die nach Größe geordneten p-Werte der 6 univariaten Tests liest man leicht im rechten Tabellenteil ab.

(4.5) $p_{(1)} = p_1 = 1/32$ $\quad p_{(2)} = p_3 = 1/32$ $\quad p_{(3)} = p_2 = 2/32$
$p_{(4)} = p_5 = 12/32$ $\quad p_{(5)} = p_4 = 14/32$ $\quad p_{(6)} = p_6 = 24/32.$

Die Abfolge der identischen p-Werte p_1 und p_3 wurde per Zufall bestimmt. Anhand der Tabelle 2 kann man leicht die multiplen p-Werte (4.3) bestimmen.

Zur Ermittlung von $p^{[1]}$ ist die Menge aller Permutationen zu bestimmen, die für irgendeine der univariaten Teststatistiken zu einem univariaten p-Wert = $p_{(1)}$ führen. Für die abhängigen Variablen 1 und 3 ist es die Permutation mit dem Laufindex 1, für die abhängigen Variablen 4 und 5 diejenige mit dem Index 16 und 8. Da für die Variablen 2 und 6 der kleinste erreichbare p-Wert 2/32 ist, kommen keine weiteren Permutationen hinzu. Damit hat man drei verschiedene Permutationen und es ergibt sich $p^{[1]} = 3/32$. Dieser Wert ist kleiner als der multiple p-Wert der Holm-Prozedur mit $6 \cdot p_{(1)} = 6/32$; insgesamt:

(4.6) $p^{[1]} = 3/32$ $\quad p^{[2]} = 3/32 < 5 \cdot p_{(2)} = 5/32$ $\quad p^{[3]} = 5/32 < 4 \cdot p_{(3)} = 8/32$
$p^{[4]} = 29/32$ $\quad p^{[5]} = 22/32$ $\quad p^{[6]} = p_{(6)} = 24/32.$

Für ein (liberales) multiples Niveau $\alpha = 5/32$ wären somit nach dem Holm-Verfahren keine, für das exakte schrittweise Verfahren dagegen die Leistungen für alle drei LPS-Untertests als signifikant verbessert anzusehen.

Die neue multiple Testprozedur bedeutet besonders in Fällen hoch korrelierter abhängiger Variablen eine deutliche Erhöhung der Teststärke, da dann für verschiedene Variable ganz überwiegend dieselben Permutationen zu extremen Werten der Teststatistik führen und damit die multiplen p-Werte kaum größer als die univariaten p-Werte ausfallen.

Es können auch einschrittige, so genannte *simultane* (Permutations-) Testprozeduren für die Komponenten einer p-dimensionalen abhängigen

Variablen angegeben werden. Wenn die univariaten Teststatistiken $S_1, \ldots, S_p$ jeweils unter der univariaten Nullhypothese alle dieselbe Verteilung besitzen, kann man – wenn große Teststatistikwerte gegen die Nullhypothese sprechen – als Teststatistik wählen:

(4.7) $S_{max} = \max \{S_1, \ldots, S_p\}$.

Die Permutationsverteilung von S_{max} wird für alle bei Gültigkeit der globalen Nullhypothese H zulässigen Vektor-Permutationen ermitteln. Dann wählt man in der diskreten Verteilungsfunktion von S_{max} den Quantilwert $s_{max,1-\alpha}$, der am nächsten an dem multiplen α-Niveau liegt, ohne es zu überschreiten. Der jeweils beobachtete s_l-Wert für die Komponente l der abhängigen Variablen wird als signifikant angesehen, wenn er nicht kleiner als $s_{max,1-\alpha}$ ist (vgl. Westfall u. Young, 1993).

Abschließende Bemerkungen

Obwohl in diesem Beitrag ganz vorwiegend Tests für lineare Hypothesen in linearen Modellen vorgestellt worden sind, gibt es eine Vielzahl von anderen statistischen Hypothesen, zu deren Testung Permutations- und Randomisierungstests eingesetzt werden können (vgl. Edgington, 1995; Hooper, 1989). Da keine Vertafelungen von kritischen Werten für diese bedingten Tests vorliegen können, kann man sehr flexibel Teststatistiken wählen, auch wenn über deren Optimalitätseigenschaften gegen bestimmte Alternativen nichts bekannt ist.

Seit es schnelle Personal Computer gibt, ist für nicht zu große Stichprobenumfänge der exakte p-Wert berechenbar. Für bestimmte (lineare) Permutationstest-Statistiken ist mit so genannten Shift-Algorithmen (Streitberg u. Röhmel, 1986; Brunner u. Munzel, 2003) die gesamte Verteilung erzeugbar, ohne dass alle zulässigen Permutationen tatsächlich ausgeführt werden müssen. Die (Netzwerk-)Algorithmen, wie sie beispielsweise im umfangreichsten Software-Paket (StatXact; Mehta u. Patel, 1999) für exakte Permutationstests verwendet werden, haben Eingang in bekannte Programmpakete wie SPSS oder SAS gefunden. Für die im Beitrag beschriebenen (und zusätzliche) verteilungsfreie ANOVA- und MANOVA-Testprobleme sind FORTRAN77-Routinen vom Autor erhältlich.

Randomisierungstests sind auch oft für die Einzelfallforschung geeignet (vgl. Edgington, 1995; Software SCRT; Onghena u. Van Damme, 1994), da

die unter den betreffenden Nullhypothesen gemachten Vertauschbarkeitsannahmen eher erfüllbar erscheinen als die Unabhängigkeitsannahmen parametrischer Testverfahren (siehe auch für Anwendungen in der kognitiven Neuropsychologie Deloche u. Willmes, 2002; Willmes, 2000).

Literatur

Anderson, M.J., ter Braak, C.J.F. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation, 73,* 85–113.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1984). Anmerkungen und Korrekturen zu Hager & Westermann. Entscheidung über statistische und wissenschaftliche Hypothesen: Probleme bei mehrfachen Signifikanztests zur Prüfung einer wissenschaftlichen Hypothese. *Zeitschrift für Sozialpsychologie, 15,* 224–229.

Brunner, E., Domhof, S., Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments.* New York: Wiley.

Brunner, E., Munzel U. (2003). *Nicht-parametrische Datenanalyse.* Berlin: Springer-Verlag.

Edgington, E.S. (1995). *Randomization tests* (3. Aufl.). New York: Marcel Dekker.

Erdfelder, E., Bredenkamp, J. (1984). Kritik mehrfaktorieller Rangvarianzanalysen. *Psychologische Beiträge, 26,* 263–282.

Deloche, G., Willmes, K. (2000). Cognitive neuropsychological models of adult calculation and number processing: the role of the surface format of numbers. *European Child & Adolescent Psychiatry, 9*(Suppl 2 II), 27–40.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Hg.), *Allgemeine experimentelle Psychologie* (S. 43–264). Stuttgart: G. Fischer.

Hager, W. (2000). About some misconceptions and the discontent with statistical tests in psychology. *Methods of Psychological Research – Online, 5,* 1–31.

Hager, W. (2002). The examination of psychological hypotheses by planned contrasts referring to two-factor interactions in fixed effects ANOVA. *Methods of Psychological Research – Online, 7,* 49–77.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Hooper, P.M. (1989). Experimental randomization and the validity of normal-theory inference. *Journal of the American Statistical Association, 84,* 576–586.

Horn, M., Vollandt, R. (1995). *Multiple Tests und Auswahlverfahren.* Stuttgart: G. Fischer.

Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *The Annals of Statistics, 14,* 336–347.

Kirk, R.E. (1995). *Experimental design: Procedures for the behavioral sciences* (3. Aufl.). Pacific Grove: Brooks/Cole.

Mardia, K.V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika, 58,* 105–121.

Mehta, C., Patel, N. (1999). *StatXact4 for Windows: User Manual.* Cambridge, MA: Cytel Software Corporation.

Miller, R.G. (1981). *Simultaneous statistical inference.* New York: Springer-Verlag.

Onghena, P., Van Damme, G. (1994). SCRT 1.1: Single-case randomization tests. *Behavior Research Methods, Instruments & Computers, 26,* 369.

Petrondas, D.A., Gabriel, K.R. (1983). Multiple comparisons by rerandomization tests. *Journal of the American Statistical Association, 78,* 949–957.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society, 4*(Suppl.), 119–130.

Pyhel, N. (1978). *Beiträge zur Theorie der Permutationstests mit Anwendung in einem einem linearen Modell.* Unveröffentlichte Dissertation: RWTH-Aachen, Aachen.

Pyhel, N. (1980). Distribution-free r sample tests for the hypothesis of parallelism of response profiles. *Biometrical Journal, 22,* 703–714.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Streitberg, B., Röhmel, J. (1986). Exact distributions for permutation and rank tests: An introduction to some recently published algorithms. *Statistical Software Newsletter, 12,* 10–17.

Welch, W.J. (1990). Construction of permutation tests. *Journal of the American Statistical Association, 85,* 693–698.

Westermann. R., Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11,* 117–146.

Westfall P.H., Young, S.S. (1993). *Resampling-based multiple testing.* New York: Wiley.

Matthias Siemer

Teststärkekonzept und Teststärkebestimmung bei Randomisierungstests

Zusammenfassung

Der vorliegende Beitrag zeigt, wie eine kohärente Definition des Teststärkebegriffs im Rahmen des Randomisierungsmodells lokaler statistischer Inferenz möglich ist. Diese Definition lokaler Teststärke basiert auf der Annahme hypothetischer Replikationen eines durchgeführten Experiments mit derselben lokalen Population. Damit wird gezeigt, dass das im Rahmen der Neyman-Pearson-Testtheorie frequentistisch konzipierte Teststärkekonzept auch im Kontext von Randomisierungstests, die kausal Hypothesen in Hinblick auf eine gegebene lokale Population testen, sinnvoll definierbar ist. Darauf aufbauend wird geschildert, wie diese theoretische Definition im Rahmen praktischer Teststärkebestimmungen in der Versuchsplanung angewendet werden kann. Ein wichtiges Ergebnis bei der Evaluation dieser randomisierungsbasierten Teststärkebestimmung ist, dass bei Vorliegen der entsprechenden Voraussetzungen die nonparametrisch bestimmte Teststärke von Randomisierungstests durch entsprechende parametrische Verfahren approximiert werden kann. Abschließend werden Alternativen bei der Teststärkebestimmung und praktische Empfehlungen diskutiert.

Der Randomisierungstest: Statistische Inferenz in lokalen Populationen

Es ist seit langem bekannt, dass es sich bei der Mehrzahl der in der experimentellen Psychologie untersuchten Versuchspersonenstichproben nicht um Zufallsstichproben aus zugrunde liegenden Populationen handelt. Eine wichtige Implikation dieser Tatsache ist, dass somit eine zentrale Voraussetzung parametrischer Signifikanztests in der Regel nicht erfüllt ist.

Gleichzeitig ist das den parametrischen Signifikanztests zugrunde liegende Inferenzmodell den meisten experimentellen Hypothesen gar nicht angemessen. In experimentellen Untersuchungen ist fast nie eine statistische Generalisierung der gefundenen Resultate auf Parameter in einer klar definierten Population intendiert. Der statistische Signifikanztest dient

vielmehr der statistischen Absicherung der gefundenen kausalen Effekte in Hinblick auf die in der konkreten Untersuchung vorliegenden Untersuchungseinheiten. Die methodische Grundvoraussetzung einer solchen lokalen kausalen Inferenz ist die randomisierte Zuordnung der Untersuchungseinheiten zu den Untersuchungsbedingungen (Bredenkamp, 1972, 1980; Erdfelder u. Bredenkamp, 1994; Hager u. Westermann, 1983).

Im deutschsprachigen Raum hat auf diesen Umstand schon frühzeitig Bredenkamp (1972; 1980) verwiesen. Im internationalen Raum setzt sich die geschilderte Sichtweise nach frühen eher vereinzelten Befürwortern (Edgington, 1995) nunmehr zunehmend durch (Ludbrook u. Dudley, 1998; Lunneborg, 2001; Maxwell u. Delaney, 1990; Reichardt u. Gollob, 1999; Siemer u. Joormann, 2003a, 2003b).

Das paradigmatische statistische Verfahren einer solchen auf Randomisierung basierenden lokalen statistischen Inferenz ist der Randomisierungstest (vgl. Bredenkamp, 1980; Edgington, 1995; Good, 1994; Kempthorne, 1955; Manly, 1997; Westermann u. Siemer, 2001; Willmes, in diesem Band). Die Logik des Randomisierungstests ist schon vergleichsweise lange bekannt und geht auf Arbeiten von Fisher (1935) und Pitman (1937) zurück. Das Prinzip des Randomisierungstests ist sehr einfach und unmittelbar einleuchtend. Der Randomisierungstest testet die Nullhypothese, dass das Treatment keinen kausalen Einfluss auf die abhängige Variable besitzt. Unter dieser Annahme sind die Ausprägungen der abhängigen Variablen für jede Versuchsperson unabhängig von der Treatmentstufe, der die Person zugewiesen wurde. Mit anderen Worten: Es wird für jede Person unter jeder Stufe des Treatments dieselbe Ausprägung auf der abhängigen Variablen erwartet. Unter dieser Annahme ist das vorliegende Ergebnis ausschließlich das Resultat der randomisierten Zuordnung der Untersuchungseinheiten. Die Wahrscheinlichkeit des vorliegenden oder eines extremeren Ergebnisses unter dieser Annahme lässt sich bestimmen, indem man die Daten permutiert oder rerandomisiert und die relative Häufigkeit von Ergebnissen mit der gleichen oder einer extremeren Ausprägung berechnet. Dies kann auf der Basis aller möglichen Permutationen oder Rerandomisierungen geschehen (exakter Randomisierungstest) oder einer zufälligen Auswahl von Permutationen (approximativer Randomisierungstest).

Obwohl die Begriffe »Randomisierungstest« und »Permutationstest« häufig synonym verwendet werden (Good, 1994; Manly, 1997) und rechnerisch zu identischen Ergebnissen führen, ist eine Unterscheidung auf konzeptueller Ebene möglich und sinnvoll (Edgington, 1995; Lunneborg, 2000). Randomisierungstests testen, wie geschildert, kausale Hypothesen über Individuen in lokalen Populationen (Lunneborg, 2001) und sind voll-

ständig durch die Randomisierungsprozedur gerechtfertigt. Permutationstests testen hingegen Hypothesen über Populationsparameter und benötigen Zufallsstichproben, aber nicht notwendigerweise randomisierte Zuordnungen (Edgington, 1995).

Da die Anwendung des Randomisierungstests sehr rechenintensiv ist und, außer bei sehr kleinen Stichproben, ohne leistungsfähige Computer nicht durchführbar ist, gab es lange Zeit keine praktischen Anwendungen des Randomisierungstests. Heutzutage stellt der Rechenaufwand keinen Hinderungsgrund mehr für den Einsatz von Randomisierungstests dar. Sowohl in der Lehre als auch in der Praxis ist jedoch nach wie vor eine überwiegende Dominanz parametrischer Verfahren festzustellen. Obwohl diese Praxis dadurch zu rechtfertigen ist, dass parametrische Verfahren und entsprechende Randomisierungstests, unter Vorliegen der parametrischen Testvoraussetzungen, zu approximativ denselben Resultaten führen (Bredenkamp, 1980; Maxwell u. Delaney, 1990), ist dennoch nach den Gründen für diese Dominanz zu fragen.

Ein erster Grund liegt sicherlich in der historischen Entwicklung, da Randomisierungstests lange Zeit keine praktikable Alternative für parametrische Tests darstellten. Dies lässt sich auch aus der vergleichsweise weiten Verbreitung von Rangsummentests ablesen, die zwar auch auf dem Rerandomisierungsprinzip basieren, aber rechnerisch wesentlich weniger aufwändig sind als etwa Randomisierungstests für Mittelwerte.

Ein weiteres Hindernis für die weitergehende Verbreitung von Randomisierungstests ist, dass sie keine offensichtliche Konzeptualisierung von Teststärke erlauben. Die Bedeutung der Teststärke für statistische Inferenz muss an dieser Stelle nicht betont werden. Auch die Wichtigkeit von Teststärke bei der Planung und Interpretation von Untersuchungen wurde im deutschsprachigen Raum schon früh erkannt, wiederum nicht zuletzt dank der Arbeiten von Bredenkamp (Bredenkamp u. Feger, 1970; Bredenkamp, 1980), und ist mittlerweile auch international erkannt (Wilkinson und The Task Force on Statistical Inference, 1999). Umso wichtiger erscheint es, dass das (zumindest anscheinende) Problem der Teststärke von Randomisierungstests konzeptuell und praktisch gelöst wird.

Teststärke und Randomisierungstests: Zwei Probleme

Das erste *Prima-facie-Problem* bei der Bestimmung der Teststärke (bzw. des β-Fehlers) von Randomisierungstests besteht darin, dass das Konzept der Teststärke über das frequentistische Testmodell von Neyman-Pearson (1928) definiert ist. In diesem Model ist die Teststärke als relative Häufigkeit definiert, mit der Replikationen einer Untersuchung, bei Gültigkeit der Alternativhypothese, zu einem signifikanten Ergebnis führen. Replikationen setzen in diesem Modell das wiederholte Ziehen von Zufallsstichproben aus einer definierten Grundgesamtheit, in der die Alternativhypothese gilt, voraus. Der Begriff der Teststärke scheint somit an das Populationskonzept und das Konzept der Zufallsstichprobe gebunden zu sein. Es ist in diesem Zusammenhang bemerkenswert, dass Fisher (1935), auf den das Prinzip des Randomisierungstests zurückgeht, eine frequentistische Interpretation von Fehlerwahrscheinlichkeiten und das Konzept des β-Fehlers stets ablehnte.

Das zweite Problem besteht darin, dass jegliche Teststärkebestimmung eine genaue Spezifikation der Alternativhypothese voraussetzt. Die Alternativhypothese ist beim Randomisierungstest jedoch weit weniger spezifiziert als bei parametrischen Tests. Da die Nullhypothese lediglich die Gleichheit der Verteilungen der Beobachtungen unter allen Treatmentstufen impliziert, besteht die Menge aller möglichen Alternativhypothesen theoretisch aus allen möglichen Abweichungen von dieser Gleichverteilung. In der Praxis ist die Menge der Alternativhypothesen jedoch durch die Festlegung der Teststatistik (z. B. Mittelwertsdifferenzen) bereits weitgehend eingeschränkt.

Die Definition lokaler Teststärke von Randomisierungstests

Wie beschrieben, muss zur Teststärkebestimmung zunächst die Alternativhypothese spezifiziert werden. Diese Spezifikation kann etwa in der Annahme konstant additiver Treatmenteffekte bestehen. Diese Annahme spiegelt in der Psychologie gängige Hypothesen wieder, die in der Regel homogene Treatmenteffekte annehmen. Die Verletzung der Annahme homogener Treatmenteffekte, zum Beispiel Ausgangswertabhängigkeit von Treatmenteffekten, wird in der Regel als Form von »Aptitude-Treatment-Interaktion« interpretiert und motiviert weitere Analysen und Untersuchungen zur

Identifikation homogener Gruppen (also Gruppen mit konstant additiven Treatmenteffekten).

Unter der Annahme einer spezifizierten Alternativhypothese (hier: konstante additive Treatmenteffekte) kann das Konzept der Teststärke von Randomisierungstests in lokalen Populationen definiert werden. Dieses Konzept lokaler Teststärke basiert auf der Annahme hypothetischer Replikationen eines durchgeführten Experiments mit derselben lokalen Population. Das Konzept geht auf den Ansatz der exakten bedingten Teststärke (*exact conditional power*) von Gabriel und Hsu (1983) zurück (vgl. auch Willmes, 1987). Danach kann die Teststärke eines Randomisierungstests konditional zu den Daten einer durchgeführten Untersuchung bestimmt werden. Die Methode kann mithilfe des Konzepts hypothetischer Replikationen beschrieben werden und erlaubt eine Explikation des Begriffs der lokalen Teststärke.

Eine hypothetische Replikation besteht in der hypothetischen erneuten randomisierten Aufteilung der lokalen Population auf Treatmentbedingungen und der anschließenden Addition des Treatmenteffekts zu der oder den Treatmentgruppen und der Bestimmung der Signifikanz der Replikation mit Hilfe eines Randomisierungstests. Die exakte lokale Teststärke lässt sich definieren als relativer Anteil an hypothetischen Replikationen mit signifikantem Resultat. Im Einzelnen beinhaltet eine hypothetische Replikation folgende Schritte (der Einfachheit halber wird dabei von einem Kontrolle-Treatment-Gruppenplan ausgegangen):

1. Die Werte der Untersuchungseinheiten unter der Treatmentbedingung werden um den (additiven) Treatmenteffekt residualisiert. Das heißt, ein konstanter Treatmenteffekt wird von der Werten aller Untersuchungseinheiten in der Treatmentgruppe subtrahiert.
2. Die Untersuchungseinheiten der Kontrollgruppe und der residualisierten Treatmentgruppe werden erneut randomisiert aufgeteilt.
3. In der neu gebildeten (hypothetischen) Treatmentgruppe wird der (additive) Treatmenteffekt hinzugefügt.
4. Die Signifikanz der hypothetischen Replikation wird mit einem Randomisierungstest bestimmt.
5. Die Schritte 2–4 werden für alle möglichen Permutationen K (oder einer zufälligen Auswahl) durchgeführt.

Auf der Basis des oben skizzierten Vorgehens lässt sich das Konzept der lokalen Teststärke exakt definieren. Allerdings handelt es sich hierbei um eine theoretische Definition, die keine unmittelbare praktische Bestimmung der Teststärke erlaubt.

Zum einen ist das skizzierte Vorgehen rechnerisch sehr aufwändig. Bei K möglichen Permutationen der Ausgangsdaten benötigt das Vorgehen insgesamt K^2 Permutationen (K Permutationen in Schritt 2 und für jede Permutation Durchführung eine Randomisierungstests mit K Permutationen in Schritt 4).

In der obigen Darstellung werden die wahren Treatmenteffekte als bekannt vorausgesetzt. Diese wahren Treatmenteffekte sind notwendig, um in Schritt 1 aus den Ausprägungen der Untersuchungseinheiten aus der Treatmentgruppe potenzielle Ausprägungen zu berechnen, die diese Untersuchungseinheiten gezeigt hätten, wenn sie der Kontrollgruppe zugewiesen worden wären. Weiterhin ist die Kenntnis der wahren Treatmenteffekte notwendig, um in Schritt 3 die Ausprägungen der abhängigen Variablen bei den Untersuchungseinheiten der neu gebildeten hypothetischen Treatmentgruppe zu berechnen.

In der Praxis sind die wahren Treatmenteffekte selbstverständlich unbekannt. Dies ist in der Regel der Grund, warum ein Experiment durchgeführt wird. Damit sind auch die potenziellen Reaktionen unbekannt, die Untersuchungseinheiten in der Treatmentgruppe gezeigt hätten, die faktisch der Kontrollgruppe zugeordnet worden sind und umgekehrt. In der Unkenntnis dieser potenziellen Reaktionen besteht gerade das prinzipielle Problem experimenteller kausaler Inferenz (Holland, 1986).

Schließlich basiert die obige Definition lokaler Teststärke auf einem bereits durchgeführten Experiment, dessen Ergebnis bereits feststeht. Ein bereits durchgeführtes Experiment wird in aller Regel nicht noch einmal mit denselben Versuchspersonen repliziert werden. Teststärkeanalysen werden jedoch in der Regel besonders sinnvoll vor einer Untersuchung durchgeführt, also als a priori Teststärkeanalysen. Darüber hinaus ist die nachträgliche Berechnung der Teststärke eines bereits durchgeführten Experiments auch mit einer Reihe von konzeptionellen und logischen Problemen behaftet (Hoenig u. Heisey, 2001; Lenth, 2001). So ist etwa die Argumentation, dass in einer vorliegenden Untersuchung ein deskriptiv vorhandener Effekt nicht zu einem signifikanten Ergebnis geführt hat, weil die Teststärke zur Aufdeckung des beobachteten Effekts zu niedrig war, logisch defizitär, da eine solche retrospektive Teststärkeanalyse bei insignifikantem Ergebnis notwendigerweise eine niedrige Teststärke ergibt.

Die obige Definition lokaler Teststärke expliziert, wie im Rahmen lokaler statistischer Inferenz Teststärke sinnvoll definiert werden kann. Die Definition der lokalen Teststärke ist dabei ebenso frequentistisch wie die Teststärke der Neyman-Pearson-Konzeption, basiert jedoch auf dem Randomisierungskonzept und dem Konzept der lokalen Population. Die Anwendung

dieses Konzepts im Rahmen der A-priori-Teststärkeplanung bedarf demgegenüber einer Reihe von Modifikationen und zusätzlicher Annahmen.

Praktische Anwendung des Teststärkekonzepts

Die praktische Anwendung des dargestellten Teststärkekonzepts im Rahmen von A-priori-Teststärkeanalysen erfordert zunächst die Erhebung von oder den Zugang zu empirischen Daten. In dieser Hinsicht ist also das dargestellte Vorgehen konditional im Hinblick auf vorhandene empirische Daten. Diese Eigenschaft ist allen nonparametrischen Ansätzen zur Ermittlung oder Schätzung von Teststärke gemeinsam und gilt beispielsweise für Teststärkeschätzungen, die auf dem Bootstrap-Ansatz basieren (z. B. Davison u. Hinkley, 1997; Lunneborg, 2000). Gerade weil nonparametrische Methoden keine Annahmen über Verteilungen beinhalten, müssen diese Verteilungen aus vorhandenen Daten geschätzt werden. Es ist aber in vielen Bereichen damit zu rechnen, dass bereits erhobene Daten verfügbar sind, etwa wenn Teststärke im Rahmen einer Serie von Experimenten geschätzt werden soll.

Die Anwendung des auf dem Randomisierungsprinzip basierenden Teststärkekonzepts erfordert es jedoch, über die lokale Population dieser zur Teststärkeschätzung herangezogenen Untersuchungseinheiten hinauszugehen. Darüber hinaus lehnt sich das vorgeschlagene Vorgehen an die Methode der »naiven« approximativen Teststärkeschätzung von Gabriel und Hsu (1983). Dieses Verfahren wird deswegen als naiv bezeichnet, weil es sich an das Vorgehen parametrischer Tests anlehnt: Zunächst wird eine Verteilung unter Annahme der Gültigkeit der Nullhypothese erstellt (Nullverteilung) und ein empirisch gefundener Wert daraufhin mithilfe dieser Nullverteilung beurteilt. Der Vorteil dieser Methode besteht darin, dass der Rechenaufwand wesentlich geringer ist als der der exakten Methode (K Permutationen). Die modifizierte Methode der naiv approximierten Teststärke ist einfach durchzuführen und besteht aus folgenden Schritten.

1. Zunächst werden die Werte der Untersuchungseinheiten unter der Treatmentbedingung um den (additiven) Treatmenteffekt residualisiert. Dieser additive Treatmenteffekt wird jedoch einfach aus den empirischen Daten geschätzt (also im Falle von Mittelwertshypothesen die deskriptive Mittelwertsdifferenz).
2. Der entstehende Datenvektor wird K-fach permutiert. Die entstehende Verteilung der Teststatistik (z. B. Mittelwertsdifferenzen) ist die empirisch geschätzte randomisierungsbasierte Nullverteilung. Der kritische

Wert dieser Verteilung bei $K(1-\alpha)$ wird ermittelt. Dieser kritische Wert ist derjenige Wert der α % der Nullverteilung abschneidet. Der kritische oder ein noch extremerer Wert hat also unter Annahme der Nullhypothese eine Wahrscheinlichkeit von α und entspricht in der Interpretation somit kritischen Werten aus bekannten Testverteilungen (z. B. Z-Werte, t-Werte).

3. Zu der in Schritt 2 unter Annahme der Nullhypothese berechneten Verteilung der Teststatistik wird im Falle additiver Effekte ein a priori festgelegter Treatmenteffekt in Einheiten der deskriptiven gemeinsamen Standardabweichung der erhobenen Daten hinzugefügt. Die entstehende Verteilung ist die Verteilung unter Annahme der Alternativhypothese.
4. Die naive Approximation der Teststärke für ein bestimmtes Signifikanzniveau α ergibt sich als relativer Anteil derjenigen Werte der in Schritt 3 berechneten Teststatistiken, die größer oder gleich dem in Schritt 2 berechneten kritischen Wert sind.

Das geschilderte Vorgehen zur Teststärkeschätzung enthält sowohl konditionale Elemente, die auf erhobenen Daten basieren, andererseits aber auch unkonditionale Aspekte, die über erhobene Daten hinausgehen.

Zu den *konditionalen* Elementen ist zu zählen, dass die Verteilung der Daten unter Annahme der Nullhypothese (Nullverteilung) aufgrund empirischer Daten geschätzt wird. Folgende Modifikationen des oben explizierten Vorgehens sind dabei möglich: Zum einen ist es möglich, dass der Datenvektor lediglich aus Untersuchungseinheiten einer unbehandelten Kontrollgruppe besteht. Dann entfällt selbstverständlich die Residualisierung einer oder mehrerer Treatmentgruppen (vgl. auch Lunneborg, 2000). Zum anderen kann die geschätzte Nullverteilung aus mehreren (Teil-)Kopien des ursprünglichen Vektors bestehen. Dies ist dann sinnvoll, falls die Teststärke für eine größere Anzahl von Untersuchungseinheiten geschätzt werden soll.

Zu den *unkonditionalen* Elementen gehört die Spezifikation der Effektstärke. Diese muss auf A-priori-Überlegungen basieren. Eine Möglichkeit besteht in der Schätzung der Effektstärke aufgrund einer größeren Anzahl früherer Studien oder einer Metaanalyse. Eine zweite Möglichkeit besteht in der Festlegung der Effektstärke auf eine praktisch bedeutsame Mindesteffektstärke, für die man die Teststärke ermitteln will. Die Spezifikation der Effektstärke enthält jedoch auch ein konditionales Element. In der psychologischen Forschung werden Effektstärken traditionell als standardisierte (an Standardabweichungen relativierte) Werte angegeben, etwa als Cohens d (Cohen, 1988). Das oben geschilderte Verfahren benötigt jedoch absolute Treatmenteffekte und mithin einen Maßstab für die a priori festgelegten

standardisierten Effektstärken. Dieser Maßstab wird in Schritt 3 als deskriptive Standardabweichung in den erhobenen Daten festgelegt, ist also konditional.

An dieser Stelle sei noch einmal betont, dass die für jede Teststärkeanalyse notwendige Spezifikation einer Effektstärke keinesfalls aufgrund erhobener Daten erfolgen sollte. Dieses Vorgehen führt zu extrem unpräzisen (und verzerrten) Schätzungen der wahren Teststärke. Der Hauptgrund besteht darin, dass Effektstärkeschätzungen auf der Basis der in der experimentellen Forschung üblichen Stichprobenzahlen extrem unpräzise sind. Einen ersten Eindruck über die sich ergebende mangelnde Präzision so gewonnener Teststärkeschätzungen kann man sich verschaffen, indem man ein Konfidenzintervall um die empirische Effektstärke berechnet (Cumming u. Finch, 2001; Fidler u. Thompson, 2001; Steiger u. Fouladi, 1997) – ein Vorgehen, das zur Interpretation von Effektstärken ohnehin stets angeraten ist (Wilkinson und The Task Force on Statistical Inference, 1999) – und nachfolgend die sich ergebende Teststärke für das untere und obere Konfidenzlimit errechnet. Trotz dieser Eigenschaft ist diese Form der auf empirischen Effekten beruhenden Post-hoc-Teststärkeschätzung erstaunlich weit verbreitet. In SPSS ist diese Methode unter der Bezeichnung »observed power« implementiert.

Schließlich beinhaltet die praktische Anwendung der permutationsbasierten Teststärkeschätzung einen nichtstatistischen Induktionsschluss: Man muss annehmen, dass die Untersuchungseinheiten in der geplanten Untersuchung denjenigen Untersuchungseinheiten gleichen, die zur Teststärkeermittlung dienten. Dieser Schluss ist insofern induktiv, als er von bekannten Daten auf unbekannte, noch zu testende Untersuchungseinheiten generalisiert. Diese Generalisierung kann beispielsweise pragmatisch dadurch gerechtfertigt sein, dass die Untersuchungseinheiten der geplanten Untersuchung auf dieselbe Art und Weise selektiert werden wie die Untersuchungseinheiten der durchgeführten Untersuchung. Die Generalisierung ist aber insofern nichtstatistisch, als sie nicht auf Annahmen über eine gemeinsame Grundgesamtheit und identischer Zufallsauswahl gegründet ist.

Die Methode der permutationsbasierten Teststärkeschätzung nach dem geschilderten Muster wurde durch Monte-Carlo-Simulationen empirisch evaluiert (Westermann u. Siemer, 2001). Die Ergebnisse zeigen durchgängig, dass permutationsbasierte Teststärkeschätzungen eine sehr gute Approximation der wahren Teststärke ergeben, wenn Verteilungsannahmen der entsprechenden parametrischen Teststärkeschätzungen erfüllt sind. Die Forderung, dass die permutationsbasierte Methode unter diesen Bedingungen die wahre Teststärke gut approximiert, kann als Minimalvoraussetzung für die Validität der Methode angesehen werden. Die Ergebnisse bedeuten

umgekehrt, dass unter Bedingungen, unter denen parametrische Teststärkeverfahren gültig sind, diese benutzt werden können, um Ergebnisse des permutationsbasierten Verfahrens zu approximieren. Ergebnisse über die Genauigkeit des Verfahrens unter Bedingungen, in denen Voraussetzungen entsprechender parametrischer Verfahren verletzt sind, stehen bisher aus. Andererseits kann man argumentieren, dass die Bedeutung von Randomisierungstests vor allem darin besteht, dass sie ein Modell für randomisierungsbasierte statistische Inferenz in lokalen Populationen bieten. Demgegenüber ist die Bedeutung von Randomisierungstests als »robuste Ersatzmethode«, wenn die Verteilungsvoraussetzungen von parametrischen Testverfahren verletzt sind, von untergeordneter Bedeutung.

Alternative Methoden zur Bestimmung der Teststärke

Eine Variante der permutationsbasierten Teststärkeschätzung besteht in einem bootstrapbasierten Verfahren. Der Bootstrap ist Randomisierungsmethoden sehr ähnlich, basiert aber auf dem Konzept des Ziehens mit Zurücklegen aus einer gegebenen Stichprobe, während Randomisierungsmethoden mit der Methode des Ziehens ohne Zurücklegen aus vorhandenen Daten arbeiten. Der Bootstrap ist prinzipiell geeignet, allgemeine nichtparametrische Teststärkeanalysen durchzuführen (Davison u. Hinkley, 1997; Efron u. Tibshiriani, 1993). Das prinzipielle Vorgehen ist dabei wie folgt: Aus einer vorhandenen Stichprobe (nach Residualisierung) werden neue Bootstrap-Stichproben mit Zurücklegen gezogen. Die Teststärke lässt sich dann als relative Häufigkeit sich ergebender signifikanter Ergebnisse (nach Hinzufügen eines festgelegten Effekts) in den Bootstrap-Stichproben schätzen. Dieses Verfahren lässt sich auch auf die Teststärkeermittlung von Randomisierungstests anwenden (Lunneborg, 2000). Die Bootstrap-Methode führt zu vergleichbaren Resultaten wie die Permutationsmethode, ist aber rechnerisch vergleichbar aufwändig wie die exakte konditionale Methode (Westermann u. Siemer, 2001).

Eine weitere Variante besteht in dem Einsatz von Monte-Carlo-Methoden zur Teststärkebestimmung von Randomisierungstests (Manly, 1997). Im Unterschied zur Permutationsmethode müssen hier jedoch nicht nur Effektstärken a priori fixiert werden, sondern es muss auch a priori eine Verteilung spezifiziert werden. Aus dieser spezifizierten Verteilung werden dann wiederholte Stichproben gezogen und zu der oder den Treatmentgruppen der spezifizierte Effekt hinzugefügt. Die Teststärke ergibt sich aus

dem relativen Anteil an signifikanten Randomisierungstests, die in den Monte-Carlo-Stichproben durchgeführt werden.

Schließlich kann die Teststärke von Randomisierungstests durch entsprechende parametrische Teststärkeschätzungen approximiert werden. Dieses Vorgehen basiert auf dem generellen Prinzip, dass in Situationen, in denen keine Zufallsauswahl vorgenommen wurde und die statistische Inferenz durch das Randomisierungsprinzip gerechtfertigt ist, die Verwendung von parametrischen Methoden dann zulässig ist, wenn die Ergebnisse der parametrischen Methode die Ergebnisse der randomisierungsbasierten Methoden approximieren (Bredenkamp, 1980; Edgington, 1995; Fisher, 1935; Maxwell u. Delaney, 1990; Pitman, 1937). Die Ergebnisse der vorliegenden Monte-Carlo-Studien (Westermann u. Siemer, 2001) zeigen, dass dies bei Vorliegen der parametrischen Voraussetzungen der Fall ist.

Abschließende Bemerkungen

Insgesamt hat der vorliegende Beitrag gezeigt, wie eine kohärente Definition des Teststärkebegriffs im Rahmen von randomisierungsbasierter lokaler statistischer Inferenz möglich ist. Damit wurde demonstriert, dass das im Rahmen der Neyman-Pearson-Testtheorie frequentistisch konzipierte Teststärkekonzept auch im Kontext von Randomisierungstests, die kausale Hypothesen in Hinblick auf eine gegebene lokale Population testen, sinnvoll definierbar ist.

Es wurde weiterhin gezeigt, wie diese Definition im Rahmen praktischer Teststärkebestimmungen in der Versuchsplanung angewendet werden kann. Als ein wichtiges Ergebnis bei der Evaluation dieser randomisierungsbasierten Teststärkebestimmung wurde geschildert, dass bei Vorliegen der entsprechenden Voraussetzungen die nonparametrisch bestimmte Teststärke von Randomisierungstests durch entsprechende parametrische Verfahren approximiert werden kann. Grundlage hierfür ist die generelle Idee, dass parametrische Verfahren insofern gerechtfertigt sind, als sie mit den Ergebnissen von randomisierungsbasierten Verfahren übereinstimmen. Im Fall von erfüllten parametrischen Voraussetzungen führen auch alle geschilderten Alternativmethoden zu approximativ gleichen Ergebnissen. Hat man hingegen Grund zu der Annahme, dass die entsprechenden Voraussetzungen nicht erfüllt sind, sollte die geschilderte, einfach durchzuführende Methode der naiv approximativen Bestimmung der Teststärke angewandt werden.

Literatur

Bredenkamp, J. (1972). *Der Signifikanztest in der Psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J., Feger, H. (1970). Kriterien für die Entscheidung über die Aufnahme empirischer Arbeiten in die Zeitschrift für Sozialpsychologie. *Zeitschrift für Sozialpsychologie, 1,* 43–47.

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

Cumming, G., Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

Davison, A.C., Hinkley, D.V. (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press.

Edgington, E.S. (1995). *Randomization tests* (3. Aufl.). New York: Dekker.

Efron, B., Tibshiriani, R.J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In W.H. Tack u. T. Herrmann (Hg.), *Enzyklopädie der Psychologie. Themenbereich B: Methodologie und Methoden. Serie I: Forschungsmethoden der Psychologie. Band 1: Methodologische Grundlagen der Psychologie.* Göttingen: Hogrefe.

Fidler, F., Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61,* 575–604.

Fisher, R.A. (1935). *Design of experiments.* New York: Hafner.

Gabriel, K.R., Hsu, C.F. (1983). Evaluation of the power of rerandomization tests, with application to weather modification experiments. *Journal of the American Statistical Association, 78,* 766–775.

Good, P. (1994). *Permutation tests: A practical guide to resampling methods for testing hypotheses.* New York: Springer-Verlag.

Hager, W., Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp u. H. Feger (Hg.), *Hypothesenprüfung. Enzyklopädie der Psychologie, Serie Forschungsmethoden* (Bd. 5, S. 24–238). Göttingen: Hogrefe.

Hoenig, J.M., Heisey, D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association, 50,* 946–967.

Lenth, R.V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55,* 187–193.

Ludbrook, J., Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician, 52,* 127–132.

Lunneborg, C.E. (2000). *Data analysis by resampling: Concepts and applications.* Pacific Grove, CA: Duxbury.

Lunneborg, C.E. (2001). Random assignment of available cases: Bootstrap standard errors and confidence intervals. *Psychological Methods, 6,* 402–412.

Manly, B.F.J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology.* London: Chapman & Hall.

Maxwell, S.E., Delaney, H.D. (1990). *Designing experiments and analyzing data. A model comparison perspective.* Pacific Grove, CA: Brooks/Cole Publishing Company.

Neyman, J., Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika, 20A,* 175–240.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society, 4*(Suppl.), 225–232.

Reichardt, C.S., Gollob, H.F. (1999). Justifying the use and increasing the power of a ttest for a randomized experiment with a convenience sample. *Psychological Methods, 4,* 117–124.

Siemer, M., Joormann, J. (2003a). Power and measures of effect size in ANOVA with fixed versus random nested factors. *Psychological Methods, 8,* 497–517.

Siemer, M., Joormann, J. (2003b). Assumptions and consequences of treating providers in therapy studies as fixed versus random effects: Reply to Crits-Cristoph, Tu, and Gallop (2003) and Serlin, Wampold, and Levin (2003). *Psychological Methods, 8,* 535–544.

Steiger, J.H., Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Mulaik u. J.H. Steiger (Hg.), *What if there were no significance tests?* (S. 221–257). Mahwah, NJ: Erlbaum.

Westermann, R., Siemer, M. (2001). Datenauswertung ohne Annahmen über Populationen und Zufallsauswahl: Permutationstests und Bootstrapping. In R. Silbereisen (Hg.), *Bericht über den 42. Kongress der Deutschen Gesellschaft für Psychologie* (S. 724–735). Göttingen: Hogrefe.

Wilkinson, L., The Task Force on Statistical Inference (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist, 54,* 595–604.

Willmes, K. (1987). *Beiträge zu Theorie und Anwendung von Permutationstests in der uni- und multivariaten Datenanalyse.* Trier: Fachbereich I der Universität Trier (unveröffentlichte Dissertation).

Teil III:
Anwendungen

Teil III:
Anwendungen

Cristina Massen

Vorteile deduktiver Forschungsstrategien am Beispiel eines Paradigmas aus der Sakkadenforschung

Zusammenfassung

Deduktives Vorgehen in der empirischen Wissenschaft ist dadurch ausgezeichnet, dass experimentelle Untersuchungen so ausgewählt werden, dass sie zwischen a priori vorliegenden konkurrierenden Theorien oder Theorieansätzen diskriminieren können. Dies setzt voraus, dass bereits adäquat formulierte und voneinander abgrenzbare Theorieansätze vorliegen, aus denen dann zwingend Hypothesen abgeleitet werden können. Häufig besteht jedoch die Schwierigkeit einer deduktiven Vorgehensweise gerade darin, ungenügend oder nicht eindeutig formulierte Theorieansätze in Beziehung zueinander zu setzen, Gemeinsamkeiten und Unterschiede zu identifizieren, und zu prüfen, ob aus ihnen überhaupt unterschiedliche Vorhersagen ableitbar sind. Im folgenden Beitrag wird an einem Beispiel aus der Forschung zu sakkadischen Augenbewegungen gezeigt, wie ein solcher Prozess des Vergleichs und der Gegenüberstellung heterogener Theorieansätze in einer konkreten Forschungssituation aussehen und welche Vorteile eine anschließende deduktive Überprüfung der Ansätze mittels eines Entscheidungsexperiments haben kann.

Einleitung

In der Forschung zu sakkadischen Augenbewegungen lässt sich vielfach ein befundzentriertes Vorgehen beobachten, bei dem die Untersuchung der Auswirkung einer experimentellen Manipulation auf bestimmte abhängige Variable im Vordergrund der Betrachtung steht und theoretische Aussagen post hoc aus den experimentellen Befunden abgeleitet werden. Dieses Vorgehen birgt die Gefahr voreiliger wissenschaftlicher Schlüsse in sich, da leicht übersehen werden kann, dass zwar die konstruierte Theorie oder theoretische Aussage mit den empirischen Befunden im Einklang steht, jedoch möglicherweise andere, nicht in Betracht gezogene Erklärungsmöglichkeiten existieren. Solange das Hauptaugenmerk auf dem gewonnenen Befund liegt und sich für diesen eine Erklärung »anbietet«, besteht für den Forschenden wenig Veranlassung, nach alternativen Erklärungen zu suchen. Das Ergebnis dieses be-

fundzentrierten Vorgehens ist häufig eine Forschungssituation, in der heterogene Erklärungsversuche, die zu einzelnen Befunden konstruiert worden sind, völlig ohne Bezug nebeneinander stehen und in der eine Bewertung der einzelnen Erklärungsversuche sehr schwierig wird. Im Gegensatz zum befundzentrierten Vorgehen ist deduktives Vorgehen in der empirischen Wissenschaft dadurch ausgezeichnet, dass das Hauptaugenmerk der Betrachtung nicht auf dem empirischen Befund, sondern auf konkurrierenden Theorien oder theoretischen Aussagen liegt. Experimentelle Untersuchungen werden so ausgewählt, dass sie zwischen diesen a priori vorliegenden konkurrierenden Theorien oder theoretischen Aussagen diskriminieren können. Liegen noch keine adäquat formulierten Theorien vor, ist der Forschende gezwungen, sich einen Überblick über existierende Theorieansätze zu verschaffen, diese in Beziehung zueinander zu setzen und zu prüfen, ob aus ihnen gegenteilige Vorhersagen ableitbar sind, die zu einer Entscheidung zwischen den Theorieansätzen beitragen können. Wichtig ist hierbei, dass eine positive Entscheidung für einen Theorieansatz immer nur vorläufig sein kann, sie kann durch spätere negative Entscheidungen immer wieder umgestoßen werden (Popper, 1982). Der Fortschritt besteht damit im Wesentlichen darin, dass bestimmte Theorieansätze falsifiziert werden, also von der weiteren Betrachtung ausgeschlossen werden können.

Der Beitrag soll beispielhaft darstellen, wie ein solcher Prozess der Gegenüberstellung von ohne Bezug nebeneinander stehenden Theorieansätzen in einer konkreten Forschungssituation aussehen kann und wie durch die sinnvolle Auswahl eines geeigneten Experiments zu einer Entscheidung zwischen den Theorieansätzen beigetragen werden kann.

Gegenstand der Betrachtung ist dabei die Generierung sakkadischer Augenbewegungen, also ruckartiger, schneller Bewegungen der Augäpfel, die dazu dienen, interessierende Teile des Gesichtsfelds in das Zentrum fovealen Sehens zu bringen. Sakkaden können fast reflexartig durch plötzlich auftauchende oder saliente neue Stimuli im Gesichtsfeld ausgelöst werden, sie können jedoch auch zu Intentionen und Zielen des Handelnden in Beziehung stehen und willentlich generiert werden. Eine interessante Frage ist dabei, inwieweit der Mensch Kontrolle über seine okulomotorische Orientierungsreaktion, seinen »Blickreflex« ausüben kann. Eine Möglichkeit zur Untersuchung dieser Frage bietet die Antisakkadenaufgabe, die erstmals 1978 von Hallett beschrieben wurde und sich mittlerweile zu einem insbesondere von klinischer und neuropsychologischer Seite (vgl. Everling u. Fischer, 1998, für einen Überblick) viel untersuchten Paradigma entwickelt hat.

In der Antisakkadenaufgabe wird die Versuchsperson instruiert, den Blickreflex zu einem links oder rechts in der Peripherie des Gesichtsfelds

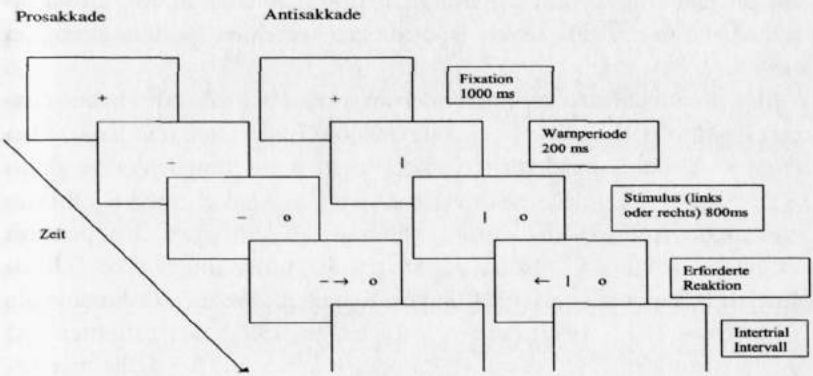


Abbildung 1: Abfolge der Ereignisse in einer typischen Antisakkadenaufgabe (rechts) im Vergleich zur Prosakkadenaufgabe (links).

auftauchenden Stimulus zu unterdrücken und stattdessen eine Sakkade auf die gegenüberliegende Seite des Fixationspunkts auszuführen. Abbildung 1 veranschaulicht die Abfolge der Ereignisse in einer typischen Antisakkadenaufgabe (rechts) im Vergleich zur Prosakkadenaufgabe (links).

Nach einem Fixationsintervall, währenddessen die Versuchsperson das Fixationskreuz anblicken soll, wird in randomisierter Abfolge links oder rechts des Fixationspunkts ein Stimulus präsentiert. In der Prosakkadenaufgabe ist die Versuchsperson instruiert, dem Blickreflex zu diesem Stimulus hin zu folgen und eine Sakkade zu diesem Ort auszuführen. In der Antisakkadenaufgabe dagegen soll sie die Sakkade zum visuellen Zielreiz hin unterlassen und stattdessen eine Sakkade zur anderen Seite generieren. In der so genannten »gap«-Bedingung wird dabei der Fixationspunkt kurz vor oder gleichzeitig mit dem Auftreten des peripheren Stimulus gelöscht, in der »overlap«-Bedingung bleibt dieser bestehen. Abhängige Variable in der Antisakkadenaufgabe sind typischerweise die Latenz der Reaktion, gemessen ab dem Zeitpunkt der Stimuluspräsentation, und der Prozentsatz fehlerhafter Prosakkaden zum Reiz hin. Des Weiteren werden häufig die Korrekturzeiten (Zeit vom Ende der ersten fehlerhaften Sakkade zum Reiz hin bis zum Beginn der zweiten, korrigierenden Sakkade) und Sakkadenamplituden berichtet. Gesunde Probanden können die Antisakkadenaufgabe ausführen, wobei allerdings ein gewisser Prozentsatz von reflexhaften Prosakkaden zum Reiz hin auftritt, der von Charakteristika der Aufgabe abhängt und sich auch nach mehrtägiger Übung nicht verändert (Everling u. Fischer, 1998). Mokler und Fischer (1999) fanden in ihrer Untersuchung,

dass ungefähr 96 Prozent der fehlerhaften Prosakkaden in der Antisakkadenaufgabe durch eine zweite Sakkade zur korrekten Position korrigiert werden.

Im Unterschied zu diesen Befunden an gesunden Probanden haben erstmals Guitton et al. (1985) signifikant erhöhte Fehlerraten und längere Latenzen in der Antisakkadenaufgabe bei Patienten mit frontalen im Vergleich zu temporalen Läsionen berichtet. In der Folgezeit hat es eine Vielzahl von Studien zur Antisakkadenleistung von anderen klinischen Gruppen mit vermuteten frontalen Defiziten gegeben, wobei unter anderem bei Schizophrenen (Clementz et al., 1994; Fukushima et al., 1988; Fukushima et al., 1994; Rosse et al., 1993; Sereno u. Holzman, 1995), bei Patienten mit Zwangsstörungen (Tien et al., 1992) oder solchen mit der Alzheimer-Erkrankung (Currie et al., 1991; Fletcher u. Sharpe, 1986) eine defizitäre Leistung in der Antisakkadenaufgabe nachgewiesen werden konnte.

Erklärungsansätze zur Antisakkade: Zum Stand der Forschung

Trotz des großen Interesses an der Antisakkadenaufgabe insbesondere im klinischen und neuropsychologischen Bereich existieren bis heute nur einige wenige Arbeiten, die sich mit der Frage nach den in dieser Aufgabe wirksamen inhibitorischen Mechanismen befasst haben. Der Stand der Forschung lässt sich hierbei als ein typisches Produkt befundzentrierten Vorgehens beschreiben: Heterogene Theorieansätze, die primär um einzelne Befunde herum konstruiert worden sind, stehen weitgehend ohne Bezug nebeneinander. Im Folgenden soll etwas näher auf diese einzelnen Ansätze eingegangen werden.

Ein früher Erklärungsansatz zur Leistung in der Antisakkadenaufgabe stammt von Hallett und Adams (1980). Diese Autoren vergleichen die Antisakkadenaufgabe mit der »pulse over return«-Version des Doppelsprungparadigmas (vgl. z. B. Becker u. Jürgens, 1979), in der kurz nach dem ersten Sprung eines Fixationspunkts zu einer Position nach rechts oder links ein zweiter Sprung dieses Stimulus auf die kontralaterale Seite erfolgt, wobei die Probanden instruiert sind, jedem Sprung des Zielreizes mit den Augen zu folgen. Als grundlegende Beobachtung in diesem Paradigma lässt sich feststellen, dass die sakkadische Reaktion der Versuchsperson von dem zeitlichen Abstand (D) zwischen der Präsentation des zweiten Zielsprungs und dem Beginn ihrer Reaktion auf den ersten Sprung abhängt. Wird der zweite

Zielsprung sehr kurz nach dem ersten präsentiert und/oder die Versuchsperson hat im Mittel lange sakkadische Latenzen auf den ersten Zielsprung, dann ist der Abstand D groß, der zweite Zielsprung kann die Reaktion auf den ersten Sprung noch beeinflussen und die Primärsakkade der Versuchsperson landet direkt auf der zweiten, endgültigen Zielposition. Wird der zweite Zielsprung erst nach einem längeren Zeitintervall präsentiert und/oder die Versuchsperson hat im Mittel sehr kurze Latenzen auf den ersten Zielsprung, dann ist D klein, die erste Reaktion kann nicht mehr gestoppt werden und die Versuchsperson führt zunächst eine Sakkade zum ersten Zielort und dann eine zweite zum endgültigen Zielort aus. Systematische Variationen des zeitlichen Abstands zwischen erstem und zweitem Zielreiz haben in diesem Paradigma ergeben, dass bei jedem Probanden die Wahrscheinlichkeit für eine Primärsakkade zur ersten Zielposition genau 50 Prozent beträgt, wenn man als zeitlichen Abstand der beiden Zielreize einen Wert wählt, der der mittleren sakkadischen Latenz eines Probanden auf einen einzelnen Zielreiz hin minus einem konstanten Wert von 134 ms entspricht. Umgekehrt ausgedrückt, ist der zeitliche Abstand zwischen beiden Zielreizen bei einem Probanden so zu wählen, dass die zeitliche Überlappung zwischen dem angenommenen ersten und zweiten Programmierprozess 134 ms beträgt.

Hallett und Adams (1980) sehen eine strukturelle Ähnlichkeit zwischen diesem Paradigma und der Antisakkadenaufgabe in dem Sinne, dass auch in der Antisakkadenaufgabe auf einen peripheren visuellen Zielreiz hin eine sakkadische Reaktion zur kontralateralen Seite (allerdings ohne zweiten visuellen Zielreiz) gefordert ist. Ausgehend von der Beobachtung, dass sich in ihrem Datensatz die Latenzen von Antisakkaden fast perfekt durch die Regressionsgleichung

$$(1)\ \overline{A} = 2.01\ \overline{P} - 144.5$$

beschreiben ließen, wobei $\overline{A}$ die mittlere (empirisch ermittelte) Antisakkadenlatenz und $\overline{P}$ die mittlere (empirisch ermittelte) Prosakkadenlatenz eines Probanden ist, gelangen sie zu folgender Erklärung der in der Antisakkadenaufgabe ablaufenden Prozesse. Es wird angenommen, dass zunächst automatisch und ungewollt eine reflexhafte Prosakkade zum Zielreiz hin programmiert wird. Dieser Prozess wird von der Programmierung einer zweiten Prosakkade zu dem Sakkadenziel auf der kontralateralen Seite gefolgt. Wenn die Programmierung der zweiten Prosakkade rechtzeitig erfolgt und die zeitliche Überlappung beider Prosakkadenprozesse groß genug ist, kann die erste reflexhafte Prosakkade gestoppt werden und als sakkadische

Reaktion erfolgt eine korrekte Antisakkade mit einer mittleren Latenz, wie sie in Gleichung 1 angegeben ist. Diese mittlere Latenz von korrekt ausgeführten Antisakkaden ($\overline{A}$) setzt sich zusammen aus der Summe der Latenzen zweier zeitlich variabler Prosakkadenprozesse, die beide durch den gleichen Mittelwert $\overline{P}$ beschrieben werden können (sodass ihre Summe im Mittel 2 $\overline{P}$ beträgt) und die ein konstantes Intervall in der Zeit überlappen (144.5 ms). Dieses konstante Überlappungsintervall soll in etwa dem aus dem Doppelsprungparadigma bekannten konstanten Überlappungsintervall von 134 ms entsprechen.[1] Ist die Programmierung der zweiten Prosakkade verzögert und das Überlappungsintervall dementsprechend kleiner, resultiert zunächst eine fehlerhafte Sakkade zum Reiz hin, die dann von einer Korrektursakkade zur anderen Seite gefolgt wird. Abgesehen von der Tatsache, dass der Ansatz von Hallett und Adams (1980) und die von ihnen aufgestellte Gleichung (1) einige Unstimmigkeiten enthalten, auf die an dieser Stelle nicht weiter eingegangen werden kann (insbesondere gehen die Autoren nicht auf die individuell unterschiedlichen Fehlerraten in der Antisakkadenaufgabe ein, die in die Überlegungen mit einbezogen werden müssten, vgl. dazu Massen, 2002), werden sie von den Autoren post hoc aus der von ihnen gefundenen Regressionsgleichung abgeleitet und nicht konsequent deduktiv durch die Generierung und Testung weiterer Hypothesen überprüft, die von dem Ausgangsbefund unabhängig sind. Dass ein solches Vorgehen zu falschen Schlussfolgerungen führen kann, wird im nächsten Abschnitt näher erörtert.

Weitere Erklärungsansätze zu den inhibitorischen Mechanismen in der Antisakkadenaufgabe stammen aus der Neurophysiologie. Schlag-Rey et al. (1997) haben in einem Experiment an Rhesusaffen untersucht, wie die visuell-rezeptiven Neuronen einer Hemisphäre, die durch das Auftreten des Stimulus in der Antisakkadenaufgabe erregt werden, die Bewegungsneuronen der entgegengesetzten Hemisphäre aktivieren können, die für das Ausführen der Sakkade zur anderen Seite zuständig sind. Dieses Problem könnte durch eine Verschiebung der rezeptiven Felder, wie sie in anderen Untersuchungen nachgewiesen werden konnte (vgl. Duhamel et al., 1992), gelöst werden. Nach dieser Hypothese würde die Instruktion zur Antisakkade bewirken, dass sich die rezeptiven Felder der Neuronen ändern und der Stimulus nun genau diejenigen Neuronen der ipsilateral zur Stimulusseite liegenden Hemisphäre aktiviert, die auch für die Programmierung einer Sakkade zu einem Zielort der kontralateralen Seite zuständig sind. Die-

1 Überlegungen zur Erklärung der Abweichung der 144.5 ms zu den 134 ms des Doppelsprungparadigmas finden sich in Hallett und Adams (1980).

se Erklärung impliziert, dass der Stimulus keine visuelle Aktivität in der kontralateralen Hemisphäre mehr auslöst und damit auch der Reflex, eine Prosakkade zum Reiz hin auszuführen, nicht mehr vorhanden wäre. Die andere Erklärungsmöglichkeit ist dagegen, dass sich die rezeptiven Felder der Neuronen nicht ändern und stattdessen das Sakkadenziel auf der dem Stimulus gegenüberliegenden Seite in irgendeiner Weise berechnet und an Neuronen mit den entsprechenden Bewegungsfeldern (die in der ipsilateral zum Stimulus gelegenen Hemisphäre liegen) weitergegeben wird. In diesem Fall würden sowohl diese aktivierten Neuronen als auch Neuronen der kontralateralen Hemisphäre, die durch den Stimulus direkt und automatisch aktiviert werden, miteinander konkurrieren und parallel feuern, wobei, wie die Autoren vermuten, die Neuronen mit den stärkeren Entladungsraten den Konkurrenzkampf gewinnen.

Die Autoren konnten in ihrem Experiment zeigen, dass Neuronen der supplementären Augenfelder in der Antisakkadenaufgabe ihre rezeptiven und Bewegungsfelder beibehalten und dass Bewegungsneuronen vor Antisakkaden stärker feuerten als vor Prosakkaden mit denselben Trajektorien. Dieses Ergebnis wird von den Autoren als Beleg für die Hypothese gewertet, dass in der Antisakkadenaufgabe eine parallele Programmierung der reflexhaften Prosakkade und der korrekten Antisakkade stattfindet, von denen sich das stärkere Programm durchsetzt.

Der Ansatz von Schlag-Rey et al. (1997) enthält deutlich mehr deduktive Anteile als derjenige von Hallett und Adams, da zwei sich widersprechende theoretische Aussagen, die Hypothese einer Verschiebung der rezeptiven Felder und die Hypothese der parallelen Programmierung zweier sakkadischer Programme, gegeneinander getestet werden. Als Ergebnis dieses Vorgehens lässt sich eindeutig feststellen, dass die Hypothese einer Verschiebung der rezeptiven Felder als empirisch widerlegt betrachtet werden kann und dass auch physiologische Evidenz für eine automatische Programmierung der reflexhaften Prosakkade in der Antisakkadenaufgabe spricht, wie sie schon von Hallett und Adams (1980) angenommen wurde. Schlag-Rey et al. (1997) stellen keinen Bezug zu der Arbeit von Hallett und Adams (1980) her, ihre Ergebnisse scheinen jedoch eher dafür zu sprechen, dass die Absage der reflexhaften Prosakkade in der Antisakkadenaufgabe nicht durch ein mit einer zeitlichen Verzögerung erstelltes zweites Prosakkaden-Programm, das dem ersten in seinem Zeitbedarf entspricht, geleistet wird, sondern durch die unterschiedlich starke neuronale Aktivität zweier parallel ablaufender, aber vollkommen unterschiedlicher Programmierprozesse. Welcher von beiden Ansätzen zutrifft, kann aufgrund der Daten von Schlag-Rey et al. (1997) jedoch nicht eindeutig entschieden werden. Im

nächsten Abschnitt wird näher erläutert werden, welche unterschiedlichen Vorhersagen sich aus beiden Ansätzen ergeben und wie diese experimentalpsychologisch geprüft werden können. Zuvor soll jedoch noch auf weitere physiologische Arbeiten eingegangen werden, die einen anderen inhibitorischen Mechanismus in der Antisakkadenaufgabe nahe legen.

Im Gegensatz zu den bisher diskutierten Arbeiten argumentieren Everling et al. (1998, vgl. auch Everling et al., 1999; Everling u. Munoz, 2000), dass die reflexhafte Prosakkade in der Antisakkadenaufgabe durch hemmende Prozesse in der Instruktionsphase, also noch vor der Präsentation des Stimulus, verhindert wird. Everling et al. (1999) nahmen hierzu Einzelzellableitungen an verschiedenen Neuronentypen im Colliculus superior von Rhesusaffen während der Ausführung von Pro- und Antisakkaden vor, wobei Pro- und Antisakkaden in gemischten Blöcken gefordert wurden und die Farbe des Fixationskreuzes jeweils anzeigte, ob eine Pro- oder Antisakkade ausgeführt werden sollte. Die Autoren fanden in Antisakkadendurchgängen im Vergleich zu Prosakkadendurchgängen während der Instruktionsphase eine erhöhte Aktivität von Fixationsneuronen (Neuronen, die tonische Aktivität in Fixationsintervallen zeigen) und eine reduzierte Aktivität von visuell-rezeptiven und so genannten *Build-up*-Neuronen (Neuronen, die ihre Aktivität zum Ende eines Fixationsintervalls langsam steigern und dann begleitend zu einer Sakkade zu ihrem motorischen Feld feuern). Aus diesen unterschiedlichen Aktivitätsmustern von visuell-rezeptiven, *Build-up*- und Fixationsneuronen in der Instruktionsphase von Pro- und Antisakkaden schließen die Autoren, dass die Unterdrückung der reflexhaften Antwort in der Antisakkadenaufgabe von aktiven Hemmungsprozessen noch vor der Präsentation des Stimulus abhängt. Sie sehen diese Schlussfolgerung außerdem gestützt durch eine weitere Untersuchung (vgl. Everling et al., 1998), in der die Aktivität von *Build-up*-Neuronen in Antisakkadendurchgängen mit fehlerhafter Prosakkade mit der in Durchgängen ohne fehlerhafte Prosakkade, also korrekt ausgeführter Antisakkade verglichen wurde. Die Autoren fanden eine erhöhte Aktivität von *Build-up*-Neuronen während der Instruktionsphase von Durchgängen mit fehlerhafter Prosakkade, was wiederum als Beleg dafür gewertet wird, dass das Auftreten eines Fehlers in der Antisakkadenaufgabe von Prozessen vor der Stimuluspräsentation abhängt.

Auch was diesen Ansatz betrifft, lässt sich wieder feststellen, dass nicht der Versuch unternommen wird, verschiedene Hypothesen in die Diskussion einzubeziehen und deren Vorhersagen direkt miteinander zu vergleichen. Im vorliegenden Fall führt dies dazu, dass die Autoren übersehen, dass ihre Befunde genauso von der Hypothese von Hallett und Adams

(1980) oder der Hypothese der parallelen Programmierung erklärt werden können. So ist aus anderen Untersuchungen bekannt, dass die Aktivität von *Build-up*-Neuronen vor der Stimuluspräsentation auch in der Prosakkadenaufgabe variiert, wobei erhöhte Aktivität mit kürzeren Latenzen von Prosakkaden einhergeht. Nimmt man nun mit Hallett und Adams (1980) oder der Wettlauf-Hypothese der parallelen Programmierung an, dass in der Antisakkadenaufgabe zunächst reflexhaft eine Prosakkade zum Reiz hin programmiert wird und das Stoppen dieser Reaktion (ob nun indirekt durch die parallele Programmierung der Antisakkade oder durch die zweite Prosakkade) bei den schnellen Prosakkaden nicht gelingt, wäre ebenfalls ein Zusammenhang zwischen der Aktivität von *Build-up*-Neuronen vor korrekten und fehlerhaften Antisakkaden beobachtbar, da eine höhere Aktivität von *Build-up*-Neuronen mit schnellen Prosakkaden, die zum Fehler führen, einherginge. Der funktionale Mechanismus, der hier zum Fehler führt, wäre in diesem Fall jedoch nicht eine zu niedrige inhibitorische Aktivität vor der Stimuluspräsentation, sondern der Wettlauf zwischen der Absage des Reflexes und der automatisch programmierten Prosakkade nach der Stimuluspräsentation.

Ableitung eines Entscheidungsexperiments

Zusammenfassend lässt sich festhalten, dass das weitgehend befundzentrierte Vorgehen in diesem Forschungsbereich zwar einige interessante Befunde und Erklärungsansätze zu den Inhibitionsmechanismen in der Antisakkadenaufgabe hervorgebracht hat, dass der momentane Forschungsstand jedoch als unbefriedigend zu bezeichnen ist, solange nicht der Versuch unternommen wird, zu einer Entscheidung zwischen den Ansätzen zu kommen. Hier ist also deduktives Vorgehen gefordert, das zunächst einmal mit der wichtigen Frage beginnen muss, inwiefern sich aus den existierenden Ansätzen unterschiedliche Beobachtungsaussagen ableiten lassen, deren empirische Prüfung zu einer Entscheidung beitragen könnte (vgl. Bredenkamp, 1980).

Vergleicht man zunächst einmal den Ansatz von Hallett und Adams (1980) und die Hypothese einer parallelen Programmierung von Pro- und Antisakkade sensu Schlag-Rey et al. (1997), so lässt sich feststellen, dass beide eine automatische Programmierung und anschließende Absage der reflexhaften Prosakkade behaupten. Nach Hallett und Adams (1980) wird die reflexhafte Prosakkade durch die nachfolgende Programmierung einer zweiten Prosakkade gestoppt, wohingegen Schlag-Rey et al. (1997) anneh-

men, dass parallel zu der Programmierung der reflexhaften Sakkade direkt eine Antisakkade programmiert wird. Eine Entscheidung zwischen beiden Ansätzen läuft auf die Frage hinaus, ob sich experimentelle Manipulationen finden lassen, die sich nur auf die Latenzen von visuell gesteuerten Prosakkaden auswirken, ohne die Latenzen von Antisakkaden zu beeinflussen (oder umgekehrt). Dies würde gegen die von Hallett und Adams (1980) postulierte Beziehung $\overline{A} = 2.01\ \overline{P} - 144.5$ und ihre Annahme zweier gleichartiger, zeitlich überlappender Programmierprozesse sprechen, während es mit der Vorstellung zweier voneinander unabhängiger und selektiv manipulierbarer Programmierprozesse vereinbar ist.

Die Hypothese eines Wettlaufs des endogenen Programms gegen das reflexhafte Programm sagt außerdem voraus, dass eine selektive Beschleunigung des reflexhaften Prozesses zu einer Erhöhung der Fehlerrate in der Antisakkadenaufgabe führen sollte, wohingegen eine selektive Beschleunigung des endogenen Programms zu einer Reduktion der Fehlerrate führen sollte. Die Auswirkung einer experimentellen Manipulation auf das angenommene reflexhafte Programm ist dabei nicht direkt beobachtbar, da es nur im Fall eines Fehlers im Verhalten sichtbar wird. Man kann jedoch sinnvollerweise annehmen, dass der Effekt einer experimentellen Variablen auf den angenommenen reflexhaften Prozess in der Antisakkadenaufgabe demjenigen entspricht, den sie auf Prosakkaden in einer Prosakkadenaufgabe hat (für eine Diskussion dieser Annahme vgl. Massen, 2002).

Ein Test zwischen beiden Ansätzen läuft also auf die Identifikation experimenteller Variablen hinaus, die sich unterschiedlich auf die Latenzen von Pro- und Antisakkaden (bzw. visuell-gesteuerten und endogen generierten Sakkaden) auswirken und gleichzeitig das oben beschriebene Muster in den Fehlerraten bei Antisakkaden zeigen, das nur von der Hypothese der parallelen Programmierung vorhergesagt wird. Betrachtet man nun einen solchen Test im Kontext der Hypothese einer aktiven Inhibition der reflexhaften Prosakkade noch vor der Stimuluspräsentation, so lässt sich zunächst feststellen, dass diese Hypothese ebenfalls erklären könnte, warum eine Beschleunigung der reflexhaften Prosakkade zu einer Erhöhung der Fehlerrate in der Antisakkadenaufgabe führt. So ist bekannt, dass die Latenz von Prosakkaden mit der Grundaktivität von sakkadenbezogenen motorischen Neuronen im Colliculus superior vor der Stimuluspräsentation korreliert (Dorris et al., 1997). Nimmt man nun an, dass in Bedingungen mit kürzeren Prosakkaden-Latenzen eine stärkere Grundaktivität von relevanten motorischen Neuronen vorliegt, die schwerer aktiv zu unterdrücken ist, wäre ebenfalls eine höhere Fehlerrate in Antisakkadendurchgängen zu beobachten. Auf der anderen Seite sagt die Hypothese der aktiven Inhibition

keinen Zusammenhang zwischen der Geschwindigkeit der intentionalen Programmierung der Antisakkade und der Fehlerrate in der Antisakkadenaufgabe voraus, da die Fehlerrate lediglich vom Erfolg der Unterdrückung der reflexhaften Antwort abhängen und nicht mit der Generierung der Antisakkade in Zusammenhang stehen sollte.

Eine gute Möglichkeit, das Wettlaufmodell der parallelen Programmierung zu prüfen, besteht in der Untersuchung der Auswirkung von »inhibition of return« auf die Fehlerraten in der Antisakkadenaufgabe. Mit dem Terminus »inhibition of return« (IOR) wird dabei der Befund bezeichnet, dass ein vor einem Zielreiz präsentierter visueller Hinweisreiz, der bei kurzen Zeitintervallen zwischen Hinweisreiz und Zielreiz (bis ca. 300 ms) schnellere Reaktionen auf den Zielreiz bewirkt (Posner, 1980), bei längeren Zeitintervallen eine gegenteilige Wirkung entfaltet und eine räumliche Reorientierung zum Zielreiz erschwert (vgl. Posner und Cohen, 1984). »Inhibition of return« kann sowohl exogen durch die vorherige Präsentation eines visuellen Hinweisreizes am Ort des Zielreizes ausgelöst werden, als auch endogen durch die vorherige Planung oder Ausführung einer Sakkade zum Ort des Zielreizes (vgl. Rafal et al., 1989), wobei das Sakkadenziel nicht durch einen exogenen visuellen Hinweisreiz angezeigt wird, sondern beispielsweise durch einen Pfeil am Ort des Fixationspunkts. Rafal et al. (1994) haben den IOR-Effekt auf die Latenzen von Pro- und Antisakkaden untersucht und gefunden, dass bei exogener Induktion von »inhibition of return« die Latenzen von Pro- und Antisakkaden im gleichen Ausmaß verlängert sind, wenn der Stimulus für die jeweilige Sakkade am Ort des Hinweisreizes erscheint (im Vergleich zu der Bedingung, in der der Stimulus für die Sakkade auf der anderen Seite erscheint). Sie schließen aus diesem Befund, dass bei exogener IOR-Induktion frühe Verarbeitungsprozesse, die sowohl Bestandteil der Programmierung von Prosakkaden als auch der von Antisakkaden sind (z. B. die Detektion des Stimulus) erschwert werden. Wird »inhibition of return« dagegen endogen durch eine vorherige Sakkade erzeugt, zeigt sich ein signifikant stärkerer Einfluss der Manipulation auf die Latenzen von Prosakkaden im Vergleich zu denen von Antisakkaden, wenn der Stimulus für die jeweilige Sakkade am Zielort der vorherigen Sakkade erscheint. Dies wird von den Autoren darauf zurückgeführt, dass bei vorheriger Ausführung einer endogenen Sakkade zusätzlich zur erschwerten Detektion des Stimulus ein motorischer Alternationsbias wirksam wird, der einen Wechsel der Sakkadenrichtung nahe legt und eine Sakkade zur anderen Seite erleichtert.[2]

2 Evidenz für die Existenz eines solchen Alternationsbias lieferte schon eine Untersuchung von Posner et al. (1985), in der Probanden nach der IOR-Induktion freigestellt wurde, zu einem von zwei bilateral und peripher dargebotenen Zielreizen eine Sakkade auszuführen.

Soll demnach eine Prosakkade zum Zielort der vorigen Sakkade ausgeführt werden, ist die Detektion des Stimulus erschwert und der motorische Alternationsbias behindert zusätzlich die erneute Sakkade zu diesem Ort, sodass sich beide Effekte addieren. Wird dagegen der Stimulus für die Antisakkade am Zielort der vorigen Sakkade präsentiert, ist zwar die Detektion des Stimulus erschwert, der motorische Alternationsbias erleichtert jedoch die geforderte motorische Reaktion auf die andere Seite, sodass sich beide Effekte gegenseitig aufheben (im Vergleich zu der Bedingung, in der Zielort der ersten Sakkade und Stimulus für die Antisakkade auf unterschiedlichen Seiten erscheinen). Insgesamt ergäbe sich somit bei IOR-Induktion durch eine vorherige Sakkade für Prosakkaden ein großer Unterschied in den Latenzen zwischen den beiden Hinweisreiz-Bedingungen, wohingegen bei den Antisakkaden der Unterschied nicht existent oder zumindest signifikant kleiner sein sollte.

Zusammenfassend lassen sich aus dieser Befundlage die folgenden Vorhersagen bezüglich der Hypothese der parallelen Programmierung und den Fehlerraten in der Antisakkadenaufgabe, die von Rafal et al. (1994) nicht berichtet werden, ableiten. Wird »inhibition of return« exogen durch einen visuellen Hinweisreiz erzeugt, sind sowohl der angenommene reflexhafte als auch der endogene Programmierprozess in der Antisakkadenaufgabe von der erschwerten Wahrnehmung des Stimulus betroffen und verlangsamt, es sollte sich daher kein Unterschied in der Fehlerrate ergeben. Bei IOR-Induktion durch eine vorherige Sakkade ist dagegen in der Bedingung, in der der Stimulus am Zielort der vorigen Sakkade präsentiert wird, der reflexhafte Prozess gegenüber der anderen Bedingung verlangsamt, nicht jedoch der endogene Prozess, sodass hier eine reduzierte Fehlerrate in der Antisakkadenaufgabe zu erwarten ist. Abbildung 2 zeigt die Ergebnisse eines Experiments (vgl. Massen, in Vorb.), in dem »inhibition of return« durch einen visuellen Hinweisreiz exogen erzeugt wurde. Der visuelle Hinweisreiz wurde dabei vorher entweder am Ort des Stimulus in der Pro- und Antisakkadenaufgabe dargeboten (valider Hinweisreiz), an einer neutralen Position in der Mitte des Stimulusdisplays (neutraler Hinweisreiz) oder auf der dem Stimulusort gegenüberliegenden Seite (invalider Hinweisreiz). Entsprechend der Vorhersage zeigt sich zum einen, dass ein valider Hinweisreiz die Latenzen von Pro- und Antisakkaden im Vergleich zu den beiden anderen Bedingungen signifikant er-

---

Probanden wählten signifikant häufiger den Zielreiz, der nicht »inhibition of return« unterlag.

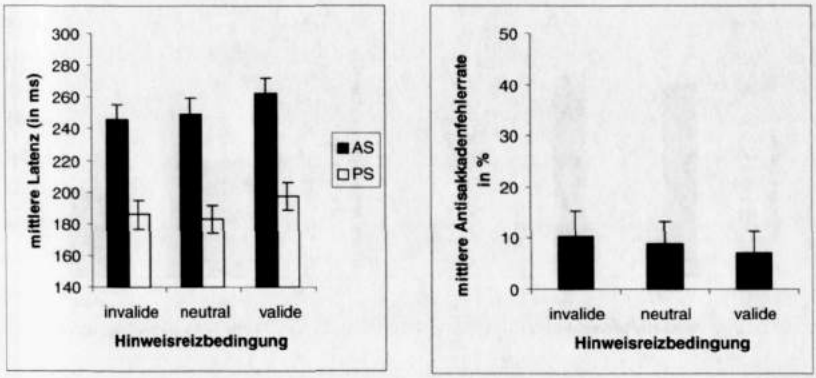


Abbildung 2: Mittlere Latenzen von Antisakkaden (AS) und Prosakkaden (PS) (links) sowie Antisakkadenfehlerraten (rechts) in Abhängigkeit von der Hinweisreiz-Bedingung in Experiment 1.

höht und zwar im selben Ausmaß (die Interaktion der Faktoren Sakkadentyp und Bedingung ist statistisch nicht signifikant).

Die Fehlerraten in Antisakkadendurchgängen unterscheiden sich hypothesenkonform nicht signifikant zwischen den IOR-Bedingungen.

In einem weiteren Experiment (Massen, in Vorb.) wurde »inhibition of return« durch eine vorherige endogene Sakkade zum späteren Stimulusort (valider Hinweisreiz) oder auf die dem Stimulusort gegenüberliegende Seite (invalider Hinweisreiz) induziert. Wie in Abbildung 3 zu sehen, unterscheiden sich in diesem Fall die Latenzen von Prosakkaden zwischen den Hinweisreiz-Bedingungen stärker als diejenigen von Antisakkaden, wobei die Interaktion der Faktoren Sakkadentyp und Hinweisreiz-Bedingung statistische Signifikanz erreicht.

Entsprechend der Vorhersage zeigt sich bei den Fehlerraten von Antisakkaden diesmal ein signifikanter Unterschied zugunsten einer niedrigeren Fehlerrate in der Bedingung mit validem Hinweisreiz, da in dieser Bedingung das reflexhafte Programm im Vergleich zum endogenen Antisakkadenprogramm stärker verlangsamt wird.

Vergleicht man des Weiteren die Fehlerraten bei Antisakkaden zwischen den beiden berichteten Experimenten aus Massen (in Vorb.), so wird ersichtlich, dass die Fehlerrate im zweiten Experiment signifikant höher ist als im ersten. Dieses Befundmuster erlaubt einen weiteren Test der Hypothese der parallelen Programmierung über die beiden Experimente hinweg. Es lässt sich vorhersagen, dass entweder die Latenzen von Prosakkaden in Experiment 2 selektiv beschleunigt sind, oder die Latenzen von Antisakka-

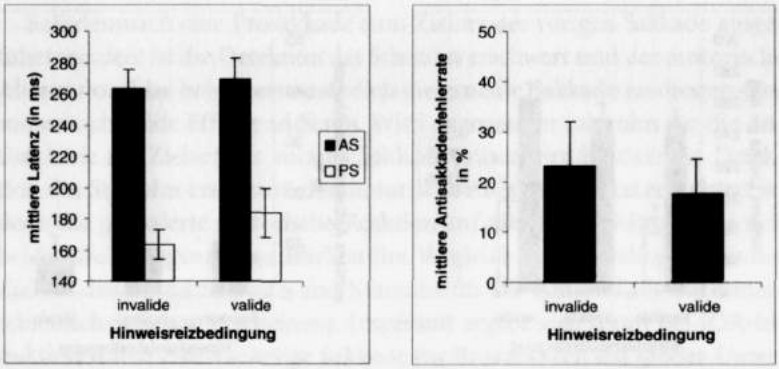


Abbildung 3: Mittlere Latenzen von Antisakkaden (AS) und Prosakkaden (PS) (links) sowie Antisakkadenfehlerraten (rechts) in Abhängigkeit von der Hinweisreiz-Bedingung in Experiment 2.

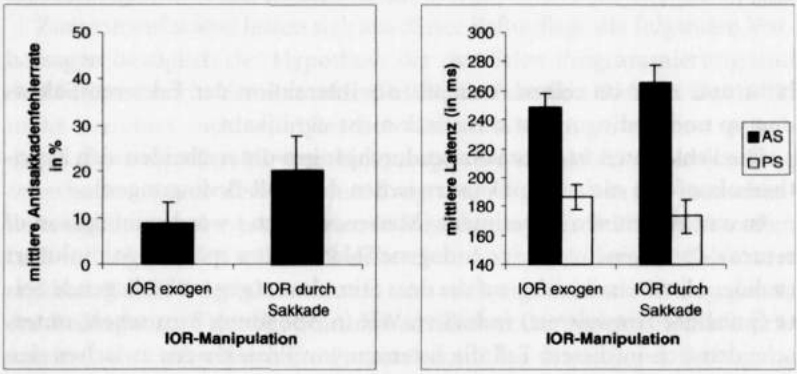


Abbildung 4: Mittlere Latenzen von Antisakkaden (AS) und Prosakkaden (PS) (rechts) sowie Antisakkadenfehlerraten (links), gemittelt über die jeweiligen Bedingungen, im Vergleich zwischen den beiden Experimenten.

den selektiv verlangsamt (oder beides). Mit anderen Worten wird für die Sakkadenlatenzen eine Interaktion zwischen dem Faktor Sakkadentyp (Prosakkade vs. Antisakkade) und dem Faktor Experiment (IOR durch exogenen Hinweisreiz vs. IOR durch endogene Sakkade) vorhergesagt. In Abbildung 4 sind die Fehlerraten bei Antisakkaden der beiden Experimente, jeweils gemittelt über die einzelnen Bedingungen, sowie die mittleren Latenzen von Pro- und Antisakkaden dargestellt.

Wie vorhergesagt, zeigt sich für die Latenzen eine signifikante Interaktion zwischen den Faktoren Experiment und Sakkadentyp, die deskriptiv sowohl auf kürzere Prosakkaden-Latenzen als auch auf längere Antisakkaden-Latenzen in Experiment 2 zurückzuführen ist. Zwar ist unklar, ob die

verlängerten Antisakkadenlatenzen und verkürzten Prosakkadenlatenzen in Experiment 2 auf Stichprobenunterschiede zwischen den beiden Experimenten zurückgehen oder auf die experimentelle Manipulation (vgl. z. B. Bredenkamp, 1996) zurückzuführen sind; diese Frage kann jedoch in unserem Kontext vernachlässigt werden, da die Vorhersage aus der Hypothese der parallelen Programmierung unabhängig von dieser Frage getestet werden kann.

Diskussion der Befunde im Kontext der Hypothesen

Die dargestellten empirischen Befunde sind nicht mit der Hypothese von Hallett und Adams (1980) vereinbar. Nach dieser Hypothese sollte jede Manipulation, die die Latenzen von Prosakkaden im Mittel um einen bestimmten Betrag × verlängert, die Latenzen von Antisakkaden um 2x verlängern. Abbildung 3 zeigt, dass entgegen dieser Vorhersage die IOR-Manipulation einen größeren Effekt auf die Prosakkadenlatenzen als auf die Antisakkadenlatenzen hat, und Abbildung 4 macht deutlich, dass eine Verlängerung der Prosakkadenlatenzen sogar mit einer Verkürzung der Antisakkadenlatenzen in derselben Bedingung einhergehen kann.

Betrachtet man diese Befunde aus dem Blickwinkel der Hypothese der aktiven Inhibition, so müsste man die unterschiedlichen Antisakkadenfehlerraten zwischen den IOR-Bedingungen in Experiment 2 (vgl. Abbildung 3) auf Unterschiede in der inhibitorischen Aktivität der Probanden in der Instruktionsphase zurückführen. Konkret müsste dabei angenommen werden, dass Probanden weniger aktive inhibitorische Aktivität vor der Stimuluspräsentation zeigen, wenn sie vorher eine Sakkade zu einem Zielort auf der dem späteren Stimulusort gegenüberliegenden Seite ausgeführt haben, als wenn diese Sakkade zur selben Seite gegangen ist. Da die Probanden nicht wissen können, auf welcher Seite der Stimulus im nächsten Durchgang erscheinen wird und die Wahrscheinlichkeit für jede Alternative 50 Prozent beträgt, muss jedoch davon ausgegangen werden, dass die inhibitorische Aktivität in jedem Durchgang gleich ist beziehungsweise nicht von der Präsentationsseite des folgenden Stimulus abhängen kann. Auch die Hypothese der aktiven Inhibition versagt also bei der Erklärung der oben dargestellten Ergebnisse, sie ist allerdings nicht gänzlich unvereinbar mit der Hypothese der parallelen Programmierung, wenn man annimmt, dass eine Erhöhung der Fixationsaktivität vor der Stimuluspräsentation den reflexhaften Prozess zugunsten des endogenen Programms abschwächen

könnte. Trotzdem bliebe in diesem Fall zu erklären, warum ein Zusammenhang zwischen der Geschwindigkeit der Generierung des endogenen Antisakkadenprogramms (vgl. Massen, in Vorb.) und der Fehlerrate in der Antisakkadenaufgabe existiert.

Zusammenfassend lässt sich also feststellen, dass sowohl die Hypothese von Hallett und Adams (1980) als auch die (häufig implizit gemachte) Annahme einer aktiven Hemmung des Reflexes als entscheidendem Inhibitionsmechanismus in der Antisakkadenaufgabe aufgrund der Befunde zurückgewiesen werden müssen, sodass beim jetzigen Kenntnisstand nur die Hypothese eines Wettlaufs zweier paralleler Programmierprozesse in der Antisakkadenaufgabe beibehalten werden kann.

Schlussfolgerungen

Die vorangegangenen Betrachtungen haben deutlich gemacht, welche Vorteile ein deduktives Vorgehen in einem bestimmten Forschungsgebiet mit sich bringt. Wenn es geschickt gewählt ist, vermag ein einziges Experiment (oder ein experimenteller Ansatz) ein ganzes Bündel von Gesetzesaussagen zurückzuweisen, die ansonsten in der wissenschaftlichen Diskussion verbleiben und möglichen Erkenntnisfortschritt verhindern könnten. Damit soll jedoch nicht gesagt sein, dass das induktive Vorgehen, soweit es sich auf befundzentrierte Forschungsstrategien bezieht, grundsätzlich abzulehnen ist. Insbesondere bei der Generierung von Gesetzesaussagen und Hypothesen in neuen Forschungsbereichen erscheint es notwendig und unersetzlich, da neue Hypothesen häufig erst durch empirische Befunde inspiriert werden. Ab einem bestimmten Punkt der theoretischen Entwicklung in einem Gebiet kann jedoch ein deduktiver Zugang einen größeren Erkenntnisgewinn mit sich bringen und die Gefahr von voreiligen und fehlerhaften wissenschaftlichen Schlüssen reduzieren.

Literatur

Becker, W., Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision Research, 19*, 967–983.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1996). Grundlagen experimenteller Methoden. In E. Erdfelder, R. Maus-

feld, T. Meiser u. G. Rudinger (Hg.), *Handbuch Quantitative Methoden* (S. 37–46). Weinheim: Psychologie Verlags Union.

Clementz, B.A., McDowell, J.E., Zisook, S. (1994). Saccadic system functioning among schizophrenia patients and their first-degree biological relatives. *Journal of Abnormal Psychology, 103,* 277–287.

Currie, J., Ramsden, B., McArthur, C., Maruff, P. (1991). Validation of a clinical antisaccadic eye movement test in the assessment of dementia. *Archives of Neurology, 48,* 644–648.

Dorris, M.C., Paré, M., Munoz, D.P. (1997). Neuronal activity in monkey superior colliculus related to the initiation of saccadic eye movements. *Journal of Neuroscience, 17,* 8566–8579.

Duhamel, J.R., Colby, C.L., Goldberg, M.E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science, 255,* 90–92.

Everling, S., Dorris, M.C., Klein, R.M., Munoz, D.P. (1999). Role of primate superior colliculus in preparation and execution of antisaccades and prosaccades. *Journal of Neuroscience, 19,* 2740–2754.

Everling, S., Dorris, M.C., Munoz, D.P. (1998). Reflex suppression in the antisaccade task is dependent upon prestimulus neural processes. *Journal of Neurophysiology, 80,* 1584–1589.

Everling, S., Fischer, B. (1998). The antisaccade: A review of basic research and clinical studies. *Neuropsychologia, 36,* 885–899.

Everling, S., Munoz, D.P. (2000). Neuronal correlates for preparatory set associated with prosaccades and antisaccades in the primate frontal eye field. *Journal of Neuroscience, 20,* 387–400.

Fletcher, W.A., Sharpe, J.A. (1986). Saccadic eye movement dysfunction in Alzheimer's disease. *Annals of Neurology, 20,* 464–471.

Fukushima, J., Fukushima, K., Chiba, T., Tanaka, S., Yamashita, I., Kato, M. (1988). Disturbances of voluntary control of saccadic eye movements in schizophrenic patients. *Biological Psychiatry, 23,* 670–677.

Fukushima, J., Fukushima, K., Miyasaka, K., Yamashita, I. (1994). Voluntary control of saccadic eye movements in patients with frontal cortical and parkinsonian patients in comparison with that in schizophrenics. *Biological Psychiatry, 36,* 21–30.

Guitton, D., Buchtel, H.A., Douglas, R.M. (1985). Frontal lobe lesions in man cause difficulties in suppressing reflexive glances and in generating goal-directed saccades. *Experimental Brain Research, 58,* 455–472.

Hallett, P.E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research, 18,* 1279–1296.

Hallett, P.E., Adams, B.D. (1980). The predictability of saccadic latency in a novel voluntary oculomotor task. *Vision Research, 20,* 329–339.

Massen, C. (2002). Exekutive Kontrolle und sakkadische Augenbewegungen: Inhibitionsmechanismen in der Antisakkadenaufgabe. In F. Wilkening, F. Caspar, O. Güntürkün, T. Rammsayer, V. Sarris u. F. Strack (Hg.), *Psychologia Universalis, Neue Reihe, Band 29.* Lengerich: Pabst Science Publishers.

Massen, C. (im Druck). Parallel programming of exogenous and endogenous components in the antisaccade task. *Quarterly Journal of Experimental Psychology, A.*

Mokler, A., Fischer, B. (1999). The recognition and correction of involuntary prosaccades in an antisaccade task. *Experimental Brain Research, 125,* 511–516.

Popper, K.R. (1982). *Logik der Forschung* (7. Aufl.). Tübingen: Mohr.

Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3–25.

Posner, M.I., Cohen, Y. (1984). Components of visual orienting. In H. Bouma u. D. Bouwhuis (Hg.), *Attention and performance X* (S. 531–556). London: Erlbaum.

Posner, M.I., Rafal, R.D., Choate, L., Vaughn, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology, 2,* 211–228.

Rafal, R., Calabresi, C.W., Brennan, C.W., Sciolto, T.K. (1989). Saccade preparation inhibits reorienting to recently attended locations. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 673–685.

Rafal, R., Egly, R., Rhodes, D. (1994). Effects of inhibition of return on voluntary and visually guided saccades. *Canadian Journal of Experimental Psychology, 48,* 284–300.

Rosse, R.B., Schwartz, B. L., Kim, S.Y., Deutsch, S.I. (1993). Correlation between antisaccade and Wisconsin Card Sorting Test performance in schizophrenia. *American Journal of Psychiatry, 150,* 333–335.

Schlag-Rey, M., Amador, N., Sanchez, H., Schlag, J. (1997). Antisaccade performance predicted by neuronal activity in the supplementary eye field. *Nature, 390,* 398–400.

Sereno, A.B., Holzman, P.S. (1995). Antisaccades and smooth pursuit eye movements in schizophrenia. *Biological Psychiatry, 37,* 394–401.

Tien, A.Y., Pearlson, G.D., Machlin, S.R., Bylsma, F.W., Hoehn-Saric, R. (1992). Oculomotor performance in obsessive-compulsive disorder. *American Journal of Psychiatry, 149,* 641–646.

Silke Hamm

Prognose der Kurzzeitgedächtnisspanne durch verschiedene Invarianzhypothesen

Zusammenfassung

Der vorliegende Beitrag beschäftigt sich mit drei verschiedenen Invarianzhypothesen des Kurzzeitgedächtnisses für lexikalisches Material. Zwei der Invarianzhypothesen wurden innerhalb des Recall-Paradigmas formuliert (Baddeley et al., 1975; Dosher u. Ma, 1998), die dritte Hypothese stammt aus dem Sternberg-Rekognitionsparadigma (Cavanagh, 1972). Anhand von zwei Experimentreihen mit verschiedenen Stimulusmaterialien werden zuerst diese drei Hypothesen auf ihre Gültigkeit hin überprüft und systematisch mit Bezug auf das Modell von Hulme et al. (1999) miteinander verglichen. In einem weiteren Schritt wird getestet, ob diese Invarianzhypothesen des Kurzzeitgedächtnisses integrierbar sind in eine vierte Invarianzhypothese zur Schätzung von Kurz- und Langzeitgedächtnisspanne, die modifizierte Total-Time-Hypothese (Bredenkamp, 1993). Die Ergebnisse sprechen für die Gültigkeit und Überführbarkeit der Invarianzhypothesen sowie für das Modell von Hulme et al. (1999).

Invarianzhypothesen

Unter Invarianzhypothesen versteht man hypothetische Gesetzmäßigkeiten, die postulieren, dass eine Funktion verschiedener Parameter über Materialien oder Versuchsbedingungen hinweg konstant bleibt. »Invarianz« bedeutet dabei nicht, dass keinerlei Varianz auftritt, sondern dass die Unterschiede zwischen Individuen und Materialien nicht bedeutsam sind. Im Folgenden wird zuerst auf drei Invarianzhypothesen des Kurzzeitgedächtnisses eingegangen, dann auf die modifizierte Total-Time-Hypothese und ihre Ableitung.

Invarianzhypothese nach Baddeley

Innerhalb des Arbeitsgedächtnis-Modells von Baddeley (2000) werden vier Komponenten des Kurzzeitgedächtnisses postuliert: zwei modalitätsspezifische Subsysteme (Phonologische Schleife und Visuell-räumlicher Notizblock), ein multimodaler Zwischenspeicher (episodischer Speicher) und ein Koordinationssystem (Zentrale Exekutive). Das Phänomen des Wortlängeneffekts (WLE) besagt, dass mehr Items behalten werden, wenn es sich um kurze und schnell artikulierbare Stimuli handelt. Baddeley et al. (1975; Baddeley et al., 2002) erklären diesen Effekt durch folgendes Modell: Die Items werden im Phonologischen Speicher, der ein Teil der Phonologischen Schleife darstellt, verarbeitet. Dort sind die phonologischen Repräsentationen einem Zerfallsprozess mit der Zeit ausgesetzt. Die Dauer der Gedächtnisspur ist unabhängig von der Materialart und liegt bei etwa zwei Sekunden (Schweickert u. Boruff, 1986). Das Verfallen der Gedächtnisspuren kann durch den aktiven Rehearsal-Prozess der Phonologischen Schleife verhindert werden. Dieser Prozess der inneren Artikulation wird durch die äußere Artikulationsgeschwindigkeit operationalisiert. Sind Items also kurz und schnell auszusprechen, dann können innerhalb des fixen Intervalls von zwei Sekunden mehr Items vor dem Verfall bewahrt werden als bei langsam auszusprechenden Stimuli, sodass für kurze Items die Kurzzeitgedächtnisspanne (KZGS) höher ist. Die Kurzzeitgedächtnisspanne ist definiert als die Anzahl von Items, die unmittelbar nach einem Lernversuch wiedergegeben werden können. Die konstante Dauer der Gedächtnisspur setzt sich nach Baddeley formal aus dem Produkt von Kurzzeitgedächtnisspanne und Artikulationszeit pro Item zusammen. Empirisch wird dieser Zusammenhang in unterschiedlichen Untersuchungsparadigmen gefunden, beispielsweise im Sprachvergleich (Chincotta u. Underwood, 1997), im Vergleich verschiedener Materialien (Baddeley et al., 1975) oder bei verschiedenen Alterskohorten (Hasselhorn, 1988). Allerdings verschwindet der Zusammenhang zwischen Artikulationszeit und Kurzzeitgedächtnisspanne, wenn beispielsweise die phonologische Ähnlichkeit (Conrad, 1964), die Wortkomplexität (Service, 1998), die Konkretheit (Walker u. Hulme, 1999) oder der Assoziationsgehalt (Lovatt et al., 2000) variiert werden.

Invarianzhypothese nach Dosher und Ma

Der Nachteil der Erklärung zum Wortlängeneffekt von Baddeley et al. (1975) ist die alleinige Berücksichtigung von Vergessen mit der Zeit, nicht aber auch von Interferenzprozessen während des Abrufs. Die Relevanz von

Interferenzprozessen kristallisierte sich beispielsweise in der Untersuchung von Cowan et al. (1992) mit heterogenen Lernlisten heraus. Die Erstellung von heterogenen Lernlisten, die zum Teil aus kurz, und zum Teil aus lang auszusprechenden Worten bestehen, ermöglicht den direkten Test der Interferenztheorie gegen die zeitliche Verfallstheorie. Dosher und Ma (1998) vertreten einen Ansatz, in den sowohl Vergessen mit der Zeit als auch Interferenz eingehen. Angenommen wird, dass Interferenzprozesse während des Abrufs wirken und durch Zeitmessungen der Wiedergabe-Zeiten während des Abrufs in einem Gedächtnisexperiment adäquat erfasst werden können. Als Prädiktor der Kurzzeitgedächtnisspanne wird also statt der Artikulationszeit die Reproduktionszeit pro Item verwendet. Empirisch wird von einer materialunabhängigen, konstanten Gesamtwiedergabe-Zeit berichtet, die sich aus dem Produkt von Kurzzeitgedächtnisspanne und Reproduktionszeit pro Item ergibt (Dosher u. Ma, 1998). Diese postulierte Invarianz konnte für englische (Dosher u. Ma, 1998) und deutsche Probanden (Pbn) unter verschiedenen Untersuchungsbedingungen nachgewiesen werden, doch zeigte sich keine Invarianz über verschiedene Sprachgruppen (Hamm, 2001). Als Vorteile der Invarianzhypothese nach Dosher und Ma (1998) sind folgende Punkte zu benennen:

- Theoretisch scheint sie eher adäquat, weil sowohl Vergessen mit der Zeit als auch Vergessen aufgrund von Interferenz berücksichtigt wird.
- Der empirische Gehalt der Invarianzhypothese ist im Vergleich zur Invarianzhypothese nach Baddeley et al. (1975) größer, weil sie sich auf einen größeren Anwendungsbereich bezieht und somit leichter falsifiziert werden kann. Beispielsweise wird der Effekt Phonologischer Ähnlichkeit (EPÄ) bei Baddeley (2000) unabhängig vom Wortlängeneffekt durch Interferenz im Phonologischen Speicher erklärt. Nach Dosher und Ma (1998) könnte der Effekt Phonologischer Ähnlichkeit durch verlangsamte Reproduktionszeiten wegen vermehrter Interferenz interpretiert werden. Dies wurde empirisch bestätigt (Hamm, 2001).
- Methodisch ist die Erfassung der Reproduktionszeit als Prädiktor der Kurzzeitgedächtnisspanne sparsamer, da kein zusätzliches Experiment zur Erfassung der Artikulationszeit durchgeführt werden muss. Zwar kann auch die Artikulationszeit in einem Gedächtnisexperiment erfasst werden, doch belegen Studien (Cowan, 1994; Hulme et al., 1999), dass die online und die offline erfasste Artikulationszeit erheblich voneinander abweichen.

Invarianzhypothese nach Cavanagh

Die Invarianzhypothese nach Cavanagh (1972) bezieht sich auf das Sternberg-Rekognitionsparadigma (Sternberg, 1966). Vorgegeben werden Ziffern- oder Buchstabenreihen verschiedener Länge, danach folgt ein Testitem, das möglichst schnell dahingehend beurteilt werden soll, ob es Element der vorigen Reihe war oder nicht. Nach der Size-Hypothese wird dieser Vergleichsprozess folgendermaßen beschrieben: Das Testitem wird Merkmal für Merkmal mit den Items der präsentierten Liste verglichen. Die Reaktionszeit der Antwort steigt linear mit der Listenlänge. Für die Annahme einer exhaustiven und gegen eine selbstterminierende Suche spricht, dass die Regressionsgeraden mit der unabhängigen Variablen Listenlänge und der abhängigen Variablen Reaktionszeit bei positivem Set (Testitem war in der Liste enthalten) und negativem Set (Testitem war nicht enthalten) parallel verlaufen. Angenommen wird eine Konstante C, die sich aus der Multiplikation von Kurzzeitgedächtnisspanne und Suchzeit pro Item ergibt. Ist die Suchzeit für ein Material sehr kurz, so ist die Kurzzeitgedächtnisspanne hoch. Empirisch wird dieser Zusammenhang teils gestützt (Puckett u. Kausler, 1984; Bredenkamp, 2004), teils wird er nicht gefunden (Hulme et al., 1999).

Zusammenhang zwischen den drei Invarianzhypothesen des Kurzzeitgedächtnisses

Bisher stehen die drei Invarianzhypothesen relativ unverbunden nebeneinander. Der Frage nach Integration kann durch Anwendung des Modells von Hulme et al. (1999) nachgegangen werden. Die Response-Zeit in einem Gedächtnisexperiment wird konzeptuell in drei Komponenten zerlegt: Präparationsintervall, Artikulationszeit und Pausenzeiten. Als Präparationsintervall wird die Zeit zwischen Aufforderung zur Wiedergabe und tatsächlichem Sprechbeginn definiert. Nach diesem Modell wird im Präparationsintervall ein Selektionsprozess der Gedächtnisspuren angenommen, der mit den Sternberg-Suchzeiten korrelieren soll, während in den Pausenzeiten halb verfallene Gedächtnisspuren rekonstruiert werden können. Demnach müssten die Prozesse im Präparationsintervall im Zusammenhang zur Cavanagh-Hypothese stehen, die Prozesse während der Sprechzeit mit der Baddeley-Hypothese (wobei bereits auf Unterschiede zwischen offline und online erfasster Zeit hingewiesen wurde) und die dritte Variable, die Gesamtwiedergabezeit nach Dosher und Ma (1998), sollte sich aus Prozes-

sen während Präparationsintervall, Artikulationszeit und Pausenzeit zusammensetzen.

Nach der Beschreibung der drei Invarianzhypothesen des Kurzzeitgedächtnisses und des Hulme-Modells wird im Folgenden auf eine vierte Invarianzhypothese, die modifizierte Total-Time-Hypothese (mTTH), und deren Ableitung eingegangen.

Die modifizierte Total-Time-Hypothese

Bredenkamp (1975; 1993) verknüpfte zwei weitere Invarianzhypothesen miteinander: die Total-Time-Hypothese (TTH, Cooper u. Pantle, 1967) und die Hypothese einer konstanten Langzeitgedächtnisspanne von Nevelsky (1970). Die Total-Time-Hypothese besagt, dass die Gesamtlernzeit für ein bestimmtes Lernmaterial eine Konstante ist, unabhängig von der zeitlichen Segmentierung der Lernversuche. Werden beispielsweise zum Bewältigen eines Materials 60 Sekunden benötigt, so macht es keinen Unterschied, ob sechsmal zehn Sekunden oder eine Minute lang gelernt wird. Formal besteht die Gesamtlernzeit aus dem Produkt der Dauer der Lernversuche und der Anzahl der Lernversuche. Kintsch (1970) berichtet über empirische Evidenz für diesen Zusammenhang. Nevelsky (1970) postuliert eine konstante Langzeitgedächtnisspanne, wenn der Gesamtinformationsgehalt die Kapazität des Kurzzeitgedächtnisses überschreitet: Die in einem Lerndurchgang in Abhängigkeit von der Darbietungszeit aufgenommene Information ist demnach eine Konstante und errechnet sich aus dem Quotienten der Gesamtinformationsmenge für das Material und der Anzahl benötigter Lernversuche. Nevelsky (1970) berichtet über empirische Evidenz. Aus der Verknüpfung zwischen Total-Time-Hypothese und Nevelskys Hypothese wird abgeleitet, dass die Menge aufgenommener Information linear mit der Darbietungszeit steigt (Bredenkamp, 1975). Diese Annahme kann weder theoretischen Überlegungen noch empirischer Evidenz standhalten, sodass Bredenkamp (1975) die modifizierte Total-Time-Hypothese formulierte: Die Anzahl der Lernversuche für eine bestimmte Darbietungszeit und ein bestimmtes Material, dessen Gesamtinformationsgehalt wiederum die Kapazität des Kurzzeitgedächtnisses übertrifft, ergibt sich aus dem Produkt der Konstanten a und dem Quotienten aus der Konstanten b und der Darbietungszeit (Bredenkamp, 1975). Dabei entspricht a der minimal erforderlichen Anzahl von Lernversuchen, um das Lernkriterium im Kurzzeitgedächtnis zu erreichen. Der Quotient aus der zweiten Konstanten b und der Darbietungszeit beschreibt die Anzahl der Lernversuche, die zu-

sätzlich zum Erreichen des Kriteriums im Kurzzeitgedächtnis benötigt werden, um den Gesamtinformationsgehalt zu verarbeiten. Von der Dauer der Darbietungszeit unbeeinflusst bleiben a und b konstant, variieren jedoch zwischen verschiedenen Materialien. Durch die modifizierte Total-Time-Hypothese wird ein hyperbolischer Zuwachs der aufgenommenen Information mit der Darbietungszeit vorausgesagt. Die Äquivalenzhypothese besagt, dass die Asymptote, der sich diese Funktion annähert, als Kurzzeitgedächtnisspanne in bit interpretiert werden kann. Diese Annahme bietet die Möglichkeit zur Integration der bereits beschriebenen Invarianzhypothesen des Kurzzeitgedächtnisses in die modifizierte Total-Time-Hypothese, wodurch der empirische Gehalt erhöht wird. Zusätzlich können strengere Prüfkriterien an das Modell angelegt werden (Bredenkamp, 1993). Die Gültigkeit der modifizierten Total-Time-Hypothese und der Äquivalenzhypothese belegen zahlreiche Ergebnisse unter verschiedenen Versuchsbedingungen bei sinnfreien Konsonant-Vokal-Verbindungen als Stimulusmaterial (Bredenkamp u. Klein, 1998; Bredenkamp u. Hamm, 2001), ebenso bei lexikalischem Material (Hamm, 2001). Allerdings zeigte sich für Zahlenmaterial trotz Gültigkeit der modifizierten Total-Time-Hypothese, dass die Asymptote der Lernfunktion die Kurzzeitgedächtnisspanne signifikant überschätzt. Dies wird mit der spezifischen Beschaffenheit des Zahlenmaterials, insbesondere mit den Möglichkeiten paralleler visueller und phonologischer Kodierung und effektiven Chunkingprozessen begründet (Hamm, 2001). Zudem können durch die modifizierte Total-Time-Hypothese zusätzlich die Lernrate (Bredenkamp u. Klein, 1998) und die Langzeitgedächtnisspanne in bit (Bredenkamp, 1993) geschätzt werden, die allerdings nicht Gegenstand dieses Beitrags sind.

Im Folgenden wird die Gültigkeit der Invarianzhypothesen bei lexikalischem Material überprüft, wobei alle drei Invarianzhypothesen systematisch in Hinblick auf das Modell von Hulme et al. (1999) verglichen werden. Die zweite Frage zielt auf die Überführbarkeit dieser Invarianzhypothesen in die modifizierte Total-Time-Hypothese ab.

Überblick über die Experimente

Jede Versuchsperson (Vp) führte drei Experimente durch (Artikulationszeit-Experiment, Gedächtnisexperiment und Sternberg-Experiment), die im Folgenden berichtet werden. Die Teilnahme erfolgte freiwillig und in Einzelsitzungen, die ungefähr eine Stunde dauerten. Als Erstes wurde im-

mer ein Artikulationszeit-Experiment vorgegeben, die Reihenfolge der beiden anderen Experimente wurde zwischen den Versuchspersonen randomisiert. Analysen deckten keinen Reihenfolge-Effekt auf. Es fungierte nur eine Person als Versuchsleiter.

Material

In Experiment 1 wurden phonologisch unähnliche Worte als Stimulusmaterial verwendet. Zur Vermeidung von Konfundierungen mit graphemischer Komplexität oder Wortkomplexität und zur Maximierung der Ähnlichkeit zur zweiten Experimentreihe wurden vierbuchstabige Substantive gewählt. Die Frequenz der Worte beträgt im Durchschnitt 177 Nennungen pro 1 Million Worte (Baayen et al., 1995). Bildhaftigkeit (5.45), Bedeutungsgehalt (4.42) und Konkretheit (5.19), beurteilt auf einer siebenstufigen Skala, wurden zur Parallelisierung mit dem Zahlenmaterial im Durchschnitt möglichst hoch gehalten. Folgende zehn Wörter wurden verwendet (Hamm, 2001):

BANK ZIEL LUFT BAUM CHEF KIND PLAN WERT FORM BUCH

Experiment 2 wurde mit Zahlenmaterial durchgeführt. Wegen phonologischer Ähnlichkeit zur ZWEI wurde die DREI ausgeschlossen. Die SIEBEN wurde durch die NULL ersetzt, da das Stimulusmaterial homogen und einsilbig sein sollte. Folgende acht Stimuli wurden verwendet:

0 1 2 4 5 6 8 9

Probanden

An Experiment 1 und 2 nahmen je 50 Probanden teil. Dabei handelte es sich meist um Psychologiestudenten, die für ihre Teilnahme eine Versuchspersonenbescheinigung oder 4 € erhielten. Die Teilnahme an einer Experimentreihe ist Ausschlusskriterium für die Teilnahme an der zweiten Experimentreihe. Der Altersdurchschnitt lag bei 24.3 Jahren, 78 Prozent der Teilnehmer waren Frauen. In der zweiten Experimentreihe brachen zwei Versuchspersonen das Experiment ab, sodass hier nur 48 gültige Daten vorliegen.

Artikulationszeit-Experiment

Aufgabe des Probanden ist, die präsentierten Reihen so schnell wie möglich laut vorzulesen. Auf einem 17-Zoll-Computerbildschirm eines Pentium 100 wird nach einem visuellen ACHTUNG-Signal (weiße Schrift, schwarzer Hintergrund) und nach Vorgabe dreier Töne (400 Hz, Dauer: 200 ms, Interstimulusintervall: 800 ms) eine Sequenz variabler Länge präsentiert. Diese erscheint mit schwarzer Schrift in einem weißen Rahmen in der Mitte des Bildschirms, zentriert mit einer Höhe von 0.6 Zentimetern pro Item. Insgesamt sind 18 Sequenzen vorzulesen. Auf eine Auswertung der ersten drei Beispiele wird verzichtet, sodass je fünf Reihen mit sechs, acht und zehn Items in die Analysen eingehen. Die Erstellung einer Sequenz erfolgt nach dem Randomisierungsprinzip *mit* Zurücklegen. Als Restriktion wird das Verbot von mehr als zweimaliger Wiederholung eines Items hintereinander eingeführt. Nach Beenden des Vorlesens, das durch Tastendruck des Versuchsleiters markiert wird, wird der Bildschirm schwarz. Dann erscheint die Frage »Wiedergabe korrekt j/n?«. Trat beim Vorlesen ein Versprecher oder eine Verzögerung, beispielsweise wegen Räusperns, auf, wird »n« eingegeben, das heißt, diese Sequenz wird nicht ausgewertet, sondern eine neue Reihe gleicher Länge vorgegeben. Bei korrekter Wiedergabe (»j«), wird die Zeit gespeichert. Zwei Sekunden nach Eingabe von j oder n erscheint wiederum das ACHTUNG-Signal, gefolgt von drei Tönen: Ein neuer Versuchsdurchgang beginnt. Über eine computerinterne Uhr wird die Zeit zwischen Darbietung der Reihe und dem Tastendruck des Versuchsleiters als Signal des Vorleseendes erfasst. Die Artikulationszeit pro Item errechnet sich aus der Regression mit der unabhängigen Variablen Listenlänge und der abhängigen Variablen der Vorlesezeit.

Kurzzeitgedächtnis-Experiment

Der Proband soll sich Sequenzen variabler Länge merken und nach einer Pause von zwei Sekunden vollständig und in der richtigen Reihenfolge wiedergeben. Nach einem ACHTUNG-Signal (weiße Schrift, schwarzer Hintergrund, Dauer: 2.5 Sekunden) werden die Items in einem mittig präsentierten weißen Balken sukzessive von links nach rechts dargeboten, je um 0.5 Zentimeter verschoben (schwarze Schrift, Höhe 0.6 Zentimeter, Dauer: 800 ms, Interstimulusintervall: 200 Sekunden). Auf das letzte Item folgt eine Pause von zwei Sekunden, in der nur der weiße Balken zu sehen ist. Dann fordert ein auf der rechten Seite des Balkens gezeigtes WIEDERGA-

BE-Signal zum Recall auf. Das Ende der Wiedergabe wird durch den Versuchsleiter per Tastendruck markiert, ebenso die Korrektheit der Antwort. Nach Eingabe des Versuchsleiters erscheint wiederum das ACHTUNG-Signal und dann die nächste Reihe. Wurde die Sequenz richtig wiedergegeben, so ist die folgende Liste um ein Item länger, bei fehlerhafter Reproduktion verkürzt sich die Reihe um ein Item. Die Stimuli werden zufällig *mit* Zurücklegen aus dem Itempool gezogen. Als Restriktion gilt das Verbot von mehr als zwei Wiederholungen eines Items nacheinander und das Verbot von auf- oder absteigenden Dreierreihen bei Zahlenmaterial. Das Experiment ist nach dem zehnten Wechsel zwischen richtig/falschen respektive falschen/richtigen Antworten beendet. Die Kurzzeitgedächtnisspanne ist definiert durch den Mittelwert dieser Wendepunkte. Zusätzlich wird die Wiedergabezeit gemessen, die das Zeitintervall zwischen Präsentation des WIEDERGABE-Signals und Beendigung des Recalls umfasst. Die Reproduktionszeit pro Item eines Probanden ergibt sich aus der Wiedergabezeit aller richtig reproduzierten Sequenzen, relativiert auf die Itemlänge.

Sternberg-Experiment

Aufgabe des Probanden ist, so schnell wie möglich zu entscheiden, ob ein Testreiz in einer vorher gezeigten Liste enthalten war oder nicht. Zuerst wird ein Ton (400 Hz, Dauer: 400 ms) und dann ein ACHTUNG-Signal (weiße Schrift, schwarzer Hintergrund, Dauer: 1 Sekunde) präsentiert, dann folgt die erste Sequenz bestehend aus zwei, drei oder vier Items (weiße Schrift, schwarzer Hintergrund). Die Auswahl der Stimuli erfolgt nach dem Randomisierungsprinzip *ohne* Zurücklegen. Die Items werden nacheinander um 0.5 Zentimeter versetzt dargeboten (Dauer: 800 ms, Interstimulusintervall: 200 ms). Nach einer Pause von zwei Sekunden, in der der Bildschirm schwarz ist, und nach einem Ton (400 Hz, Dauer: 400 ms) erscheint mittig der Testreiz. Handelt es sich um ein altes Item, soll möglichst schnell die Taste »ä« gedrückt werden, bei einem neuen Reiz die Taste »a«. Die Taste für »ä« liegt auf der Tastatur rechts, die Taste für »a« weit links, sodass eine gute Unterscheidbarkeit gewährleistet ist. Um die Gedächtnisanforderungen gering zu halten, wurde über die Tasten ein Hinweisschild »alt« oder »neu« geklebt. Zudem wurden zwei Übungsdurchgänge absolviert und auf eine Auswertung der ersten drei Versuchsdurchgänge verzichtet. Die Zeit zwischen Präsentation des Testitems und Tastendruck wird computergestützt erfasst. Jeweils in 50 Prozent der Fälle pro Listenlänge ist das Testitem vorher präsentiert worden, in den anderen 50 Prozent dagegen handelt es

Tabelle 1: Aufbau des Sternberg-Experiments: Anzahl der Versuchsdurchgänge pro Probanden

Item	Memory Load 2	3	4
... alt (positives Set)	10	10	10
... neu (negatives Set)	10	10	10

sich um ein neues Item. Durch die intraindividuelle Variation von Memory Load (Listenlänge) und Set (positiv: Testreiz war enthalten in der Reihe, negativ: Testreiz ist neu) absolviert jede Versuchsperson 60 Durchgänge, wie in Tabelle 1 veranschaulicht. Unter »Reaktionszeit« wird im Folgenden die Latenzzeit der Antwort verstanden, die ein Proband benötigt, um ein Item als »alt« oder »neu« zu klassifizieren. Mit »Suchzeit« ist jeweils der Anstieg der Regressionsgeraden gemeint, mit der Prädiktorvariablen Listenlänge und der Kriteriumsvariablen Reaktionszeit.

Ergebnisse

Im Folgenden werden zuerst die deskriptiven Ergebnisse und die Auswertungen zu den drei Invarianzhypothesen des Kurzzeitgedächtnisses vorgestellt. Dann folgt ein Vergleich dieser drei Invarianzhypothesen nach dem Modell von Hulme et al. (1999) sowie die Verknüpfung mit der modifizierten Total-Time-Hypothese.

Deskriptive Ergebnisse

Für jeden Probanden wurden einzeln Artikulationszeit, Reproduktionszeit und Suchzeit pro Item errechnet. Für jede dieser Analysen wurde nach dem Tukey-Kriterium eine Ausreißeranalyse vorgenommen. Als Ausreißer definiert wurden alle Werte, die größer als Q3 + 1.5 * (Q3–Q1) oder kleiner als Q3–1.5 * (Q3–Q1) sind, wobei Q3 das dritte und Q1 das erste Quartil bezeichnet (Tukey, 1977). Für die Suchzeiten nach Cavanagh ergab sich das

Tabelle 2: Deskriptive Ergebnisse von Experiment 1 (Wortmaterial)

Wortmaterial	Artikulationszeit	Reproduktionszeit	Suchzeit	KZGS
Mittelwert	412.58	718.86	71.58	4.45
Varianz	5488	21536	3352	0.30

Tabelle 3: Deskriptive Ergebnisse von Experiment 2 (Zahlenmaterial)

Zahlenmaterial	Artikulationszeit	Reproduktionszeit	Suchzeit	KZGS
Mittelwert	308.51	547.73	61.46	6.16
Varianz	3623	12329	2471	0.53

Problem, dass bei fünf Pbn negative Anstiege der Regressionsgeraden gefunden wurden, das heißt, die Reaktionszeit bei einer Sequenzlänge von vier Items war schneller als bei zwei Items. Diese Werte wurden aus der Analyse ausgeschlossen. In Tabelle 2 sind für das Wortmaterial die deskriptiven Werte und die Varianz abgetragen, in Tabelle 3 für das Zahlenmaterial. Die Einheit der Zeitmessungen ist Millisekunden, die Kurzzeitgedächtnisspanne ist in Anzahl der Items angegeben.

In den folgenden Auswertungen werden Experiment 1 und 2 gemeinsam berichtet.

Analyse zur Invarianzhypothese von Baddeley

Als Erstes wird auf die *Invarianzhypothese nach Baddeley* eingegangen. Eine einfaktorielle Varianzanalyse mit der unabhängigen Variablen Materialart und der abhängigen Variablen Artikulationszeit – unter gegebener Varianzhomogenität ($t(1,96) = 1.72$) – zeigt, dass die Artikulationszeit zwischen den beiden Experimentreihen bedeutsam variiert: $F(1,96) = 57.98, p < .05$. Die Aussprache für das Zahlenmaterial ist signifikant schneller. Eine zweite einfaktorielle Varianzanalyse mit dem Materialtyp als unabhängiger Variablen und der Kurzzeitgedächtnisspanne als abhängiger Variablen deckt zwischen den Materialarten signifikante Unterschiede auf: $F(1,96) = 173.66, p < .05$. Varianzhomogenität ist gegeben ($t(1,96) = 1.74$). Die schnellere Artikulationsgeschwindigkeit einer Materialart geht mit einer höheren Kurzzeitgedächtnisspanne einher. Wenn diese Invarianzhypothese für alle Individuen gilt und die spezifischen Ausprägungen der Dauer der Gedächtnisspur, Artikulationszeit oder Kurzzeitgedächtnisspanne variieren können, dann bezieht sich das Postulat einer Invarianz nicht auf die arithmetischen Mittel, sondern auf die Relation der geometrischen Mittel (Bredenkamp, 1993). Die Analyse zur Überprüfung der Invarianzhypothese von Baddeley bezieht sich im Materialvergleich also auf die Summe der logarithmierten Artikulationszeiten und logarithmierten KZGS-Werte als abhängige Variable. Es werden keine bedeutsamen Unterschiede zwischen den Experimentreihen aufgedeckt: $F(1,96) = 0.74$. Die Power beträgt .97 bei $\alpha =$

0.05 und mittlerer Effektstärke nach Cohen $f^2 = 0.15$ (Erdfelder et al., 1996). Deskriptiv liegt die Dauer der Gedächtnisspur für Wortmaterial bei 1.83 Sekunden und bei Zahlenmaterial bei 1.89 Sekunden. Diese Werte liegen innerhalb des von Schweickert und Boruff (1986) angegebenen Intervalls und stimmen gut mit den Daten von Hamm (2001) überein, die mit identischer Versuchsapparatur und Stimuli an einer anderen Studentenkohorte erhoben wurden (Zahlenmaterial: 1.85 Sekunden; unähnliche Worte: 1.86 Sekunden).

Analyse zur Invarianzhypothese nach Dosher und Ma

Die *Invarianzhypothese nach Dosher und Ma* (1998) wird durch zwei Analysen überprüft. Nachdem sich bedeutsame Unterschiede in Abhängigkeit der Materialart für die Kurzzeitgedächtnisspanne gezeigt haben, deckt eine weitere Varianzanalyse auch signifikante Unterschiede in der Reproduktionszeit für die beiden Materialien auf: $F(1,96) = 42.03, p < .05$. Varianzhomogenität ist gegeben ($t(1,96) = 3.50, p > .05$). Die schnellere Reproduktionszeit einer Materialart geht mit einer größeren Kurzzeitgedächtnisspanne einher. Die zweite Analyse prüft die Konstanz der Gesamtwiedergabezeit, operationalisiert durch die Summe der logarithmierten Reproduktionszeit und der logarithmierten Kurzzeitgedächtnisspanne (Bredenkamp, 1993). Zwischen den beiden Materialarten werden in einer Varianzanalyse – bei gegebener Varianzhomogenität ($t(1,96) = 1.10$) – keine bedeutsamen Unterschiede aufgedeckt: $F(1,96) = 1.66, \alpha = 0.05, \beta = 0.03, f^2 = 0.15$ (Erdfelder et al., 1996). Empirisch ergibt sich für die Gesamtwiedergabe-Zeit bei Wortmaterial ein Wert von 3.16 Sekunden und bei Zahlenmaterial ein Wert von 3.36 Sekunden. Diese Daten sind etwas niedriger als das von Dosher und Ma (1998) angegebene Intervall, entsprechen aber ziemlich genau den Angaben einer vorherigen Untersuchung (Hamm, 2001): 3.28 Sekunden für unähnliche Worte und 3.37 Sekunden für numerisches Stimulusmaterial.

Analyse zur Invarianzhypothese von Cavanagh

Zur Prüfung der *Cavanagh-Hypothese* werden drei Analysen vorgenommen: Zuerst werden für jedes Stimulusmaterial getrennt die Reaktionszeiten in Abhängigkeit von der Listenlänge auf einen linearen und quadratischen Trend geprüft, dann werden die Cavanagh-Konstanten innerhalb eines Experiments in Abhängigkeit des Sets verglichen. Die Überprüfung

zur Abhängigkeit vom Set ergibt sich aus der Annahme einer exhaustiven Durchmusterung der Reihe, unabhängig davon, ob das Item enthalten ist (positives Set) oder nicht (negatives Set). Zuletzt werden die Cavanagh-Konstanten der beiden Materialarten gegenübergestellt.

Für beide Stimulusarten und beide Sets konnte durchgängig gezeigt werden, dass mit steigendem Memory Load die Reaktionszeit linear und nicht quadratisch steigt. Die Werte des Beta-Fehlers wurden jeweils mittels GPOWER errechnet (Erdfelder et al., 1996).

- Experiment 1, positives Set: linearer Trend $F(1,43) = 73.64, p < .05$ quadratischer Trend $F(1,43) = 0.50$, n. s.; $\alpha = 0.05$, $\beta = 0.24$, $f^2 = 0.15$
- Experiment 1, negatives Set: linearer Trend $F(1,43) = 44.93, p < .05$ quadratischer Trend $F(1,43) = 0.41$, n. s.; $\alpha = 0.05$, $\beta = 0.24$, $f^2 = 0.15$
- Experiment 2, positives Set: linearer Trend $F(1,44) = 63.54, p < .05$ quadratischer Trend $F(1,44) = 3.98$, n. s., $\alpha = 0.05$, $\beta = 0.25$, $f^2 = 0.15$
- Experiment 2, negatives Set: linearer Trend $F(1,44) = 68.05, p < .05$ quadratischer Trend $F(1,44) = 1.67$, n. s.; $\alpha = 0.05$, $\beta = 0.25$, $f^2 = 0.15$

Weder in Experiment 1 mit Wortmaterial noch in Experiment 2 mit Zahlen als Stimuli fand sich zwischen den Cavanagh-Konstanten der beiden Sets ein bedeutsamer Unterschied. Um die Annahme der Nullhypothese zu erschweren, wird $\alpha = 0.10$ gesetzt. Experiment 1: $F(1,39) = 0.63$, $\alpha = 0.10$, $\beta = 0.22$, $f^2 = 0.15$, Experiment 2: $F(1,43) = 1.45$, $\alpha = 0.10$, $\beta = 0.19$, $f^2 = 0.15$ (Erdfelder et al., 1996). Wegen des Fehlens bedeutsamer Unterschiede wird im Folgenden unter Suchzeit nur noch die Suchzeit bei positivem Set verstanden. Eine einfaktorielle Varianzanalyse mit der abhängigen Variablen »Summe aus logarithmierter Suchzeit und logarithmierter KZGS« deckt zwischen den Experimenten keine bedeutsamen Unterschiede auf, Varianzhomogenität ist gegeben ($t(1,82) = 0.044$): $F(1,82) = 0.048$ bei $\alpha = 0.10$, $\beta = 0.03$, $f^2 = 0.15$. Deskriptiv liegt die Cavanagh-Konstante in Experiment 1 bei 384 ms und in Experiment 2 bei 382 ms. Diese Werte stimmen sehr gut mit den von Klein (1995) ermittelten Werten überein: abstrakte Worte: 324 ms, Zahlen: 382 ms.

Vergleich der drei Invarianzhypothesen

Im Folgenden werden erst die Korrelationen zwischen Artikulationszeit, Reproduktionszeit, Suchzeit und Kurzzeitgedächtnisspanne bestimmt. Dann werden die drei Zeitmessungen hinsichtlich des Modells von Hulme et al. (1999) miteinander verglichen.

Tabelle 4: Ergebnisse der Korrelationsschätzungen über Experiment 1 und 2

	Artikulationszeit	Reproduktionszeit	Suchzeit
KZGS	–.592**	–.581**	–.204*
Artikulationszeit		.412**	.004
Reproduktionszeit			.179*

In Tabelle 4 werden die Korrelationen nach Pearson der verschiedenen Parameter dargestellt. Insgesamt gehen 93 gültige Daten in die Auswertung ein. »**« bedeutet, dass die einseitig getestete Korrelation bei $p < .01$ signifikant ist, »*« bei $p < .05$ bei einseitigen Tests.

Nach dem Korrelations-Differenzen-Test von McNemar (1962) ergeben sich folgende Unterschiede:

- Der Korrelationsvergleich zwischen Artikulationszeit – KZGS und Reproduktionszeit – KZGS deckt keine bedeutsamen Unterschiede auf: $t(93) = 0.05$.
- Die Korrelation zwischen Artikulationszeit – KZGS ist wesentlich höher als die Korrelation zwischen Suchzeit – KZGS mit $t(93) = -3.34$ für $\alpha = 0.05$.
- Die Korrelation zwischen Reproduktionszeit – KZGS ist ebenfalls signifikant höher als die Korrelation Suchzeit – KZGS, $t(93) = 3.56$ bei $\alpha = 0.05$.

Nach dem Modell von Hulme et al. (1999) bestehen systematische Beziehungen zwischen den Prädiktoren: Artikulationszeit und Suchzeit sollen unabhängig voneinander sein, während die Reproduktionszeit sowohl die Suchzeit als auch die Artikulationszeit enthalten soll. Für die Unabhängigkeit von Artikulationszeit und Suchzeit spricht die Nullkorrelation, dargestellt in Tabelle 4. Zur Prüfung der Bedeutung der Reproduktionszeit wird aus der Korrelation zwischen Reproduktionszeit und Kurzzeitgedächtnisspanne die Artikulationszeit auspartialisiert. Der partielle Korrelationskoeffizient beträgt $-.46$, $p < 0.01$. Der Korrelationskoeffizient zwischen Reproduktionszeit und Kurzzeitgedächtnisspanne sinkt auf $r = -.56$, $p < 0.01$, wenn die Suchzeit auspartialisiert wird. Werden Artikulationszeit und Suchzeit gleichzeitig aus der Reproduktionszeit auspartialisiert, so bleibt dennoch eine signifikante Korrelation zwischen Reproduktionszeit und Kurzzeitgedächtnisspanne bestehen ($r = -.44$, $p < 0.01$). Dies wird nach dem Modell von Hulme et al. (1999) den Prozessen zugeschrieben, die während der Pausenzeiten ablaufen. Wird die Kurzzeitgedächtnisspanne in einer Regressionsanalyse durch die Prädiktoren Artikulationsrate, Reproduktionsrate und Suchrate vorhergesagt, so findet man ein r^2 von .46. »Rate«

bedeutet jeweils die reziproke Zeitmessung. Bei Verwendung allein von Artikulationsrate und Suchrate als Prädiktoren sinkt nach der Semipartialkorrelation höherer Ordnung r^2 signifikant auf .34, wie der F-Test zeigt: $F(2,68) = 7.56$, $\alpha = 0.05$, $\beta = .17$, $f = 0.15$ (Erdfelder et al., 1996).

Integration der drei Invarianzhypothesen in die modifizierte Total-Time-Hypothese für Wortmaterial

Bevor auf die Integration der drei bereits beschriebenen Invarianzhypothesen in die modifizierte Total-Time-Hypothese eingegangen werden kann, wird kurz das an anderer Stelle beschriebene Langzeitgedächtnis-Experiment erläutert, aus dem die empirischen Daten zur Überprüfung der modifizierten Total-Time-Hypothese entnommen wurden (Hamm, 2001).

Aufgabe in dem Langzeitgedächtnis-Experiment ist, sich eine Reihe bestehend aus zehn Items zu merken und vollständig in der richtigen Reihenfolge wiederzugeben. Die Items werden zentriert und en bloc für 10, 20 oder 30 Sekunden in schwarzer Schrift in einem mittig gezeigten, weißen Balken präsentiert. Direkt nach Ablauf der Darbietungszeit erscheint auf der rechten Seite des Balkens das Wiedergabe-Signal. Nach der Reproduktion entscheidet der Versuchsleiter per Tastendruck, ob die Antwort richtig ist oder nicht. Jeder Proband erhält nur eine einzige Reihe, die Zuordnung zu der Darbietungszeit erfolgt randomisiert. Der Versuch ist beendet, wenn die Sequenz zweimal hintereinander richtig wiedergeben werden konnte (Hamm, 2001).

In den folgenden Auswertungen werden aus den drei Konstanten nach Baddeley, Dosher und Ma sowie Cavanagh die Höhe der Kurzzeitgedächtnisspanne prognostiziert und in bit umgerechnet. Dann folgt eine Integration dieser KZGS-Werte in die modifizierte Total-Time-Hypothese durch Einsetzen der prognostizierten KZGS-Werte. Für die Konstante b bei Wortmaterial ergibt sich ein Wert von $b = 25.75$ (Hamm, 2001). Der Itempool umfasst zehn Items, der Informationsgehalt pro Item beträgt 3.32 bit. Bei einer Sequenzlänge von zehn Items resultiert also der Gesamtinformationsgehalt von 33.20 bit (Hamm, 2001). Die prognostizierte Anzahl der Lernversuche für die Darbietungszeit von 10, 20 und 30 Sekunden mit dem Gesamtinformationsgehalt von 33.20 bit und den durch die Invarianzhypothesen prognostizierten Kurzzeitgedächtnis-Werten wird mit der empirischen Anzahl von Lernversuchen durch einen F-Test verglichen. Die empirischen und prognostizierten Werte sind in Tabelle 5 aufgelistet. Nach Baddeley ergibt sich eine geschätzte Höhe der Kurzzeitgedächtnisspanne von 15.26 bit, nach Dosher und Ma von 15.06 bit und nach Cavanagh von 17.39 bit.

Tabelle 5: Empirische und prognostizierte Kurzzeitgedächtnisspannen-Werte in bit zur Integration der Invarianzhypothesen

		Prognose durch die Invarianzhypothese von ...		
Darbietungszeit	empirisch	Baddeley	Dosher u. Ma	Cavanagh
10 Sekunden	3.95	4.75	4.78	4.49
20 Sekunden	3.37	3.46	3.49	3.20
30 Sekunden	1.95	3.04	3.06	2.77

Die Fehlervarianz zwischen den Darbietungszeiten ist $s_e^2 = 6.60$ (Hamm, 2001). Die Abweichungsvarianz zwischen dem empirischen und den nach Baddeley prognostizierten Werten beträgt $s_{emp}^2 = 17.48$. Der empirische F-Wert ist $F(2,54) = 2.65$. Der kritische F-Wert liegt bei einer mittleren Effektstärke, $f^2 = 0.15$ und $\alpha = 0.05$ bei $F(2,54) = 3.17$ (Erdfelder et al., 1996). Für die Prognose der Asymptote der Lernfunktion durch die Hypothese von Dosher und Ma ergibt sich eine Abweichungsvarianz von $s_{emp}^2 = 18.34$, mit einem F-Wert von $F(2,54) = 2.78$. Die quadrierte und aufsummierte Differenz zwischen empirischen und nach der Cavanagh-Hypothese prognostizierten Werten, relativiert auf die Anzahl der Freiheitsgrade, liegt bei $s_{emp}^2 = 9.41$. Wie auch bei den anderen beiden Prognosen ist der F-Wert $F(2,54) = 1.43$ niedriger als die kritische Grenze. Deskriptiv ergibt sich die beste Übereinstimmung zwischen der Kurzzeitgedächtnisspanne in bit als Asymptote der modifizierten Total-Time-Hypothese und den Prognosen nach der Cavanagh-Hypothese.

Diskussion

Die vorgestellten Experimente dienen der Überprüfung der Gültigkeit der Invarianzhypothesen von Baddeley et al. (1975), Dosher und Ma (1998) und Cavanagh (1972) sowie deren Integration in die modifizierte Total-Time-Hypothese (Bredenkamp, 1993) und das Modell von Hulme et al. (1999).

Für alle drei Invarianzhypothesen des Kurzzeitgedächtnisses sind die Voraussetzungen der Gültigkeit erfüllt: Eine höhere Kurzzeitgedächtnisspanne geht mit schnellerer Artikulationszeit (Baddeley), schnellerer Reproduktionszeit (Dosher und Ma) sowie schnellerer Suchzeit (Cavanagh) einher. Sowohl die Dauer der Gedächtnisspur, die Gesamtwiedergabe-Zeit als auch die Konstante C variieren zwischen Experiment 1 und 2 nicht systematisch. Vergleiche der empirischen Werte mit bereits veröffentlichten Werten zeigen hohe Übereinstimmungen, die für die Validität der hier angewendeten Verfahren sprechen. Allein für die Cavanagh-Hypothese treten problema-

tische Befunde auf. Zwar wächst hypothesenkonform mit steigendem Memory Load die Reaktionszeit linear und die Regressionsgeraden zwischen positivem und negativem Set verlaufen parallel, doch ist bei einigen Probanden der Anstieg in der Sternberg-Suchzeit negativ. Über diesen Effekt berichtete ebenfalls Bredenkamp (in Vorb.). Entscheidend könnte hier das experimentelle Design gewesen sein: Entgegen dem originalen Vorgehen von Sternberg (1966) wussten die Probanden in Experiment 1 und 2 genau, wie viele und welche Items im Gesamtitempool enthalten sind. Dieses Wissen konnte im Artikulationszeit-Experiment erworben werden, das immer als erstes durchgeführt wurde, und könnte verstärkt zu Strategie-Effekten im Umgang mit dem Stimulusmaterial geführt haben, in dem Sinne, dass sich das Kriterium der Antworttendenz verschiebt, je nachdem wie lang die präsentierte Liste ist (Klein, 1995).

Der zweite zentrale Befund bezieht sich auf den Vergleich der drei Invarianzhypothesen, der mit dem Modell nach Hulme et al. (1999) voll kompatibel ist. Wie die Auswertungen zeigen, besteht hypothesenkonform ein Zusammenhang zwischen Suchzeit und Reproduktionszeit sowie zwischen Artikulationszeit und Reproduktionszeit, nicht aber auch zwischen Suchzeit und Artikulationszeit. Für die Existenz und Relevanz von Prozessen, die während der Pausenzeiten ablaufen und die nach Hulme et al. (1999) mit Rekonstruktionsprozessen in Zusammenhang gebracht werden, spricht die Tatsache, dass die Reproduktionszeit über die Artikulationszeit und Suchzeit hinaus Varianz an der Prognose der Kurzzeitgedächtnisspanne aufklärt. Auffällig ist, dass die Prognose der Kurzzeitgedächtnisspanne durch die Suchzeit wesentlich schlechter ausfällt als durch die Artikulations- oder Reproduktionszeit. Ein Vergleich der Reliabilität der Zeitmessungen, definiert durch das Cronbach α als Interkorrelation zwischen den Items, verdeutlicht, dass die Reliabilität für die Artikulationszeit am höchsten ist ($\alpha = 0.96$ sowohl in Experiment 1 als auch 2). Die Reliabilitätsschätzung der Reproduktionszeit ist niedriger mit $\alpha = 0.86$ in Experiment 1 und $\alpha = 0.82$ in Experiment 2. Für die Suchzeiten ergibt sich für Experiment 1 der Wert von $\alpha = 0.50$ und für Experiment 2 $\alpha = 0.49$. Obwohl kein Effekt der Reihenfolge in der Versuchsdurchführung für die abhängigen Variablen Suchzeit, Reproduktionszeit, Artikulationszeit und Kurzzeitgedächtnis entdeckt wurde, ist dennoch eine deutliche Reduktion der Reliabilitätsmaße für die Suchzeit zu erkennen, wenn das Sternberg-Experiment als Letztes durchzuführen war. Zwischen den 60 Versuchsdurchgängen dieses Experiments war keine Pause vorgesehen. Drei Gründe sprechen dafür, dass die niedrige Reliabilität auf Ermüdungserscheinungen zurückzuführen ist: Erstens, zwei Probanden brachen das Experiment während des Absolvierens der Sternberg-Aufgabe als letztes

Experiment ab. Zweitens, nach dem Tukey-Kriterium mussten häufiger Reaktionszeiten im Sternberg-Experiment ausgeschlossen werden als bei den anderen beiden Experimenten. Dies liegt zum einen sicherlich daran, dass bei den anderen beiden Experimenten nicht so viele Versuchsdurchgänge absolviert wurden, andererseits liegt die Vermutung nahe, dass sich die Versuchspersonen bei Überanstrengung einfach eine »Auszeit« genommen haben. Drittens, Analysen auf Ebene einzelner Versuchspersonen zeigen, dass die Reaktionszeiten in der zweiten Testhälfte, die definiert ist als die Durchgänge 31–60, signifikant länger sind. Dies ist kein Effekt der Listenlänge, da er bei mehreren Versuchspersonen nachgewiesen werden konnte und zwischen den Versuchspersonen die Variablen Set und Listenlänge individuell und randomisiert zugewiesen wurden.

Da eine geringe Reliabilität die Schätzung der Höhe des Zusammenhangs zwischen zwei Variablen reduziert, ist aus der Gegenüberstellung der drei Invarianzhypothesen und dem deutlich schlechteren Abschneiden der Suchzeit bei der Prognose der Kurzzeitgedächtnisspanne, nicht eindeutig ein Vorteil der Hypothesen von Baddeley oder Dosher und Ma abzulesen. Reanalysiert man die Daten, indem man nur Zeitmessungen mit Werten von Cronbachs α .50 berücksichtigt, und für das Sternberg-Experiment allein die Daten der ersten Testhälfte, ergeben sich die folgenden Korrelationskoeffizienten (Tabelle 6).

Wie man im Vergleich zu Tabelle 4 sieht, ändert sich das Ergebnismuster durch Einführung des Reliabilitätskriteriums nur in einem Punkt: Die Korrelation der Suchzeit mit der Kurzzeitgedächtnisspanne weicht statistisch nun nicht mehr in ihrer Höhe von der Korrelation zwischen Artikulationszeit und Kurzzeitgedächtnisspanne ab ($t = -1.44, p < 0.5$). Der Zusammenhang zwischen Reproduktionszeit und Kurzzeitgedächtnisspanne dagegen ist noch immer deutlich höher ($t = -3.12, p < .05$). Die Reanalysen unter Berücksichtigung der Reliabilität belegen also, dass die Überlegenheit zumindest der Artikulationszeit gegenüber der Suchzeit zur Prognose der Kurzzeitgedächtnisspanne kein substanzieller Effekt ist. Ferner sprechen auch die bereinigten Daten für die Gültigkeit des Modells von Hulme et al. (1999).

Die abschließende Überprüfung, ob eine Integration der Invarianzhypothesen des Kurzzeitgedächtnisses in die modifizierte Total-Time-Hypothese

Tabelle 6: Ergebnisse der Korrelationsschätzungen unter Berücksichtigung der Reliabilität ($n = 43$)

	Artikulationszeit	Reproduktionszeit	Suchzeit
KZGS	–.562**	–.723**	–.317*
Artikulationszeit		.315**	.009
Reproduktionszeit			.267*

für Wortmaterial möglich ist, zeigt, dass auch bei lexikalischem Material die Asymptote der Lernfunktion durch alle drei Invarianzhypothesen des Kurzzeitgedächtnisses treffend geschätzt werden kann. Dieses Ergebnismuster bestätigt die empirischen Ergebnisse früherer Studien, die mit Konsonant-Vokal-Verbindungen durchgeführt wurden (Bredenkamp u. Hamm, 2001; Bredenkamp, in Vorb.) und spricht wiederum für die simultane Gültigkeit aller vier Hypothesen. Die modifizierte Total-Time-Hypothese bietet jedoch gegenüber den drei Invarianzhypothesen des Kurzzeitgedächtnisses den Vorteil, dass nicht nur die Kurzzeitgedächtnisspanne adäquat geschätzt werden kann, sondern darüber hinaus auch die Langzeitgedächtnisspanne, die in diesem Beitrag nicht berücksichtigt wurde (vgl. Bredenkamp, 1993; Hamm, 2001).

Summa summarum kristallisierten sich drei zentrale Befunde heraus. Erstens, trotz Problemen im experimentellen Design mit dem Sternberg-Experiment konnte gezeigt werden, dass die drei Invarianzhypothesen des Kurzzeitgedächtnisses gleichzeitig gültig sind. Zweitens, der Vergleich der einzelnen Hypothesen spricht für das Modell von Hulme et al. (1999) und drittens, eine Überführung der Invarianzhypothesen in die modifizierte Total-Time-Hypothese ist auch für lexikalisches Material möglich.

Literatur

Baayen, R.H., Piepenbrock, R., Gulikers, L. (1995). *The Celex lexical data base (CD-Rom).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 147–423.

Baddeley, A.D., Chincotta, D., Stafford, L., Turk, D. (2002). Is the word length effect in STM entirely attributable to output delay? Evidence from serial recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 55A,* 353–369.

Baddeley, A.D., Thomson, N., Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14,* 575–589.

Bredenkamp, J. (1975). Die modifizierte Invarianzhypothese und die Hypothese einer konstanten Langzeit-Gedächtnisspanne. *Psychologische Beiträge, 17,* 483–496.

Bredenkamp, J. (1993). Die Verknüpfung verschiedener Invarianzhypothesen im Bereich der Gedächtnispsychologie. *Zeitschrift für Experimentelle und Angewandte Psychologie, 40,* 368–385.

Bredenkamp, J. (2004). Cavanagh's hypothesis within the context of other invariance hypotheses: Theory and data. In C. Kaernbach, E. Schröger u. H. Müller (Hg.), *Psychophysics beyond sensation: Laws and invariants of human cognition.* Mahwah, NJ: Erlbaum.

Bredenkamp, J., Klein, K.-M. (1998). Experimental tests of a model connecting three invariance hypotheses on learning and memory processes. *Zeitschrift für Psychologie, 206,* 107–124.

Bredenkamp, J., Hamm, S. (2001). Further experimental tests of invariance hypotheses on learning and memory processes. *Zeitschrift für Psychologie, 209,* 227–244.

Cavanagh, J.P. (1972). Relation between the immediate memory span and the memory search rate. *Psychological Review, 79,* 525–530.

Chincotta, D., Underwood, G. (1997). Digit span and articulatory suppression: A cross-linguistic comparison. *European Journal of Cognitive Psychology, 9,* 89–96.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology, 55,* 75–84.

Cooper, E.H., Pantle, A.J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin, 68,* 221–234.

Cowan, N. (1994). Mechanisms of verbal short-term memory. *Current Directions in Psychological Science, 3,* 185–189.

Cowan, N., Day, L., Saults, J.S., Keller, T.A., Johnson, T., Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language, 31,* 1–17.

Dosher, B.A., Ma, J.J. (1998). Output loss or rehearsal loop? Output-time versus pronunciation-time limits in immediate recall for forgetting-matched materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 316–335.

Erdfelder, E., Faul, F., Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers, 28,* 1–11.

Hamm, S. (2001). *Synthese von Kurz- und Langzeitgedächtnisspanne. Experimentelle Überprüfung eines mathematischen Modells.* Aachen: Shaker.

Hasselhorn, M. (1988). Wie und warum verändert sich die Gedächtnisspanne über die Lebensspanne? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 20,* 322–337.

Hulme, C., Newton, P., Cowan, N., Stuart, G., Brown, G. (1999). Think before you speak: Pauses, memory search, and trace reintegration processes in verbal memory span. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 447–463.

Kintsch, W. (1970). *Learning, memory, and conceptual processes.* New York: Wiley.

Klein, K.-M. (1995). *Experimentelle Untersuchungen zu zwei Invarianzhypothesen des Kurzzeitgedächtnisses.* Bonn: Pace.

Lovatt, P., Avons, S.E., Masterson, J. (2000). The word-length effect and disyllabic words. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 53A,* 1–22.

McNemar, Q. (1962). *Psychological statistics.* New York: Wiley.

Nevelsky, P.B. (1970). Comparative investigation of the short- and long-term memory span. In K.H. Pribram u. D.E. Broadbent (Hg.), *Biology of memory* (S. 21–28). New York: Academic Press.

Puckett, J.M., Kausler, D.H. (1984). Individual differences and models of memory span: A role for memory search rate? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 72–82.

Schweickert, R., Boruff, B. (1986). Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 419–425.

Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 51A,* 283–304.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153,* 652–654.

Tukey, J.W. (1977). *Exploratory data-analysis.* Reading: Addison-Wesley.

Walker, I., Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1256–1271.

Bianca Vaterrodt-Plünnecke

Die Prozess-Dissoziations-Prozedur mit der lexikalischen Entscheidungsaufgabe: Bestandsaufnahme und Validierung eines multinomialen Modells[1]

Zusammenfassung

In diesem Kapitel wird ein multinomiales Modell vorgestellt, das die Messung von bewussten, unbewussten und Antworttendenz-Prozessen ermöglichen soll. Es handelt sich um eine spezielle Variante des PD+-Modells für Wort-Nonwort-Entscheidungen (Vaterrodt-Plünnecke et al., 1996), das in der Vergangenheit bereits häufig angewendet wurde, obwohl eine konsequente Validierung der Modellparameter noch ausstand. Dieser Beitrag ist eine Reaktion auf dieses Desiderat, indem die Validität aller Modellparameter über systematische experimentelle Manipulationen geprüft wird.

Einleitung

Die unterschiedlichen Verfahren zur Trennung von kontrollierten (bewussten) und automatischen (unbewussten) informationsverarbeitenden Prozessen werden seit vielen Jahren eingehend diskutiert. Jacoby stellte 1991 die Prozess-Dissoziations-Prozedur vor, die es ermöglicht, im Rahmen einer Aufgabe über ein Messmodell die kontrollierten und automatischen Anteile kognitiver Prozesse zu erfassen. Diese Methode fand schnell viele Anhänger, rief aber auch zahlreiche Kritiker auf den Plan. Die Prozess-Dissoziations-Prozedur beeindruckte dadurch, dass sie die Probleme des bis dahin üblichen Dissoziationsparadigmas zu überwinden schien und gleichzeitig in vielen Bereichen der kognitiven Psychologie anwendbar war. Kritik zog vor allem das Messmodell auf sich. Insbesondere wurden die Art der

1 Die hier dargestellten Untersuchungen wurden im Rahmen des DFG-Projekts (Va131/2–1) in Zusammenarbeit mit Herrn Prof. Dr. Jürgen Bredenkamp durchgeführt. Wir danken Herrn Dr. Thomas Krüger, Frau Elfi Stirner, Frau Sonja Boddenberg, Herrn Alexander Herschung und Frau Claudia Vieth für ihre Mitarbeit. Herrn Prof. Dr. Edgar Erdfelder, Herrn Dr. Arndt Bröder und Frau Anna Zaunbauer sowie zwei anonymen Gutachtern danke ich für wertvolle Anmerkungen zu einer früheren Fassung dieses Kapitels.

Berücksichtigung zufälliger Prozesse (Raten), die Annahme der stochastischen Unabhängigkeit von kontrollierten und automatischen Prozessen und die fehlende statistische Testbarkeit des Modells kritisiert. Die Kritik war fruchtbar. Für viele der Probleme sind mittlerweile Lösungsansätze entwickelt worden (z. B. Buchner et al., 1995; Erdfelder u. Buchner, 1998; Vaterrodt-Plünnecke et al., 2002). Nicht nur das Messmodell, sondern auch die Prozedur selbst wurde in der Vergangenheit vielfältig erweitert und modifiziert (Krüger, 1999).

Die Grundidee ist jedoch gleich geblieben: Die Prozess-Dissoziations-Prozedur soll es ermöglichen, die Anteile kontrollierter und automatischer Prozesse zu bestimmen. Diese Anteile werden über ein Messmodell aus der Leistung bei einer Aufgabe unter zwei Bedingungen erschlossen. Die eine Bedingung ist so gestaltet, dass kontrollierte und automatische Prozesse gemeinsam zur Lösung der Aufgabe beitragen (»Facilitation«). Die andere Bedingung ist so gestaltet, dass nur die kontrollierten Prozesse zur korrekten Lösung beitragen, während automatische Prozesse – in Abwesenheit von kontrollierten Prozessen – zu Fehlern führen (»Interferenz«). Die Art und Weise, wie im Rahmen verschiedener Prozess-Dissoziationstechniken diese beiden Bedingungen realisiert werden, unterscheidet sich erheblich. Daher verwundert es nicht, dass durch die verschiedenen Prozeduren ganz unterschiedliche Arten von kontrollierten und automatischen Prozessen erfasst werden, die sich zudem mehr oder weniger gut bewährt haben (Erdfelder u. Buchner, 2003).

Das PD+-Modell (Vaterrodt-Plünnecke et al., 1996) integriert verschiedene Varianten der Prozess-Dissoziationstechniken in einem Rahmenmodell. Generell lässt sich das PD+-Modell für viele Bereiche spezifizieren, in denen sich zur Erfassung von kontrollierten und automatischen Prozessen, Interferenz- und Facilitationsbedingungen erzeugen lassen und zudem Antwort- oder »Ratetendenzen« eine Rolle spielen. Eine dieser Varianten soll in diesem Beitrag genauer betrachtet und diskutiert werden. Es handelt sich um die lexikalische Entscheidungsaufgabe und ein multinomiales Modell, das erstmals 1994 dargestellt wurde (Vaterrodt-Plünnecke, 1994).

Die multinomiale Modellierung kognitiver Prozesse geht auf Batchelder und Riefer (1999; Riefer u. Batchelder, 1988) zurück. Sie ist der deduktivistischen Forschungsmethodologie in hohem Maß verpflichtet, da von einer Theorie ausgehend ein Messmodell formuliert wird, das empirisch prüfbar ist. Ein solches Modell macht die psychologischen Prozesse, die der Theorie zugrunde liegen, explizit und liefert zugleich die statistische Testmethode. Damit ist der Ansatz auch gegen Kritiken resistent, die einen fehlenden Bezug zwischen Modell und statistischem Werkzeug monieren (Bröder,

2000). Innerhalb der deduktivistischen Theorie des Experiments berührt dies die wichtige Frage nach der *Validität des statistischen Schlusses*. Über die A-priori-Festlegung von α, β und der Effektgröße und der daraus resultierenden Stichprobengröße (vgl. Bredenkamp, 1972; 1980), zum Beispiel auch durch Kompromiss-Power-Analysen, soll die Validität des statistischen Schlusses im Rahmen der multinomialen Modellierung gesichert werden.

Die Begriffe »bewusst« und »unbewusst« werden über das Modell »implizit definiert« (Westermann, 1987). Dabei wird das Messmodell im Sinne von Bochenski (1954) als axiomatisches System verstanden, welches eine syntaktische Definition der Begriffe (»Zeichen«) »Bewusstsein« und »Unbewusstsein« liefert. Die semantische Bedeutung dieser Begriffe muss dann in systematischen Validierungsstudien überprüft werden (vgl. Bredenkamp, 1998). Validierungsstudien, wie sie in diesem Beitrag dargestellt werden, betreffen ein zentrales Konzept der deduktivistischen Theorie des Experiments: die *Variablenvalidität*. Nur wenn die Variablen einer Untersuchung das messen, was sie gemäß der zu prüfenden Theorie messen sollen, können sie einen Beitrag zu einer strengen Hypothesenprüfung leisten. Damit sollen Scheinfalsifikationen und Scheinbestätigungen verhindert werden (Bredenkamp, 1980).

Nur unter der Voraussetzung der Modellgültigkeit *und* der Validität der Parameter des Modells können schließlich die interessierenden substanzwissenschaftlichen Fragestellungen untersucht werden.

Dieser Beitrag illustriert exemplarisch, wie eine Überprüfung der Variablenvalidität im Rahmen stochastischer Messmodelle, die in den letzten zehn Jahren zunehmend Eingang in die kognitionswissenschaftliche Forschung gefunden haben, erfolgen kann.

Die dargestellten Untersuchungen haben die Validierung des PD+-Modells für lexikalische Entscheidungsaufgaben zum Ziel. Dazu wird im folgenden Abschnitt das Modell beschrieben und kurz auf Untersuchungen eingegangen, die im klinischen Bereich durchgeführt wurden. Obwohl das Modell in der Vergangenheit bereits häufig angewendet wurde, mangelt es an Untersuchungen, die systematisch auf die Validierung der Modellparameter abzielen. Genau dies beabsichtigen die Experimente, die im dritten Abschnitt dargestellt werden. Im letzten Abschnitt werden die wichtigsten Ergebnisse zusammengefasst und im Hinblick auf die Validität des Modells diskutiert.

Beschreibung und Anwendung des Modells

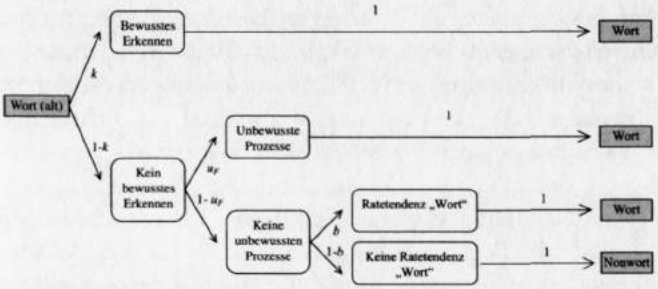
Die Prozess-Dissoziations-Prozedur zur Trennung automatischer und kontrollierter Prozesse im Rahmen einer lexikalischen Entscheidungsaufgabe sieht vor, dass die Versuchspersonen in einer Studierphase bestimmte Wörter lesen, die per Zufall aus einem Wörterpool ausgewählt werden (Beispiel: Gelesen wird: *Dozent, Tinte;* nicht gelesen wird: *Blume, Karton*). Nach der Darbietung der Wörter folgt die lexikalische Entscheidungsaufgabe, bei der es sich um einen wahrnehmungsnahen indirekten Gedächtnistest handelt. Die Probanden sollen entscheiden, ob es sich bei einem nur wenige Millisekunden präsentierten Item um ein Wort oder um ein Nonwort handelt. Die Gedächtnisnutzung spielt für die Wörter, die in der Lernphase gezeigt wurden sowie für die Anagramme (Nonwörter) dieser Wörter, die in der Studierphase zwar selbst nicht gezeigt wurden aber eine gewisse Ähnlichkeit zu den Studierwörtern aufweisen, eine Rolle. Für nicht dargebotene Wörter und deren Anagramme sollte die Gedächtnisnutzung keine Rolle spielen. In der Testphase werden also Wörter aus der Lernphase (*Dozent, Tinte; »alte Wörter«*) und deren Anagramme (*Zedont, Tenti; »alte Nonwörter«*) sowie nicht gelesene Wörter (*Blume, Karton; »neue Wörter«*) und deren Anagramme (*Blemu, Torkan; »neue Nonwörter«*) präsentiert.

Analog den Annahmen der Prozess-Dissoziations-Prozedur wird davon ausgegangen, dass die vorangehende Präsentation eines Worts Gedächtnisspuren hinterlässt. Die unwillkürliche Nutzung dieser Spuren kann einerseits die Aufgabenlösung in der Testphase unterstützen (Facilitation) und andererseits die Lösung der Aufgabe behindern (Interferenz). Dazu soll das PD+-Modell betrachtet werden, das sich für die beschriebene Prozedur gemäß Abbildung 1 spezifizieren lässt (vgl. Vaterrodt-Plünnecke, 1994; Krüger, 1999).

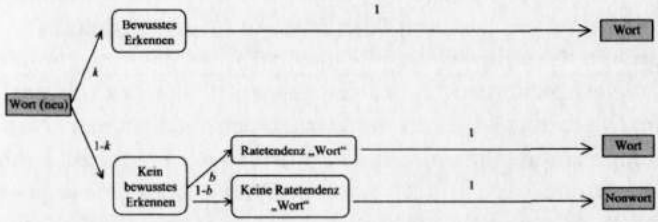
Das Modell enthält vier Parameter: Der Parameter k soll wahrnehmungsnahe bewusste Diskriminationsprozesse erfassen, die Parameter u_F und u_I beziehen sich auf den Einfluss unbewusster Gedächtnisleistungen in der Facilitations- und Interferenzbedingung und Parameter b soll die Antworttendenz repräsentieren.

Teilbaum T1 beschreibt die Facilitationsbedingung, in der ein in der Studierphase gelesenes Wort (z. B. *Dozent, Tinte*) in der Testphase wenige Millisekunden dargeboten wird. Es gibt verschiedene Möglichkeiten, in dieser Bedingung zu einer korrekten Antwort zu gelangen: Entweder die Versuchsperson erkennt bewusst, dass es sich bei dem nur sehr kurz dargebotenen Item um ein Wort handelt. Die Wahrscheinlichkeit dafür beträgt k. Falls die Versuchsperson nicht in den Zustand des bewussten Erkennens

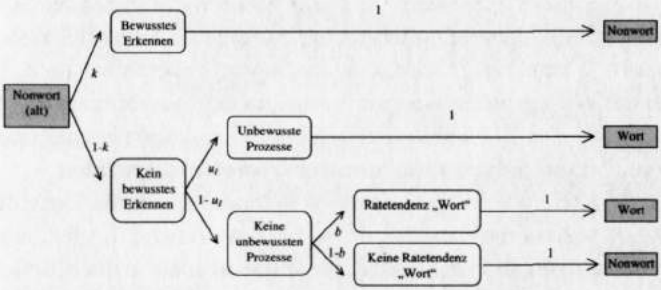
Teilbaum 1: Wort (alt) Facilitation



Teilbaum 2: Wort (neu)



Teilbaum 3: Nonwort (alt) Interferenz



Teilbaum 4: Nonwort (neu)



Abbildung 1: Spezifikation des PD+-Modells für Wort-Nonwort-Entscheidungen (vgl. PD+-Modell von Vaterrodt-Plünnecke et al., 1996; Vaterrodt-Plünnecke, 1994).

gelangt, dann können unbewusste Prozesse als Folge einer erhöhten perzeptuellen Flüssigkeit für alte Wörter zur korrekten Antwort führen. Die Wahrscheinlichkeit dafür beträgt $(1\text{-}k)\cdot u_F$. Es gibt noch eine dritte Möglichkeit, die korrekte Lösung in der Facilitationsbedingung zu erzielen: Falls die Versuchsperson weder bewusst noch unbewusst zur richtigen Lösung kommt, dann hat sie noch die Möglichkeit, »Wort« zu raten. Die Wahrscheinlichkeit dafür beträgt $(1\text{-}k)\cdot(1\text{-}u_F)\cdot b$.

In der Testbedingung, die unter Teilbaum T2 beschrieben ist, werden neue Wörter (z. B. *Blume, Karton*) präsentiert. Da für diese Wörter keine Erleichterung der Aufgabenlösung als Folge erhöhter perzeptueller Flüssigkeit zu erwarten ist, kann hier die korrekte Lösung nur aufgrund einer bewussten Diskriminationsleistung (k) oder – falls bewusstes Erkennen nicht möglich ist – auf der Basis von Rateprozessen ($(1\text{-}k)\cdot b$) erfolgen.

In der Interferenzbedingung (Teilbaum T3) werden Anagramme (z. B. *Zedont, Tenti*) präsentiert, die zu den gelesenen Wörtern (*Dozent, Tinte*) gehören. Die korrekte Antwort der Versuchsperson lautet hier »Nonwort«. Die vorangegangene Präsentation des Worts in der Lernphase kann dann dazu führen, dass das entsprechende Anagramm, das ja eine gewisse Ähnlichkeit zum gelesenen Wort aufweist, mit einer Wahrscheinlichkeit von u_I fälschlicherweise als Wort bezeichnet wird. In diesem Fall hat die perzeptuelle Flüssigkeit – im Sinne eines unbewussten Einflusses der Gedächtnisrepräsentation – die Behinderung der Aufgabenlösung zur Folge.

Teilbaum T4 stellt die Bedingung dar, in der Anagramme (z. B. *Blemu, Torkan*) der Wörter präsentiert werden, die in der Studierphase nicht gelesen worden sind (z. B. *Blume, Karton*). Hier ist keine Interferenz (und auch keine Facilitation) aufgrund unbewusster Prozesse zu erwarten.

In der klinischen Forschung erfreuen sich neben anderen Varianten der Prozess-Dissoziations-Prozedur auch die Anwendung der lexikalischen Entscheidungsaufgabe und das dargestellte Modell großer Beliebtheit (Scholz et al., 1997; Windmann u. Krüger, 1998; Ott et al., 2000). Man erhofft sich durch solche Anwendungen, Aufschluss über kognitive Auffälligkeiten psychiatrischer Gruppen zu erhalten. So wird beispielsweise davon ausgegangen, dass Angst-Patienten im Vergleich zu depressiven Patienten zu einem erhöhten Anteil impliziter Informationsverarbeitung neigen, während depressive Patienten einen erhöhten Anteil expliziter Informationsverarbeitung aufweisen. Scholz et al. (1997; Scholz, 1997) wendeten eine akustische Variante der lexikalischen Entscheidungsaufgabe bei Patienten mit somatoformen Störungen an, um zu prüfen, ob deren Verarbeitungsprozesse mit denen von Angstpatienten vergleichbar sind. Die Ergebnisse ihrer Untersuchungen legen nahe, dass Patienten mit somatoformen

Störungen bei störungsspezifischem Wortmaterial, ähnlich wie Angstpatienten, zu einem erhöhten Anteil impliziter Verarbeitung neigen. Windmann und Krüger (1998) fanden bei der Anwendung des oben dargestellten Modells, dass Patienten mit Panikstörungen im Vergleich zu gesunden Probanden keine Unterschiede hinsichtlich der unbewussten Prozesse und der Diskriminationsleistung zeigten, jedoch unterschiedliche Antworttendenzen aufwiesen. Zaunbauer und Bredenkamp (2002) konnten bei Fibromyalgiepatienten und einer gesunden Kontrollgruppe keine Unterschiede hinsichtlich der Verarbeitung von supra- und subliminal geprimten Adjektiven mit unterschiedlicher Valenz nachweisen. Allerdings fielen die Parameterschätzungen insgesamt sehr gering aus, was eine Interpretation der Daten erschwerte.

Validierungs-Untersuchungen

Die Interpretation der vorgefundenen Effekte setzt nicht nur voraus, dass das Modell eine gute Anpassung an die vorliegenden Daten aufweist, sondern auch, dass das Modell valide ist. Validierungsstudien sind erforderlich, um die psychologische Bedeutung der einzelnen Modellparameter abzusichern. Die Parameter sollten tatsächlich der »Deutung« entsprechen, die man ihnen zuspricht. Die Validierung muss über die experimentelle Manipulation verschiedener unabhängiger Variablen erfolgen, deren Einfluss auf kontrollierte, automatische und Rate-Prozesse man als bekannt voraussetzen kann. Es muss geprüft werden, ob sich die Variation einzelner unabhängiger Variablen tatsächlich nur auf den sie betreffenden Modellparameter auswirkt. Oft sind die Validierungsuntersuchungen direkt in die Modellprüfung einbeziehbar.

Die derzeit vorliegenden Validierungsstudien zur Prozess-Dissoziations-Prozedur müssen als unvollständig bewertet werden. Ziel der folgenden Untersuchungen ist es, dieses Defizit in Bezug auf die lexikalische Entscheidungsaufgabe zu reduzieren, indem systematische Manipulationen aufgrund des psychologisch relevanten Hintergrundwissens durchgeführt werden, um deren Auswirkungen auf die Parameterschätzungen zu überprüfen. Die Experimente dienen somit der Absicherung der psychologischen Bedeutung der Modellparameter k, u_I/u_F und b. Angenommen wird, dass sich der Parameter k auf das bewusste Diskriminieren bezieht, die u-Parameter auf den unbewussten Einfluss der Gedächtnisrepräsentation und b auf die Antworttendenz.

Experiment 1: Variation der Darbietungszeit zur Validierung des Parameters *k*

In der ersten Untersuchung sollte über die Manipulation der Präsentationsdauer in der Testphase die Validität des Diskriminationsparameters *k* geprüft werden. Unterschiedlich lange Präsentationszeiten der Items in der Testphase sollten mit einer unterschiedlich guten Wahrnehmbarkeit und damit Diskriminierbarkeit von Wörtern und Nonwörtern einhergehen. Eine Präsentationszeit von 56 ms (4 Kathodenstrahlzyklen) sollte eine höhere Ausprägung des Parameters k zur Folge haben als eine Präsentationszeit von 42 ms (3 Kathodenstrahlzyklen), was wiederum ein höheres *k* bedingen sollte als eine Präsentationsdauer von 28 ms (2 Kathodenstrahlzyklen). Die Ausprägung der anderen Parameter *b*, u_F und u_I sollte von einer derartigen Variation wenig beeinflusst werden.

Versuchspersonen. An der Untersuchung nahmen 36 studentische Versuchspersonen teil, die per Zufall einer der 3 Darbietungsbedingungen zugeteilt wurden. Die Bedingungen wurden interindividuell variiert. Für die Teilnahme an dieser Untersuchung wurden die Versuchspersonen entweder mit 5 DM bezahlt oder sie erhielten eine Versuchspersonenbescheinigung.

Material. Als Material dienten 64 Wörter, die hinsichtlich ihrer Wortlänge (5–6 Buchstaben), Konkretheit, Bildhaftigkeit und Worthäufigkeit vergleichbar sind (Baschek et al., 1977). Außerdem wurden Anagramme dieser 64 Wörter gebildet. Hierbei handelte es sich um Wortapproximationen zweiter Ordnung (Perrig, 1992), was bedeutet, dass die Anagramme noch aussprechbar sind, jedoch nicht auf den ersten Blick dem entsprechenden Wort zugeordnet werden können. Die Hälfte der 64 Wörter wurde in der Lernphase per Zufall ausgewählt und präsentiert. Die komplett randomisierte Präsentation der 32 Wörter aus dem Pool der 64 Wörter gewährleistet, dass Unterschiede in den Bedingungen nie auf eventuell nicht berücksichtigte Unterschiede im Itemmaterial zurückzuführen sind. In allen Versuchen waren die physikalischen und optischen Kontextbedingungen der Studierphase analog zu denen der Testphase gestaltet. Dadurch sollte eine möglichst optimale transferangemessene Verarbeitung sicher gestellt werden.

Durchführung. Ein Turbo-Pascal-Programm steuerte den kompletten Ablauf der Untersuchung. In der Studierphase wurden den Versuchspersonen 32 Wörter dargeboten, die zufällig aus den insgesamt 64 Wörtern ausgewählt worden waren. Die Versuchspersonen wurden instruiert, diese Wörter zu lesen. Jedes Wort wurde drei Sekunden auf dem Monitor präsentiert und mit einer Pattern-Maske überschrieben (XXXXX). Im Anschluss an die Studierphase folgte die lexikalische Entscheidungsaufgabe,

die als Wahrnehmungstest eingeführt wurde. Den Versuchspersonen wurden 32 alte und 32 neue Wörter und die dazu gehörigen 64 Anagramme in zufälliger Reihenfolge dargeboten. Die Dauer der Präsentation wurde durch eine spezielle Bildschirmansteuerung und die Pattern-Maskierung (XXXXX) der Items kontrolliert und interindividuell variiert (28 ms, 42 ms oder 56 ms). Die Versuchspersonen sollten bei den sehr kurz dargebotenen Informationen entscheiden, ob es sich um ein Wort oder um ein Nonwort handelt, indem sie die mit »W« oder »N« markierte Taste der Computertastatur drückten.

Ergebnisse. In der Bedingung mit der sehr kurzen Präsentationsdauer von 28 ms zeigte sich, dass praktisch keine Diskriminationsleistung vorhanden war. Bei der Darbietung eines Items – ganz gleich, ob alt oder neu, Wort oder Nonwort – wurde geraten. In etwas weniger als der Hälfte der Fälle wurde »Wort« ($b = .46$) geraten, in etwas mehr als der Hälfte der Fälle wurde »Nonwort« geraten. Die Parameter k und u_F/u_I wurden auf den Grenzwert 0 geschätzt. Da die 28-ms-Bedingung die Versuchspersonen einzig und allein zum »Raten« motivierte, wurde diese Bedingung bei den Analysen nicht berücksichtigt. Die Modellanpassungstests werden daher nur auf der Grundlage der 42-ms- und 56-ms-Bedingungen dargestellt.

Tabelle 1 zeigt die mittels multinomialer Modellierung erzielten Parameterschätzungen für die beiden Bedingungen.

Ausgehend von diesem Modell ($G^2 = 0.0$; $df = 0$) wurden Restriktionen eingeführt, mit denen man Aussagen über die den Restriktionen zugrunde liegenden Hypothesen und über die Anpassungsgüte des restringierten Modells machen kann. Für die gegebene Anzahl von $n = 3072$ Beobachtungen (128 [Antworten je Vp] · 24 [Anzahl der Vpn]) ergibt eine Kompromiss-Power-Analyse (Erdfelder et al., 1996; Erdfelder et al., in diesem Band) für die Aufdeckung kleiner Effekte ($w = 0.1$) ein $\alpha = \beta < .01$ (somit eine Power $1\text{-}\beta > .99$) und die kritischen Werte $G^2_{Krit} = 8.32$ für $df = 1$ und $G^2_{Krit} = 10.1$ für $df = 2$.

Die Anpassungstests für die sukzessiv restringierten Modelle entsprachen weitgehend unseren Erwartungen: So führte die erste Restriktion, die Gleichsetzung der Parameter für die Antworttendenzen (b) für die beiden Bedingungen (42 ms / 56 ms), zu keiner signifikanten Verletzung der Mo-

Tabelle 1: Parameterschätzungen und 95 %-Konfidenzintervalle für Experiment 1

Präsentationszeit in ms	k	b	u_I	u_F
42	.30 [.24 .35]	.48 [.44 .52]	.11 [–.04 .27]	.27 [.14 .40]
56	.40 [.34 .45]	.43 [.39 .48]	.20 [.04 .36]	.47 [.36 .59]

dellannahmen ($G^2 = 1.76$; $df = 1$). Demnach hatten unterschiedliche Präsentationszeiten keinen Einfluss auf die Antworttendenz. Die zweite Restriktion, welche die Gleichsetzung der Parameter für die unbewussten Prozesse (u_I bzw. u_F) über die unterschiedlichen Präsentationszeiten vorsieht, hatte ebenfalls keine signifikante Verletzung der Modellanpassung zur Folge ($G^2 = 2.6$; $df = 2$). Erwartungsgemäß führte allein die dritte Restriktion, die sich auf den Diskriminationsparameter (k) für die unterschiedlichen Präsentationszeiten bezieht, zu einer signifikanten Verletzung der Modellannahmen ($G^2 = 13.97$; $df = 1$). Das erste Experiment lässt somit den Schluss zu, dass sich eine Variation der Präsentationszeit in der Testphase auf die Diskriminationsleistung auswirkt und weder die Antworttendenz noch die unbewussten Prozesse beeinflusst. Auf die deskriptiv vorliegenden Unterschiede in den Schätzungen für die unbewussten Prozesse in den verschiedenen Bedingungen soll in der Diskussion eingegangen werden.

Experiment 2: Manipulation der Antworttendenz zur Validierung des Parameters *b*

Das zweite Experiment sah eine Manipulation des Antwortverhaltens über verschiedene Auszahlungsbedingungen zur Validierung des Antworttendenzparameters b vor. Der Parameter b soll die Tendenz, »Wort« zu raten, reflektieren. Werden die Antworten »Wort« vergleichsweise stärker belohnt als die Antworten »Nonwort«, dann sollte sich dies in einem höheren Schätzwert des b-Parameters bemerkbar machen (liberale Bedingung), als wenn die Antworten »Wort« weniger stark belohnt werden als die Antworten »Nonwort« (konservative Bedingung). Die Ausprägung des b-Parameters in einer neutralen Bedingung sollte zwischen den extremeren Formen der Manipulation liegen. Eine Variation der Antworttendenz über ein Belohnungssystem (verschiedene Auszahlungsbedingungen) sollte sich allein auf den Parameter b auswirken und keinen Einfluss auf die Parameter k, u_I und u_F haben.

Versuchspersonen. An der Untersuchung nahmen 36 studentische Versuchspersonen teil, die per Zufall einer der drei Auszahlungsbedingungen »liberal«, »neutral« oder »konservativ« zugeteilt wurden. Die Bedingungen wurden interindividuell variiert. Für ihre Teilnahme wurde den Versuchspersonen ein Betrag in Abhängigkeit von der erzielten Punktzahl ausgezahlt.

Material. Als Material dienten die 64 Wörter und Nonwörter des ersten Experiments.

Tabelle 2: Auszahlungsmatrix zur Manipulation der Antworttendenz

	Konservativ Darbietung		Neutral Darbietung		Liberal Darbietung	
Antwort	Wort	Nonwort	Wort	Nonwort	Wort	Nonwort
»Wort«	+1	−3	+1	−1	+3	−1
»Nonwort«	−1	+3	−1	+1	−3	+1

Durchführung. Analog zu Experiment 1 wurden den Versuchspersonen per Zufall 32 der 64 Wörter für drei Sekunden dargeboten. Sie wurden instruiert, diese Wörter zu lesen. Im Anschluss an die Studierphase folgte der indirekte Gedächtnistest, der als Wahrnehmungstest eingeführt wurde. Hier setzte auch die Manipulation der Antworttendenz ein. Jeder Versuchsperson wurde die ihr zugeloste Auszahlungsbedingung schriftlich erklärt. Tabelle 2 fasst die verschiedenen Bedingungen zusammen.

In der konservativen Bedingung führte eine falsche Antwort »Wort« zu einem höheren Punktabzug (−3 Punkte) als eine falsche Antwort »Nonwort« (−1 Punkt). Hingegen wurde eine korrekte Antwort »Wort« nur mit einem Punkt belohnt, während eine korrekte Antwort »Nonwort« 3 Punkte einbrachte. In der neutralen Bedingung führte jede falsche Antwort zu demselben Punktabzug (−1 Punkt), und jede korrekte Antwort wurde mit einem Punkt belohnt. In der liberalen Bedingung wurden falsche »Nonwort«-Antworten stärker bestraft als falsche »Wort«-Antworten.

Die Versuchspersonen erhielten die Information, dass die erzielte Punktzahl in Geld umgerechnet würde, das ihnen nach dem Versuch ausgezahlt werde (5 Pfennige pro Punkt in der konservativen und der liberalen Bedingung, 10 Pfennige pro Punkt in der neutralen Bedingung). Dies sollte die Versuchspersonen dazu motivieren, eine möglichst hohe Punktzahl zu erreichen. Wie in Experiment 1 wurden in der Testphase 32 alte und 32 neue Wörter und die dazu gehörenden Anagramme in zufälliger Reihenfolge dargeboten. Die Versuchspersonen sollten jeweils entscheiden, ob es sich um ein Wort oder um ein Nonwort handelte. Die Dauer der Präsentation betrug 42 ms und wurde durch die beschriebene Bildschirmansteuerung und Maskierung kontrolliert. Während der lexikalischen Entscheidungs-

Tabelle 3: Parameterschätzungen und 95 %-Konfidenzintervalle für Experiment 2

Antworttendenz	k	b	u_I	u_F
Konservativ	.14 [.09 .19]	.19 [.16 .22]	.02 [−.04 .09]	.23 [.16 .30]
Neutral	.03 [−.03 .09]	.53 [.50 .56]	.00 [−.13 .13]	.28 [.17 .39]
Liberal	.16 [.11 .22]	.74 [.71 .77]	.00 [−.25 .25]	.30 [.12 .48]

aufgabe wurde die erzielte Punktzahl nach jeder Wort-Nonwort-Entscheidung am unteren Rand des Bildschirms visuell zurückgemeldet.

Ergebnis. Tabelle 3 zeigt die mittels multinomialer Modellierung erzielten Parameterschätzungen für die verschiedenen Auszahlungsbedingungen.

Eine Kompromiss-Power-Analyse ergibt für die in diesem Experiment gegebene Anzahl von $n = 4608$ Beobachtungen (128 [Antworten je Vp] · 36 [Anzahl der Vpn]) zur Aufdeckung kleiner Effekte ($w = 0.1$) ein $\alpha = \beta < .001$ (somit eine Power $1-\beta > .999$) und den kritischen Wert $G^2_{Krit} = 14.1$ bei 2 Freiheitsgraden und $\alpha = \beta < .01$ (somit eine Power $1-\beta > .99$) den kritischen Wert $G^2_{Krit} = 17.2$ bei 4 Freiheitsgraden.

Die unterschiedliche Ausprägung des Schätzers für den Parameter b in den drei experimentellen Bedingungen entsprach genau der Erwartung: In der konservativen Bedingung fiel er geringer und in der liberalen Bedingung höher aus als in der neutralen Bedingung. Ferner hatte die Manipulation der Antworttendenz keinen signifikanten Einfluss auf die Ausprägung der Schätzer für k und u_I beziehungsweise u_F. Die Parameterrestriktionen für u_I und u_F über die drei experimentellen Bedingungen hatten keine bedeutsame Abweichung der Modellanpassung zur Folge ($G^2 = 4.9$; $df = 4$).

Zudem zeigte sich den Erwartungen entsprechend, dass die Variation der Antworttendenz keinen Einfluss auf die Schätzung des Parameters k ausübte. So führte eine Restriktion der k-Parameter über die drei Antworttendenz-Bedingungen zu keiner signifikanten Verschlechterung der Modellanpassung ($G^2 = 10.69$; $df = 2$). Dagegen hatte die Restriktion $b_N = b_K = b_L$, zur Prüfung der Gleichheit der Antworttendenz über die verschiedenen Bedingungen, eine signifikante Verschlechterung der Modellanpassung zur Folge ($G^2 = 672.4$; $df = 2$). Deskriptiv zeigte sich, dass b in der neutralen Bedingung beinahe den gleichen Wert wie in Experiment 1 hat. Entgegen den Erwartungen fiel die Schätzung für die Diskriminationsleistung k in der neutralen Bedingung tendenziell geringer aus als in den beiden anderen Bedingungen. Trotz dieser Einschränkung kann das Experiment als Beleg dafür gewertet werden, dass der Parameter b die Antworttendenz reflektiert.

Experiment 3: Variation der Darbietungsmodalität zur Validierung der Parameter für unbewusste Prozesse u_I und u_F

Das dritte Experiment galt der Validierung der u-Parameter, die die unbewussten Prozesse repräsentieren. Der Parameter u_F soll den »begünstigenden« Einfluss erhöhter perzeptueller Flüssigkeit darstellen, der das Erkennen des Worts als »Wort« erleichtert. Der Parameter u_I soll den »interferierenden«

Einfluss erhöhter perzeptueller Flüssigkeit darstellen, der das Erkennen des Anagramms behindert. Er stellt die Wahrscheinlichkeit dar, dass Anagramme, die aus dargebotenen Wörtern gebildet wurden, fälschlicherweise als »Wort« bezeichnet werden. Die Validierung dieser Parameter stellt zweifellos die größte Herausforderung dar, da sich die Frage stellt, welche experimentellen Manipulationen geeignet sind. In der Literatur wird kaum über Experimente berichtet, die direkt auf eine Manipulation der unbewussten Prozesse abzielen. Die im Rahmen der Prozess-Dissoziations-Prozedur durchgeführten Untersuchungen können zwar oftmals belegen, dass Variationen der Aufmerksamkeit oder der Verarbeitungstiefe einen Einfluss auf bewusste Prozesse haben und die unbewussten Prozesse nicht tangieren. Untersuchungen, die *allein* die unbewussten Prozesse verändern und keinen Einfluss auf die bewussten Prozesse haben, gibt es innerhalb der Prozess-Dissoziations-Forschung meines Wissens nicht.

Unser psychologisch relevantes Hintergrundwissen, das zur Validierung der *u*-Parameter herangezogen werden kann, stammt daher vorwiegend aus Ergebnissen zum Dissoziationsparadigma (Vergleich von direkt erfassten mit indirekt erfassten Gedächtnisleistungen). Die im Rahmen dieser Logik durchgeführten Experimente belegen, dass indirekte Tests, da sie in der Regel eher datengesteuert sind, stärker von Manipulationen beeinflusst werden, die die Oberflächenmerkmale der »Daten« in der Enkodier- und Abrufsituation betreffen, als direkte Tests wie Reproduktion und Rekognition, die eher konzeptuelle Gedächtnisleistungen erfassen (Challis u. Sidhu, 1993; Light et al., 1992). Nimmt man an, dass Oberflächenmerkmale einen wesentlichen Einfluss auf unbewusste Prozesse haben, dann sollte ein Wechsel der Darbietungsmodalität von der Lern- zur Testphase (z. B. auditiv/visuell) im Vergleich zu einem Nicht-Wechsel (z. B. visuell/visuell) eine geringere Ausprägung der *u*-Parameter zur Folge haben. Die Wahrscheinlichkeit kontrollierter Prozesse *k* sowie die der Basisrateprozesse *b* sollte von einem Modalitätswechsel nicht berührt werden. Wir gehen davon aus, dass sich eine derartige experimentelle Manipulation auf die Parameter u_I und u_F in gleicher Richtung auswirkt.

Versuchspersonen und Design. Der Modalitätswechsel wurde sowohl für eine Präsentationsdauer der Testitems von 56 ms als auch für eine Präsentationsdauer der Testitems von 42 ms realisiert. An der Untersuchung nahmen 64 studentische Versuchspersonen teil, die per Zufall einer der vier Bedingungen (zwei [Wechsel vs. Nicht-Wechsel] · zwei [56 vs. 42 ms]) zugeteilt wurden. Sie wurden entweder für die Teilnahme mit 5 DM bezahlt oder erhielten eine Vpn-Bescheinigung. Die Bedingungen mit je 16 Versuchspersonen wurden interindividuell variiert.

Material. Das Itemmaterial entsprach dem der beiden zuvor dargestellten Untersuchungen. Alle 64 Wörter wurden aufgenommen und digitalisiert gespeichert, was eine akustische Darbietung in randomisierter Form analog zur visuellen Präsentation ermöglicht.

Durchführung. Die Durchführung der Bedingungen ohne Modalitätswechsel (VV: Studierphase visuell / Testphase visuell) entsprach den 42- und 56-ms-Bedingungen des Experiments 1. In den Bedingungen mit Modalitätswechsel (AV: Studierphase auditiv / Testphase visuell) wurden 32 Wörter per Zufall aus dem Pool der 64 Wörter ausgewählt und in der Studierphase den Probanden über Kopfhörer auditiv präsentiert. Die Gesamtzeit der auditiven Präsentation entsprach der der visuellen Darbietung. Die Testphasen beider Bedingungen (Wechsel/Nicht-Wechsel) unterschieden sich nicht: Wörter und Nonwörter wurden visuell präsentiert.

Ergebnisse. Für die in diesem Experiment gegebene Anzahl von $n = 7168$ Beobachtungen (128 [Antworten je Vp] · 56 [Anzahl der Vpn]) ergab eine Kompromiss-Power-Analyse zur Aufdeckung kleiner Effekte ($w = 0.1$) ein $\alpha = \beta < .0001$ (somit eine Power $1\text{-}\beta > .999$) und die kritischen Werte $G^2_{Krit} = 20.7$ bei $df = 2$ und $G^2_{Krit} = 24.1$ bei $df = 4$.[2]

Tabelle 4 zeigt die mittels multinomialer Modellierung erzielten Parameterschätzungen bei Gleichsetzung der Parameter k und b über die beiden Modalitätsbedingungen.

Das Modell passt mit $G^2 = 11.16$ bei $df = 4$ gut zu den Daten. Deskriptiv fielen die Schätzer für die unbewussten Prozesse u_F und u_I sowohl in der 42- als auch in der 56-ms-Bedingung gemäß der Erwartungen aus: Ein

Tabelle 4: Parameterschätzungen und 95 %-Konfidenzintervalle für Experiment 3. Modalität AV bedeutet: Studierphase Akustisch, Testphase Visuell (Wechsel); Modalität VV bedeutet: Studierphase Visuell, Testphase Visuell (kein Wechsel)

Präsentations-Dauer in ms	Modalität	k	b	u_I	u_F
56	AV	.51 [.48 .54]	.58 [.55 .61]	00 [–.17 .17]	.45 [.33 .58]
56	VV			.05 [–.14 .24]	.54 [.42 .65]
42	AV	.28 [.24 .32]	.49 [.46 .52]	.00 [–.13 .14]	.06 [–.07 .20]
42	VV			.10 [–.04 .23]	.19 [.08 . 31]

2 Acht Versuchspersonen mussten aus der Datenanalyse nachträglich ausgeschlossen werden, da sie keinerlei Diskriminationsleistung zeigten.

Wechsel der Darbietungsmodalität von der Lern- zur Testphase führte zu einer geringeren Ausprägung der Schätzer für die unbewussten Prozesse u_I und u_F als ein Nicht-Wechsel. Die Gleichsetzung der Parameter u_F und u_I über die unterschiedlichen Modalitätsbedingungen AV und VV sollte zu einer bedeutsamen Verschlechterung der Modellanpassung führen. Tatsächlich führte diese Restriktion der Parameter u_I und u_F über die beiden Bedingungen zu keiner signifikanten Verschlechterung der Modellanpassung ($G^2 = 1.19$ bzw. $G^2 = 14.06$; $df = 2$). Beim Vergleich der Präsentationszeiten zeigte sich analog zu Experiment 1 eine Verbesserung der Diskriminationsleistung (Parameter k), wenn die Items in der Testphase länger dargeboten wurden ($G^2 = 81.6$; $df = 2$). Anders als erwartet verringerte eine höhere Präsentationsgeschwindigkeit die Parameterschätzung für die unbewussten Prozesse in der Facilitationsbedingung (u_F) signifikant ($G^2 = 65.07$; $df = 2$).

Der Schätzer für die Ratetendenz b bleibt wie erwartet von der Variation der Präsentationsdauer unberührt ($G^2 = 7.3$; $df = 2$). Entsprechend den Erwartungen bleibt auch der Parameter u_I unbeeinflusst von der Präsentationsdauer ($G^2 = 15.6$; $df = 2$).

Die Ergebnisse liefern empirische Evidenz für die Annahme, dass ein Wechsel der Darbietungsmodalität die Schätzung der unbewussten Prozesse u_I/u_F nur geringfügig beeinflusste. Stattdessen ließ sich ein signifikanter Effekt der Präsentationsdauer auf die unbewussten Prozesse u_F aufzeigen.

Diskussion der Ergebnisse

Das Ziel der berichteten Untersuchungen war die systematische Validierung der Parameter des multinomialen PD+-Modells für lexikalische Entscheidungen, das in der Vergangenheit schon häufig angewendet wurde. Spezielle experimentelle Manipulationen, deren Effekte als bekannt vorausgesetzt werden können, sollten in den jeweiligen Experimenten nur auf *einen* Parameter des Modells Einfluss nehmen und die anderen Parameter weitgehend unbeeinflusst lassen. Um zu einer Bewertung zu kommen, inwiefern dieses Vorhaben geglückt ist, sollen nun die Aspekte, die für und gegen die Validität des Modells sprechen, zusammenfassend betrachtet werden:

Für die Validität des Parameters k spricht, dass sich die Variation der Darbietungsdauer in den Experimenten 1 und 3, wie erwartet auf den Parameter für das bewusste Diskriminieren k auswirkte. Eine längere Präsentationszeit der Items in der Testphase hatte eine bessere Diskriminierbarkeit von Wörtern und Nonwörtern zur Folge als eine kürzere Präsenta-

tionsdauer. Die unterschiedlichen Präsentationszeiten hatten in Experiment 1 keinen signifikanten Einfluss auf die übrigen Parameter, sodass man sagen kann, dass k ein Indikator für die Diskriminierbarkeit ist. Es fanden sich jedoch Hinweise darauf, dass eine Variation der Darbietungsdauer in der Testphase, zumindest tendenziell, auch die unbewussten Prozesse, die sich im Parameter u_F niederschlagen, begünstigt. Dieser Befund wird weiter unten diskutiert.

Experiment 2 belegt, dass der Parameter b die Antworttendenz repräsentiert. Die Manipulation der Antworttendenz über unterschiedliche Auszahlungsbedingungen hatte einen signifikanten Einfluss auf die Schätzung des Parameters b, gleichzeitig wurden die Parameterschätzungen für die bewussten Prozesse (k) und unbewussten Prozesse (u_F, u_I) in diesem Experiment nicht tangiert. Auffällig ist, dass in der neutralen Bedingung dieser Untersuchung die Parameterschätzungen für k und u_I, im Vergleich zur 42-ms-Bedingung des ersten Experiments, relativ gering ausfielen. Diese beiden Parameterschätzungen waren auch in den konservativen und liberalen Bedingungen relativ schwach ausgeprägt. Worauf könnte die relativ geringe Ausprägung der Schätzung des Diskriminationsparameters k in dieser Untersuchung zurückgeführt werden? Denkbar sind zwei Ursachen: Einerseits könnte das Bestreben der Versuchspersonen, eine möglichst hohe Punktzahl mit möglichst geringem Aufwand zu erreichen, dafür verantwortlich sein: Der Reaktionsbewertung wird mehr Aufmerksamkeit gewidmet als der Wahrnehmbarkeit des Reizmaterials. Eine andere Ursache könnte darin liegen, dass die visuelle Rückmeldung des Punktestands am unteren Bildschirmrand nach Abgabe der Antwort, die Aufmerksamkeit von der Bildschirmmitte abzog, und der damit verbundene Blickwechsel die Diskriminationsleistung reduzierte. Eine Entscheidung darüber, worauf die geringe Diskriminationsleistung zurückzuführen ist, kann nur durch weiterführende Experimente getroffen werden.

Ziel des dritten Experiments war die Validierung der u-Parameter, die die unbewussten Prozesse reflektieren sollen. Ausgehend von Befunden, die im Rahmen der Dissoziationslogik gewonnen wurden, gingen wir davon aus, dass die unbewussten Prozesse eher datengesteuerter Natur sind und somit stärker von Veränderungen der Oberflächenmerkmale in der Enkodier- und Abrufsituation beeinträchtigt werden als die bewussten Prozesse. Daher wurde eine Untersuchung, in der ein Modalitätswechsel zwischen Enkodier- und Testphase vorgesehen war, konzipiert. Das Experiment wurde mit den Präsentationszeiten von 56 ms und 42 ms durchgeführt. Das Ergebnismuster lieferte keinen zweifelsfreien Beleg für die Validität der Parameter für unbewusste Prozesse: Zwar beeinflusste der Wechsel der Dar-

bietungsmodalität weder die Schätzer für die Antworttendenz noch die Schätzer für die bewussten Diskriminationsprozesse k. Gleichzeitig beeinflusste der Modalitätswechsel zumindest tendenziell die Ausprägung der Parameterschätzungen u_I und u_F wie prognostiziert: In beiden Zeitbedingungen fielen die Parameter u_I und u_F geringer aus, wenn die Modalität von der Lern- zur Testphase wechselte, als in der Bedingung ohne Wechsel. Allerdings waren diese Unterschiede nicht statistisch signifikant. Man muss daher konstatieren, dass der Versuch, die unbewussten Prozesse zu manipulieren, nicht vollends geglückt ist. Dies mag damit zusammenhängen, dass entweder die Idee des Modalitätswechsels zur Beeinflussung unbewusster Prozesse, die dem Hintergrundwissen der Dissoziationslogik entstammte, ungeeignet war, oder dass die Manipulation einfach nur zu schwach war, um den Effekt deutlicher hervor treten zu lassen. In künftigen Untersuchungen sollte man daher weiter der Frage nachgehen, mit welchen experimentellen Manipulationen tatsächlich allein die unbewussten Prozesse zu manipulieren sind. Möglicherweise müsste man hier auf ganz andere Techniken wie Hypnose oder Ähnliches zurückgreifen.

Auffällig und anders als erwartet verhielt sich der Parameter u_F in Experiment 3. Vergleicht man die Ergebnisse der Parameterschätzungen über die beiden Präsentationszeiten, so zeigte sich, dass die Schätzungen für den Parameter u_F in der 56-ms-Bedingung signifikant höher ausfielen als in der 42-ms-Bedingung. Angesichts dieser Resultate scheinen auch die Schätzungen von u_F in Experiment 1 kein Zufallsprodukt zu sein: Die Parameterschätzungen für u_F variieren offenbar systematisch mit der Präsentationsdauer. Über die Erklärung dieses Befunds lässt sich nur spekulieren. Möglicherweise beeinflusst eine längere Präsentationszeit nicht nur die bewussten Prozesse, sondern hat auch einen Effekt auf unbewusste Prozesse im Sinne erhöhter perzeptueller Flüssigkeit. Dieser Einfluss ist aber nur dann wirksam, wenn es sich um Wörter und nicht wenn es sich um Nonwörter handelt. Anders ausgedrückt: Die Präsentationszeit beeinflusst bewusste Prozesse nicht selektiv, sondern übt zwei Effekte gleichzeitig aus.

Die Zusammenschau aller Ergebnisse liefert kein eindeutiges Bild: Vieles spricht für die Validität der Modellparameter, einiges ist jedoch weiterhin klärungsbedürftig. Auch die geringe Ausprägung der Parameter für unbewusste Prozesse u_I in der Interferenzbedingung, zumindest in den Experimenten 2 und 3, war nicht erwartet worden. Die Präsentation eines Worts in der Lernphase hatte kaum oder überhaupt nicht dazu geführt, dass das daraus gebildete Nonwort häufiger fälschlicherweise als Wort bezeichnet wurde als ein »neues« Nonwort.

Tabelle 5: Parameterschätzungen und 95 %-Konfidenzintervalle für Experiment 4. Die Schätzer für das unähnliche Wortmaterial sind aus Experiment 1 übernommen

Präsentationszeit (ms)	Wortmaterial	k	b	u_I	u_F
42	***Ähnlich***	.19 [.13 .25]	.59 [.56 .63]	.23 [.05 .41]	.13 [–.03 .28]
42	Unähnlich	.30 [.24 .35]	.48 [.44 .52]	.11 [–.04 .27]	.27 [.14 .40]
56	***Ähnlich***	.29 [.24 .35]	.59 [.55 .63]	.42 [.21 .63]	.32 [.17 .47]
56	Unähnlich	.40 [.34 .45]	.43 [.39 .48]	.20 [.04 .36]	.47 [.36 .59]

Um dieses Problem wenigstens im Ansatz zu lösen, entschlossen wir uns zu einer weiteren Untersuchung. Basierend auf der Vermutung, dass die verwendeten Nonwörter den dazugehörigen Wörtern nicht ähnlich genug waren, um eine Interferenz zu bewirken, variierten wir das Material: Anstelle der in den Experimenten 1–3 dargebotenen relativ unähnlichen Nonwörter (z. B. Dozent – Zedont) wurden ähnliche Nonwörter (z. B. Dozent – Dezont) in der Testphase präsentiert. Ansonsten entsprach der Versuch dem ersten Experiment, wobei zwei Präsentationszeiten (42 ms und 56 ms) realisiert wurden. Insgesamt wurden 24 Vpn getestet, die per Zufall einer der zwei Bedingungen (42 vs. 56 ms) zugeteilt wurden. Tabelle 5 zeigt die mittels multinomialer Modellierung erzielten Parameterschätzungen. Zum Vergleich sind dort auch noch einmal die entsprechenden Ergebnisse aus Experiment 1 (unähnliches Wortmaterial) aufgeführt.

Beim Vergleich der Wortmaterialbedingungen fanden sich in beiden Präsentationszeitbedingungen die erwarteten Effekte: Eine erhöhte Ähnlichkeit der Nonwörter zu den Wörtern führte zu einer verstärkten Nutzung unbewusster Prozesse. Die Parameterschätzungen für die unbewussten Prozesse in der Interferenzbedingung u_I fielen bei hoher Ähnlichkeit von Wörtern und Nonwörtern deutlich höher aus als bei geringer Ähnlichkeit. Gleichzeitig verringerte sich erwartungsgemäß die Schätzung des Diskriminations-Parameters k, was mit einer stärkeren Verwechslungsmöglichkeit von Nonwörtern und Wörtern zu begründen ist. Die Parameterschätzungen für die Diskriminationsleistung k waren in den Bedingungen mit hoher Ähnlichkeit von Wörtern und Nonwörtern deutlich niedriger ausgeprägt als in den Bedingungen mit geringer Ähnlichkeit. Mit der höheren Ähnlichkeit der Nonwörter zu den Wörtern erhöhte sich zudem die Tendenz, »Wort« zu raten.

Als Fazit lässt sich festhalten, dass sich die meisten experimentellen Manipulationen zur Validierung des Modells tatsächlich nur auf den sie betreffenden Parameter des Modells auswirkten und zugleich keinen oder nur geringen Einfluss auf die übrigen Parameterschätzungen hatten. Die Ergebnisse

lassen meines Erachtens den Schluss zu, dass der Parameter k die Diskriminationsleistung und der Parameter b die Antworttendenz repräsentiert. Um die Validität des Modells abschließend bewerten zu können, ist es allerdings erforderlich, den Parametern u_I und u_E, die die unbewussten Prozesse repräsentieren sollen, noch mehr Aufmerksamkeit zu widmen. Wünschenswert wären experimentelle Manipulationen, die einen deutlichen Einfluss auf die unbewussten Prozesse haben und die anderen Parameter nicht tangieren. Dazu muss zunächst sichergestellt sein, dass die unbewussten Prozesse, sowohl in der Interferenz- als auch in der Facilitationsbedingung überhaupt eine gewisse Ausprägung aufweisen. Ein erster Schritt dahin ist die Wahl möglichst ähnlicher Nonwörter. Die Wahl ähnlicher Nonwörter begünstigte die Ausprägung der Parameterschätzungen für u_I in der Interferenzbedingung. Daher empfiehlt es sich, in weiteren Untersuchungen nur mit ähnlichen Nonwörtern zu arbeiten.

Literatur

Baschek, I.-L., Bredenkamp, J., Oehrle, B., Wippich, W. (1977). Bestimmung der Bildhaftigkeit (I), Konkretheit (C) und der Bedeutungshaltigkeit (m') von 800 Substantiven. *Zeitschrift für experimentelle und angewandte Psychologie, 24,* 353–396.

Batchelder, W.H., Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Bochenski, I.M. (1954). *Die zeitgenössischen Denkmethoden.* München: Lehnen.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1998). Die Erfassung unbewußter Verarbeitungsprozesse. In K.C. Klauer u. H. Westmeyer (Hg.), *Psychologische Methoden soziale Prozesse* (S. 30–46). Lengerich: Pabst.

Bröder, A. (2000). A methodological comment on behavioral decision research. *Psychologische Beiträge, 42,* 645–662.

Buchner, A., Erdfelder, E., Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process-dissociation framework. *Journal of Experimental Psychology: General, 124,* 137–160.

Challis, B.H., Sidhu, R. (1993). Dissociative effect of massed repetition on implicit and explicit measures of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 115–127.

Erdfelder, E., Faul, F., Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers, 28,* 1–11.

Erdfelder, E., Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127,* 83–96.

Erdfelder, E., Buchner, A. (2003). Prozess-Dissoziationsprozedur: Quo vadis? *Zeitschrift für Psychologie, 211,* 17–25.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

Krüger, T. (1999). *Die Erfassung bewußter und unbewußter Gedächtnisprozesse: Die Prozeß-Dissoziations-Prozedur. Probleme und Perspektiven einer neuen Methode.* Lengerich: Pabst.

Light, L.L., LaVoie, D., Valencia-Laver, D., Albertson-Owens, S.A., Maed, G. (1992). Direct and indirect measures of memory for modality in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1284–1297.

Ott, R., Curio, I., Scholz, O.B. (2000). Implicit memory for auditorily presented threatening stimuli: A process-dissociation approach. *Perceptual and Motor Skills, 90,* 131–146.

Perrig, W. (1992). Unbewusste Prozesse im Aufbau menschlicher Erkenntnis: Experimentelle Möglichkeiten ihrer Funktionsbestimmung. In U. Gerhard (Hg.), *Psychologische Erkenntnisse zwischen Philosophie und Empirie* (S. 159–182). Bern: Huber.

Riefer, D.M., Batchelder, W.H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318–339.

Scholz, O.B. (1997). Das Unbewußte als Informationsverarbeitungsprozess und seine Bedeutung für die moderne Klinische Psychologie. *Zeitschrift für Psychologie, 205,* 327–356.

Scholz, O.B., Ott, R., Müller-Sinik, K. (1997). Beziehung zwischen Parametern der impliziten und expliziten Informationsverarbeitung und psychosomatischen Selbstberichtsmaßen. *Verhaltenstherapie, 7,* 217–225.

Vaterrodt-Plünnecke, B. (1994). Multinomiale Modellierung impliziter Gedächtnisprozesse: Ein alternativer Ansatz. *Zeitschrift für Experimentelle und Angewandte Psychologie, 41,* 295–314.

Vaterrodt-Plünnecke, B., Krüger, T., Bredenkamp, J. (2002). Process dissociation procedure: A testable model for considering assumptions about the stochastic relation between consciously controlled and automatic processes. *Experimental Psychology, 49,* 3–26.

Vaterrodt-Plünnecke, B., Krüger, T., Gerdes, H., Bredenkamp, J. (1996). Prozeß-Dissoziations-Prozedur: Prüfbare Meßmodelle zur Erfassung von kontrollierten, automatischen und Antworttendenz-Prozessen. *Zeitschrift für Experimentelle und Angewandte Psychologie, 43,* 483–519.

Westermann, R. (1987). Wissenschaftstheoretische Grundlagen der experimentellen Psychologie. In G. Lüer (Hg.), *Allgemeine experimentelle Psychologie* (S. 5–42). Stuttgart: Fischer.

Windmann, S., Krüger, T. (1998). Subconscious detection of threat as reflected by an enhanced response bias. *Consciousness and Cognition, 7,* 603–633.

Zaunbauer, A., Bredenkamp, J. (2002). *Multinomiale Modellierung der lexikalischen Entscheidungsaufgabe nach supraliminalem und subliminalem Priming mit Adjektiven unterschiedlicher Valenz.* Berlin: 43. Kongress der Deutschen Gesellschaft für Psychologie.

Karl Christoph Klauer

Deduktivistische Ansätze beim deduktiven Schlussfolgern

Zusammenfassung

Deduktivistisches Vorgehen erfordert eine möglichst lückenlose Ableitung von zumindest statistisch prüfbaren Vorhersagen für beobachtbare Daten aus den Theorien. Das verlangt auch Annahmen darüber, wie Versuchspersonen die Prämissen eines Denkproblems im Einzelnen verstehen, und wie sie sich verhalten, wenn sie, wie bei Denkproblemen allfällig möglich, überfordert sind und die korrekte Antwort nicht wissen oder Unsicherheit darüber empfinden. Im vorliegenden Kapitel wird dargelegt, dass und wie Theorien des deduktiven Schlussfolgerns um Annahmen über unterschiedliche Interpretationsmöglichkeiten und heuristische Antworttendenzen explizit ergänzt werden können, um deduktivistisches Theorientesten zu ermöglichen. Möglichkeiten der experimentellen Kontrolle werden gleichfalls angerissen.

Einleitung

Ein traditionelles Thema der Denkpsychologie ist das deduktive Schlussfolgern. Die Forschung dazu ist einerseits durch einige wenige Aufgaben wie die syllogistischen Aufgaben, die Auswahlaufgabe von Wason und Aufgaben zum aussagenlogischen Schlussfolgern dominiert, andererseits durch Debatten um die Angemessenheit übergreifender Theorien wie die Theorie Mentaler Modelle, regelbasierte Theorien oder in jüngerer Zeit informationstheoretisch inspirierte Ansätze. Angesichts systematischer, zum Teil überraschender Fehler im menschlichen Schlussfolgern wird auch um die Frage gestritten, ob das beobachtete Verhalten und die dazu führenden Denkprozesse als rational bezeichnet werden dürfen. In diesem Kapitel sollen einige der Hürden beleuchtet werden, die sich dem stringenten Ableiten von prüfbaren Hypothesen aus den Theorien in den Weg stellen, sowie Möglichkeiten erläutert werden, solche Hürden zu überwinden. Dazu sind einige Vorarbeiten notwendig. Im Folgenden werden zunächst die oben ge-

nannten Aufgabentypen skizziert. Im Anschluss daran sollen die wichtigsten Theorien kurz vorgestellt werden.

Aufgaben

Syllogismen

Die älteste Klasse von Aufgaben, die in der Psychologie deduktiven Schlussfolgerns systematisch untersucht wird, bedient sich der klassischen Syllogismen. Syllogismen sind einfache Schlussfiguren, die aus zwei Prämissen und einer Schlussfolgerung bestehen. Ein Beispiel dafür ist der folgende Schluss:

- Alle Einwohner von Buxtehude sind Beamte.
- Alle Beamten sind pensionsberechtigt.
- Also sind alle Einwohner von Buxtehude pensionsberechtigt.

Syllogismen verwenden Prämissen und Schlussfolgerungen mit Quantoren. Auftretende Quantoren sind »Alle«, »Einige«, »Keine« und »Einige nicht«. Diese beziehen sich auf Elemente von Klassen A, B und C. Die logische Form des oben genannten Syllogismus lautet zum Beispiel: Alle A sind B. Alle B sind C. Also sind alle A C. Die Quantoren »Alle« und »Keine« werden manchmal universelle Quantoren genannt, da sie sich auf alle Elemente der zugehörigen Klasse beziehen, die Quantoren »Einige« und »Einige nicht« machen hingegen nur Aussagen über einzelne Mitglieder der zugehörigen Klassen, sie werden deswegen partikuläre Quantoren genannt. Die Bedeutung von »Alle« und »Einige« ist affirmativ, die von »Keine« und »Einige nicht« negierend. Die ersten beiden Quantoren werden daher positive, die letzten zwei negative genannt.

Syllogismen besitzen eine Figur, die durch die Abfolge der Terme A, B und C in Prämissen und Schlussfolgerung definiert ist. Beschränkt man sich auf Schlussfolgerungen mit der Reihenfolge A-C, so sind die vier möglichen Figuren für die beiden Prämissen: (1) Wie im Beispiel A-B, B-C, (2) A-B, C-B, (3) B-A, B-C und (4) B-A, C-B. Lässt man auch Schlussfolgerungen der Reihenfolge C-A zu, so ergeben sich sogar acht Figuren. Bei vier Quantoren, die den beiden Prämissen und der Schlussfolgerung zugeordnet werden können, resultieren insgesamt 512 mögliche Syllogismen, von denen die meisten keine logisch gültigen Schlüsse darstellen (Evans et al., 1993, Kap. 7).

Aufgaben, die sich der Syllogismen bedienen, unterscheiden sich im Antwortformat und in den Instruktionen. Häufig wurde die Bewertungsaufgabe verwendet, bei der Versuchspersonen angeben sollen, ob ein vorgelegter Syllogismus einen logisch gültigen Schluss darstellt oder nicht (z. B. Janis u. Frick, 1943). Bei der Produktionsaufgabe werden nur die Prämissen vorgelegt, und es wird darum gebeten, eine logisch gültige Schlussfolgerung zu ziehen, wenn eine existiert (z. B. Johnson-Laird u. Bara, 1984). Ein Kompromiss zwischen beiden Formaten stellen Aufgaben mit mehreren Antwortoptionen dar, bei denen zum Beispiel Schlussfolgerungen in der Reihenfolge A-C mit unterschiedlichen Quantoren angeboten werden und die richtige oder die richtigen auszuwählen sind (z. B. Dickstein, 1975). Weniger häufig sind die Anforderungen, die Prämissen und die darin genannten Relationen zwischen A, B und C bildlich darzustellen oder das eigene Denken beim Bearbeiten der Aufgabe laut zu verbalisieren (Bucciarelli u. Johnson-Laird, 1999).

Aufgaben zum Aussagenlogischen Schlussfolgern

Eine verwandte Klasse von Aufgaben beruht auf einfachen Schlüssen aus dem Bereich der elementaren Aussagenlogik. Ein Beispiel ist der folgende Schluss:

- Es gibt ein Dreieck an der Tafel, oder es gibt einen Kreis an der Tafel.
- Es gibt keinen Kreis an der Tafel.
- Also gibt es ein Dreieck an der Tafel.

Wieder gibt es zwei Prämissen und eine Schlussfolgerung. Die Majorprämisse verbindet zwei Aussagen p und q oder deren Negationen durch einen Junktor, im Beispiel wurde der Junktor »oder« verwendet. Die Minorprämisse ist die Aussage p beziehungsweise q oder deren Negation. Die Schlussfolgerung ist die jeweils andere Aussage oder deren Negation. Als Junktoren wurden die folgenden (und weitere) verwendet: »und«, »oder«, »entweder-oder«, »wenn-dann«, »genau dann, wenn«, »nur dann, wenn«. Die Antwortformate sind ähnlich wie bei syllogistischen Aufgaben: In einer Produktionsvariante werden nur die beiden Prämissen vorgelegt, und es wird darum gebeten, einen logisch gültigen Schluss zu ziehen, wenn es einen (nicht-trivialen) gibt. In der Bewertungsaufgabe wird eine (oder mehrere) Schlussfolgerung präsentiert, und es ist zu entscheiden, ob sich diese logisch zwingend aus den Prämissen ergibt oder nicht.

Die Auswahlaufgabe von Wason

Die Auswahlaufgabe wurde von Wason (1966) vorgestellt und wird bis heute intensiv untersucht. Ihre Faszination für Denkpsychologen ergibt sich daraus, dass die vermeintlich einfache Aufgabe systematisch falsch gelöst wird, sowie aus der Tatsache, dass sehr starke Einflüsse unterschiedlicher inhaltlicher Einbettungen beobachtet werden.

Die Aufgabe gibt es in vielen Varianten. Eine klassische Variante legt eine konditionale Regel vor, die für eine Menge von Karten formuliert ist. Jede Karte hat auf der einen Seite einen Buchstaben, auf der anderen eine Zahl. Eine typische Regel lautet: »Wenn auf der Buchstabenseite ein Vokal ist, dann ist auf der Zahlenseite eine gerade Zahl«. Der Versuchsperson werden dann vier Karten vorgelegt, von denen jeweils nur eine Seite zu sehen ist. Die sichtbaren Seiten präsentieren einen Vokal, einen Konsonanten, eine gerade Zahl und eine ungerade Zahl. Aufgabe der Versuchsperson ist es, die Karten auszuwählen, die umgedreht werden müssen, um zu entscheiden, ob die Regel gilt oder nicht. Als richtige Wahl gelten die Wahl der Karte mit dem Vokal und die Wahl der Karte mit der ungeraden Zahl auf den sichtbaren Seiten. Nur diese beide Karten können eine verletzende Instanz, also eine Karte mit einem Vokal auf der Buchstabenseite und einer ungeraden Zahl auf der Zahlenseite, darstellen. Die weitaus häufigsten Wahlen von Versuchspersonen sind die Karte mit dem Vokal alleine zu wählen, oder die Karte mit dem Vokal und die Karte mit der geraden Zahl.

Theorien

Umfangreiche empirische Untersuchungen mit diesen und anderen Aufgaben haben eine Reihe von replizierbaren empirischen Phänomenen aufgezeigt, zu deren Erklärung zum Teil spezifische, auf einzelne Aufgabenklassen zugeschnittene Theorien, aber auch Aufgaben übergreifende Theorien vorgeschlagen wurden. Aufgaben übergreifende theoretische Ansätze sind die Theorie Mentaler Modelle, regelbasierte Theorien und informationstheoretische Modelle.

Die Theorie Mentaler Modelle

Johnson-Laird (1983) hat die Theorie aufgestellt, dass Denken im Konstruieren, Manipulieren und Bewerten von mentalen Modellen besteht. Mentale Modelle sind mentale Repräsentationen von Prämissen, die denen ähneln, die beim Wahrnehmen gebildet werden. Die Theorie Mentaler Modelle postuliert drei Schritte menschlichen Schlussfolgerns:

1. Zunächst wird von den Prämissen ausgehend ein anfängliches mentales Modell gebildet, das mögliche Zustände einer Welt repräsentiert, in der die Prämissen erfüllt sind.
2. Dann wird versuchsweise eine Schlussfolgerung konstruiert, indem eine Aussage entdeckt wird, die in dem anfänglichen Modell gilt und die informativ ist, das heißt nicht einfach eine Prämisse wiederholt oder eine triviale Schlussfolgerung darstellt.
3. Schließlich wird versucht, Gegenbeispiele zu finden. Das sind Modelle, die mit den Prämissen konsistent sind, in denen aber die versuchsweise Schlussfolgerung nicht gilt. Kann kein Gegenbeispiel gefunden werden, so wird die Schlussfolgerung als gültig akzeptiert.

Im Kontext der Aussagenlogik zum Beispiel werden für die Prämisse »Wenn p, dann q« zwei anfängliche Modelle wie folgt gebildet (Johnson-Laird et al., 1992):

[p] q
...

wobei »...« das so genannte implizite Modell meint, das anzeigt, dass weitere, noch nicht explizit modellierte Situationen außer »[p] q« existieren. Die eckigen Klammern um p bedeuten, dass Situationen mit p vollständig durch den gegebenen Satz von Modellen repräsentiert sind; p tritt also nur in Verbindung mit q auf. In diesem Fall kann von der Minorprämisse p sicher auf q geschlossen werden. Bevor jedoch ein Schluss von einer Minorprämisse, die nicht vollständig repräsentiert wird (zum Beispiel von »q« oder »nicht q«) gezogen wird, muss das implizite Modell weiter expliziert werden.

Die Theorie wurde vor allem anhand ihrer Vorhersagen für die Schwierigkeit unterschiedlicher Schlüsse bewertet. Zum Beispiel wird der so genannte »modus ponens«-Schluss von »Wenn p, dann q« und »p« auf »q« wesentlich öfter korrekt gezogen als der ebenfalls gültige »modus tollens«-Schluss von »Wenn p, dann q« und »Nicht q« auf »Nicht p«. Die Schwierigkeit eines Schlusses hängt der Theorie Mentaler Modelle zufolge wesent-

lich von der Anzahl (expliziter) Modelle ab, die für das Entdecken und Prüfen einer versuchsweisen Schlussfolgerung gebildet werden müssen.

Regelbasierte Theorien

In regelbasierten Theorien (Braine et al., 1995; Braine et al., 1984; Braine u. Rumain, 1983; Rips, 1994) wird angenommen, dass Menschen über ein bestimmtes Repertoire an abstrakten Inferenzregeln wie die Modus-ponens-Regel verfügen, die sie nach bestimmten Auswahl- und Prioritätskriterien auf die Prämissen anwenden. Genauer gesagt wird angenommen, dass sich das Denken in drei Schritten vollzieht: In einem ersten Schritt werden die Prämissen in ein syntaktisches (zum Beispiel propositionales) Format übersetzt, das ihre logische Struktur widerspiegelt und auf das die Inferenzregeln zugreifen können. Dann werden die Inferenzregeln angewendet, um eine Schlussfolgerung abzuleiten. Dabei werden Strategien zur Kombination von Inferenzregeln in Ableitungsketten eingesetzt, wobei Zwischenziele gebildet und neue Prämissen generiert werden können. Im Erfolgsfall muss eine gewonnene Schlussfolgerung in das zumeist umgangssprachliche Format der Prämissen zurückübersetzt werden.

Diesen Theorien zufolge hängt die Schwierigkeit eines gegebenen Schlusses vor allem davon ab, wie kompliziert die Ableitungskette ist, die vermöge der verfügbaren Inferenzregeln von den Prämissen zu der Schlussfolgerung führt. Eine weitere Quelle von Fehlern ist mangelnde Verfügbarkeit benötigter Inferenzregeln.

Informationstheoretische Ansätze

Informationstheoretische Ansätze wurden von Oaksford und Chater (1994) sowie Chater und Oaksford (1999) vorgeschlagen. Denken wird als Entscheiden unter Unsicherheit aufgefasst und anstelle der klassischen Logik werden statistische Theorien zur Begründung optimalen Verhaltens herangezogen. Der Ansatz lässt sich anhand der Wasonschen Auswahlaufgabe verdeutlichen. Es wird angenommen, dass Menschen die Auswahlaufgabe als ein statistisches Entscheidungsproblem zwischen zwei probabilistischen Modellen auffassen. Das eine wird Abhängigkeitsmodell genannt, in ihm gilt die konditionale Regel, »Wenn p, dann q«, was zu einer stochastischen Abhängigkeit von p und q führt, da der Fall »p und nicht-q« nicht auftreten kann, seine Wahrscheinlichkeit also gleich Null ist, zumindest aber sehr viel

kleiner als aufgrund der Wahrscheinlichkeiten von p und der von nicht-q alleine zu erwarten wäre (Oaksford, 2002). Ein alternatives Modell spezifiziert den Fall, dass die Regel verletzt ist. Oaksford und Chater (1994) gehen dafür von einem Modell mit unabhängigen p- und q-Ereignissen aus. Darüber hinaus gibt es eine subjektive Ausgangswahrscheinlichkeit dafür, dass die Regel gilt, dass also das Abhängigkeitsmodell die Wirklichkeit zutreffend beschreibt.

Die Entscheidung zwischen den Modellen soll aufgrund von Daten gefällt werden. Bei der Auswahlaufgabe sind dies die Vorder- und Rückseiten der Karten, die zum Umdrehen ausgewählt werden. Solche Karten sollten ausgewählt werden, die für die Entscheidung möglichst nützlich sind, also möglichst aussagekräftige Daten versprechen. Welche Karten in diesem Sinne besonders aussagekräftig sind und also umgedreht werden sollten, beschreiben Oaksford und Chater (1994) mit einem statistischen Maß der Information. Die Theorie macht Aussagen darüber, wie subjektive Annahmen über die Wahrscheinlichkeit von p-Ereignissen und von q-Ereignissen sowie über die Wahrscheinlichkeit der Gültigkeit des Abhängigkeitsmodells die Wahlwahrscheinlichkeiten der unterschiedlichen Karten determinieren. Ähnliche probabilistische Ansätze gibt es mittlerweile auch für das syllogistische Schließen (Chater u. Oaksford, 1999).

Der Schluss von der Theorie auf die Daten

Kern des deduktivistischen Vorgehens zum Prüfen von Theorien ist das Ableiten von statistischen Hypothesen aus psychologischen (Bredenkamp, 1980, 1996). Das Ableiten von statistischen Hypothesen aus den skizzierten Theorien zum deduktiven Schließen ist nicht einfach, da wichtige Voraussetzungen unterspezifiziert sind oder zum Teil bewusst offen gelassen werden. Das führt in der Regel dazu, dass die psychologische Hypothese für die Gültigkeit der statistischen Hypothese weder notwendig, noch hinreichend ist. Sie ist nicht notwendig, weil je nachdem, mit welchen expliziten oder impliziten Hilfsannahmen die Lücke überbrückt wird, auch die Hilfsannahmen die statistische Hypothese bedingen können; sie ist nicht hinreichend, weil die offen gelassenen Prozesse die aufgrund der psychologischen Hypothese erwarteten Effekte konterkarieren können. Das reduziert sowohl die Strenge als auch die Fairness der Theorieprüfung (Erdfelder u. Bredenkamp, 1994).

Im Folgenden sollen diese Probleme anhand zweier wichtiger Lücken

verdeutlicht werden. Sie betreffen die Interpretation vorgegebener Aussagen, in der Regel also der Prämissen und Schlussfolgerungen, sowie Heuristiken, die unter Unsicherheit oder bei Überforderung zur Auswahl von Antworten in der Regel eingesetzt werden. Die Lücken betreffen demnach Eingabeprozesse, die dem Denken vorauslaufen, und Ausgabeprozesse, die zum Zuge kommen, wenn das vorgelegte Denkproblem nicht gelöst werden kann.

Interpretation

Die geschilderten Theorien konzentrieren sich vernünftigerweise auf den Prozess des Denkens selber. Sie erkennen aber an, dass die mentalen Repräsentationen, auf denen sich die Denkprozesse vollziehen, aus in der Regel sprachlich vorgegebenen Aussagen gewonnen werden und dass eine Übersetzung in das Format mentaler Repräsentationen erfolgen muss. Bei der Theorie Mentaler Modelle ist das Ergebnis der Übersetzung ein oder mehrere mentale Modelle, bei regelbasierten Theorien ist das Ergebnis ein abstraktes, hinsichtlich der logischen Form analysiertes Abbild der Prämissen, bei informationstheoretischen Ansätzen nimmt das Ergebnis die Form eines probabilistischen Modells an.

Es gibt viele Befunde, die darauf hindeuten, dass die sprachlichen Aussagen unterschiedlich übersetzt werden können. Besonders gut untersucht sind konditionale »Wenn-dann« Aussagen (Johnson-Laird u. Byrne, 2002), die eine Vielzahl sehr unterschiedlicher Interpretationen erlauben. Eine wichtige Zweideutigkeit ist zum Beispiel, ob die Aussage »Wenn p, dann q« eher konditional (p ist hinreichend für q) oder bikonditional gedeutet wird (p ist notwendig und hinreichend für q). Den zwei Möglichkeiten entsprechen unterschiedliche Repräsentationen als mentale Modelle, im regelbasierten Format oder als probabilistische Modelle. Ähnliche Interpretationsspielräume gibt es beim aussagenlogischen Schlussfolgern für die anderen Junktoren, zum Beispiel den disjunktiven, der umgangssprachlich oft mit »oder« umschrieben wird. Unterschiedliche Interpretationen spielen auch im Bereich der Syllogismen (Roberts et al., 2001) und der Auswahlaufgabe von Wason eine wichtige Rolle (Osman u. Laming, 2001).

Die gleichen Denkprozesse, auf unterschiedliche Ausgangsrepräsentationen angewendet, führen in vielen Fällen zu unterschiedlichen Ergebnissen und damit zu unterschiedlichen statistischen Hypothesen. Umgekehrt können Unterschiede, die in Denkprozessen begründet liegen, verwischt oder ins Gegenteil verkehrt werden, wenn unterschiedliche Interpretatio-

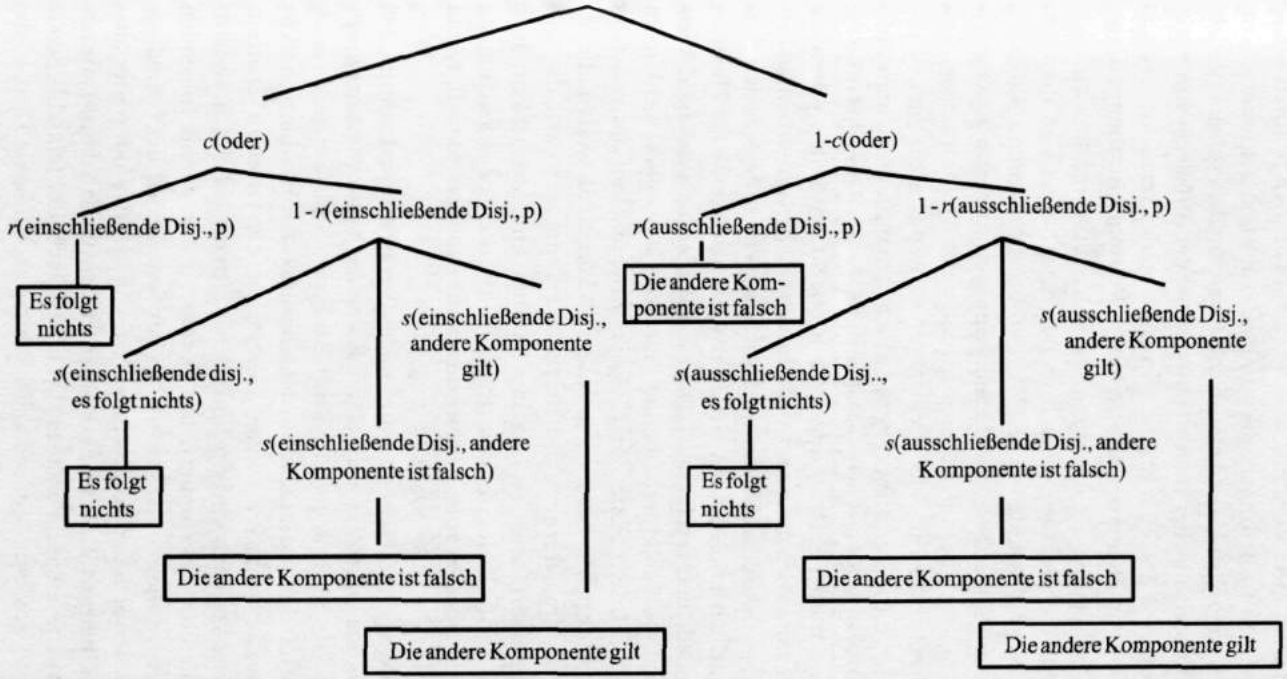


Abbildung 1: Multinomiales Verarbeitungsbaummodell für die Häufigkeiten der Antwortoptionen »q« (die andere Komponente gilt), »nicht-q« (die andere Komponente ist falsch) und »es folgt nichts« bei der Aufgabe mit Majorprämisse »p oder q« und Minorprämisse »p«. Disj. = Disjunktion; c = Wahrscheinlichkeit, mit der der Junktor »oder« im Sinne der einschließenden anstelle der ausschließenden Disjunktion interpretiert wird; r = Wahrscheinlichkeit, mit der der so interpretierte Schluss korrekt durch deduktives Denken bearbeitet wird; s = Wahrscheinlichkeiten, mit der die Antwortoptionen gewählt werden, wenn schlussfolgerndes Denken nicht zur Lösung führt.

nen möglich sind. Zwei Wege bieten sich an, um dennoch prüfbare statistische Hypothesen zu gewinnen, die zumindest mit höherer Wahrscheinlichkeit durch die psychologischen impliziert werden: die explizite Berücksichtigung unterschiedlicher Interpretationen bei der Ableitung von statistischen Hypothesen und eine genauere Instruktion der Versuchsperson, in der zum Beispiel im Zuge eines vorbereitenden Verständnistrainings die gewünschte Interpretation der Aussagen vereinbart und eingeübt wird.

Klauer und Oberauer (1995) beschritten in einer Studie zum aussagenlogischen Schlussfolgern den ersten Weg. Vorgegeben wurden Aufgaben, die wie oben geschildert, aus einer Majorprämisse mit einem aussagenlogischen Junktor, zum Beispiel »p oder q« und einer Minorprämisse, zum Beispiel p, bestehen. Aufgabe der Versuchsperson war es anzugeben, ob daraus etwas über die in der Minorprämisse nicht genannte elementare Aussage, im Beispiel q, folge, genauer gesagt, ob q, nicht-q, oder nichts über die Gültigkeit von q oder nicht-q folge. Die beobachteten Daten waren die Häufigkeiten, mit denen diese Antwortoptionen von den untersuchten Personen produziert wurden. Ziel der Untersuchung war es, die Schwierigkeiten unterschiedlicher Schlüsse empirisch zu bewerten, für die die Theorien mentaler Modelle und regelbasierte Theorien Vorhersagen machen. Allerdings sollte dafür die Interpretation der Prämissen kontrolliert werden: Im Beispiel führt die Interpretation der Majorprämisse im Sinne des ausschließenden »oder« dazu, die Antwort »nicht-q« als korrekt zu werten, die Interpretation im Sinne des einschließenden »oder« hingegen dazu, die Antwort »es folgt nichts für q« als korrekt zu werten. Eine Möglichkeit, diese Probleme zu kontrollieren, besteht in einer expliziten Berücksichtigung beider Interpretationen zum Beispiel durch ein explizites Modell, wie in Abbildung 1 für die Beispielaufgabe angegeben:

Die Abbildung stellt ein multinomiales Verarbeitungsbaummodell (Batchelder u. Riefer, 1999) für die Häufigkeiten der Antwortoptionen »q«, »nicht-q« und »es folgt nichts« der Beispielaufgabe dar. Die erste Verzweigung von oben steht für die beiden bedeutendsten Interpretationsmöglichkeiten des Junktors »oder«. Mit einer gewissen, aus den Daten zu schätzenden Wahrscheinlichkeit c wird der Junktor nämlich einschließlich, andernfalls ausschließlich interpretiert. Die Parameter der darauf folgenden Verzweigungen, r, messen die Leichtigkeit der Schlüsse auf der Grundlage der von der Versuchsperson gewählten einschließenden oder aber ausschließenden Interpretation. Beides sind unterschiedliche Schlüsse, die unterschiedlich schwierig sein können. Im linken Teilbaum (einschließende Interpretation) wird mit Wahrscheinlichkeit r(einschließende Disj., p) die dann korrekte Antwort, dass nichts folgt, erzielt, im rechten Teilbaum (aus-

schließende Interpretation) mit Wahrscheinlichkeit r(ausschließende Disj., p) als korrekte Antwort »nicht q« geschlussfolgert. Die weiteren Verzweigungen betreffen heuristische Prozesse, die laut Modell zum Zuge kommen, wenn die durch die r-Parameter erfassten Denkprozesse ohne ein definitives Ergebnis enden. Auf diese Komplikation gehe ich weiter unten noch genauer ein.

Das multinomiale Verarbeitungsbaummodell ist ein Analyseinstrument, das im vorliegenden Fall die unterschiedlichen Interpretationsmöglichkeiten explizit berücksichtigt und so Vorhersagen über die Struktur der Daten ermöglicht, die diesen Möglichkeiten Rechnung tragen. Klauer und Oberauer (1995) benutzten das Modell, um Vorhersagen der Theorie Mentaler Modelle über die Rangreihe der Schwierigkeiten verschiedener Schlüsse beim propositionalen Schlussfolgern zu prüfen. Die Vorhersagen beziehen sich auf die r-Parameter für die Leichtigkeit der Schlüsse bei einer gegebenen Interpretation. Im Großen und Ganzen ließen sich die Vorhersagen der Theorie Mentaler Modelle bestätigen. Ähnliche Modelle können im Prinzip für syllogistische Aufgaben oder die Auswahlaufgabe von Wason konstruiert werden, um den Schluss von Theorien auf statistische Hypothesen verbindlicher zu gestalten.

Eine zweite Möglichkeit, das Problem unterschiedlicher Interpretationen zu beherrschen, besteht darin, den Versuchspersonen die gemeinte logische Interpretation genauer zu erklären. Dies versuchten Klauer et al. (1997; vgl. Klauer et al., 2000; Meiser et al., 2001) im Bereich des aussagenlogischen Schlussfolgerns mithilfe eines Wahrheitswerttabellentrainings. Die Bedeutung von Junktoren beim aussagenlogischen Schlussfolgern kann durch Wahrheitswerttabellen definiert werden. Eine solche Tabelle listet für einen gegebenen Junktor J auf, welche Kombinationen von p beziehungsweise nicht-p mit q beziehungsweise nicht-q mit der Aussage p J q vereinbar sind. Mit der Aussage »p und q« zum Beispiel ist nur das gemeinsame Gelten von p und q vereinbar, alle anderen Kombinationen mit nicht-p oder nicht-q sind mit der Gültigkeit der Aussage unvereinbar. In einem vorbereitenden Training wurde mit den Versuchspersonen zunächst eine gewünschte Deutung der später verwendeten Junktoren vereinbart und anhand der Wahrheitswerttabelle eingeübt. Bei den späteren Denkaufgaben durfte dann vorausgesetzt werden, dass die gewünschte Interpretation zugrunde gelegt wurde. Tatsächlich ergab sich so eine Steigerung in der Anzahl korrekter Schlüsse, möglicherweise bei gleich bleibenden Denkprozessen einfach deswegen, weil die Interpretationen der Versuchspersonen und der Versuchsleiter stärker deckungsgleich waren als ohne das Verständnistraining. Bei relativ aufdringlichen Eingriffen wie dem Training ist natürlich auch möglich, dass

die Denkprozesse selber nach dem Training anders verlaufen als vorher. Das wäre eine Einschränkung für die Nützlichkeit dieses Ansatzes, wenn er alleine dazu dienen soll, die Interpretationen der Versuchspersonen zu kontrollieren, ohne die Denkprozesse selber zu verändern.

Antworttendenzen unter Unsicherheit

Die betrachteten Theorien gehen in der Regel davon aus, dass das menschliche Denkvermögen beschränkt ist. Schranken werden insbesondere in der begrenzten Arbeitsgedächtniskapazität gesehen und begründen, warum die Schwierigkeit eines Schlusses im Falle der Theorie Mentaler Modelle von der Anzahl der benötigten mentalen Modelle abhängen soll beziehungsweise im Fall regelbasierter Theorien von der Komplexität der Ableitungskette, die Prämissen und Schlussfolgerung verbindet. Damit ist die Möglichkeit des Scheiterns der internen Denkprozesse impliziert. Was machen aber Denker, wenn sie sich überfordert sehen, keine Schlussfolgerung begründet ableiten oder bejahen können und dennoch zu einer Antwort genötigt werden, und sei es nur die, dass nichts folge? Bei Aufgaben mit Mehrfachwahlantworten soll eine der angebotenen Optionen ausgewählt werden, beim selbsttätigen Generieren von Schlussfolgerungen muss vermutlich zwischen mental repräsentierten Möglichkeiten ausgewählt werden. In dieser Situation greifen Entscheider in der Regel zu Heuristiken, die die verfügbare Information in vereinfachender Weise nutzen, um dennoch eine begründete und vernünftig erscheinende Antwort auszuwählen (z. B. Gigerenzer u. Goldstein, 1996). Sowohl Ansätze aus der Theorie Mentaler Modelle wie auch regelbasierte Theorien lassen solche Prozesse heuristischen Ratens bei Unsicherheit im Prinzip zu, sie werden aber nicht genauer spezifiziert. Auch das stellt eine Schwierigkeit beim Ableiten von statistischen Hypothesen aus den theoretischen Aussagen dar. Zum Beispiel können zwei Schlüsse bei gleicher Wahrscheinlichkeit der Bewältigung durch interne Denkprozesse unterschiedlich schwierig scheinen, weil die Denkprozesse mit einer gewissen Wahrscheinlichkeit scheitern und die beim Raten dann bevorzugte Option im einen Fall die im normativ logischen Sinne korrekte ist, im anderen Fall eine falsche.

Überlegungen dazu, welche Informationen Rateheuristiken berücksichtigen könnten, finden sich in vielen Arbeiten zur Denkpsychologie. Zum Beispiel wird vermutet, dass Versuchspersonen unter Unsicherheit vorsichtige Antworten wie »Es folgt nichts« oder negativ formulierte Schlussfolgerungen (*negative conclusion bias*; Evans u. Handley, 1999) bevorzugen könnten. Ebenso könnten oberflächliche Ähnlichkeiten von Prämissen und Schluss-

folgerungen die Disposition beeinflussen, einen Schluss als logisch gültig zu akzeptieren, wie es im Bereich der Syllogismen zum Beispiel in der so genannten Atmosphärenhypothese formuliert wurde: Ist einer der Prämissenquantoren negativ, so werden negative Schlussfolgerungen präferiert, andernfalls positive. Ist einer der Prämissenquantoren ein partikulärer, so werden Schlussfolgerungen mit partikulären Quantoren präferiert, andernfalls solche mit universellen Quantoren (Woodworth u. Sells, 1935). Das Phänomen des so genannten *belief bias* im syllogistischen Denken (Klauer et al., 2000) schließlich beschreibt die Beobachtung, dass unabhängig von der logischen Form des Syllogismus inhaltlich plausible Schlussfolgerungen häufiger als logisch gültig akzeptiert werden als unplausible, was vermutlich zumindest zum Teil ebenfalls auf heuristisches Raten zurückgeführt werden kann. Klauer et al. (2000) haben ferner in vielen Experimenten gezeigt, dass die Wahrscheinlichkeit, einen Syllogismus als gültig zu akzeptieren, auch von *Basisrateninformation* über den zu erwartenden Anteil von gültigen Syllogismen unter den zur Beurteilung vorgelegten beeinflusst wird.

Im Kontext der informationstheoretischen Ansätze wurden in jüngerer Zeit vermehrt Experimente durchgeführt, die weitere probabilistische Kontextinformationen manipulierten, zum Beispiel die Wahrscheinlichkeit, mit der die Ereignisse p und q der konditionalen Aussage »Wenn p, dann q« auftreten (z. B. Oaksford et al., 1999; Oberauer et al., 1999). Vor dem Hintergrund des Gesagten ist plausibel, dass solche Manipulationen nicht nur Denkprozesse beeinflussen können, sondern auch Heuristiken zur Antwortfindung unter Unsicherheit.

Tatsächlich kann für die informationstheoretischen Ansätze gefragt werden, ob sie nicht genau solche Heuristiken, die bekanntermaßen für statistische und andere Kontextinformationen empfänglich sind, modellieren. Das soll am Beispiel der Wasonschen Auswahlaufgabe kurz erläutert werden. Jüngeren Darstellungen zufolge (Oaksford, 2002) modelliert der Ansatz die Wahrscheinlichkeiten, mit denen die unterschiedlichen Kartentypen gewählt werden. Die Wahrscheinlichkeiten der Wahlen der einzelnen Kartentypen werden als Funktion probabilistischer Kontextinformation über die Wahrscheinlichkeit der mit p und q gemeinten Ereignisse beschrieben. Die Wahrscheinlichkeit einer bestimmten Kombination von Karten wird dann durch das Produkt der Wahrscheinlichkeiten der gewählten Karten und dem Komplement der Wahrscheinlichkeiten für die nicht gewählten Karten modelliert. Nicht anders würde man aber die Kartenwahlen modellieren, wenn unabhängiges Raten der einzelnen Kartentypen angenommen wird (Gleißner u. Erdfelder, 1994).

Bei der Wasonschen Auswahlaufgabe sind Rateprozesse schon aufgrund

des Aufgabenformats wahrscheinlich. Wenn einer Versuchsperson nicht klar ist, was von ihr verlangt wird, oder sie nicht motiviert ist, über die Aufgabe nachzudenken, liegt nahe, dass sie aufgrund oberflächlicher Informationen rät und dass diese Rateprozesse in erster Näherung durch unterschiedliche, unabhängige Wahrscheinlichkeiten für die einzelnen Kartentypen beschrieben werden können. Das Phänomen des *matching bias* beschreibt in diesem Kontext zum Beispiel die Tatsache, dass solche Karten tendenziell bevorzugt gewählt werden, deren Aufschriften in der zu prüfenden Regel explizit benannt sind. Dieses Phänomen könnte zumindest zum Teil auf eine einfache Heuristik zurückgehen, beim Raten solche Karten zu bevorzugen, deren Aufschrift bereits einmal explizit genannt wurde.

Zumindest für einen gewissen Prozentsatz der Antworten ist also Kartenwahl durch unabhängiges Raten plausibel. Eine interessante Schlussfolgerung ist die, dass experimentelle Manipulationen zum Beispiel der Instruktionen, der inhaltlichen Einbettungen oder des Aufgabenformats nicht nur die Ratewahrscheinlichkeiten für die einzelnen Kartentypen beeinflussen können, sondern auch den Prozentsatz der durch Raten zustande gekommenen Antworten. Macht zum Beispiel die Instruktion den Versuchspersonen in der einen Bedingung deutlicher, was verlangt wird, als in der anderen, die eher zu Verwirrung führt, so darf in der letzteren Bedingung mit größerer Unsicherheit und häufigerem Raten gerechnet werden. Ein höherer Prozentsatz von Rateprozessen, modelliert als unabhängige Wahlen der vier Kartentypen, dürfte sich in der Regel in einer höheren Dispersion der Antworthäufigkeiten über die sechzehn möglichen Kartenkombinationen äußern, wodurch die durch Denken ermittelten systematisch auftretenden Kartenkombinationen weniger prägnant hervortreten. Eine informelle Inspektion der einschlägigen Studien erweckt in der Tat den Eindruck, dass ein häufiger Effekt von experimentellen Manipulationen darin besteht, die Streuung der Kartenwahlen über die sechzehn Möglichkeiten hinweg zu verändern. Solche Effekte werden aber durch die vorherrschenden Auswertungsstrategien, die sich in der Regel auf die Analyse einiger weniger, besonders interessierender Auswahlkombinationen oder auf die Analyse von Randwahrscheinlichkeiten der Wahlen der einzelnen Kartentypen beschränken, nicht aufgedeckt.

Durch Instruktionen und geeignete Antwortformate, vermutlich auch durch motivationale Faktoren und Entlastung der Versuchsperson, kann der Einfluss verfälschender Ratetendenzen in vielen Fällen gemildert, aber wahrscheinlich nicht eliminiert werden. Für die Kontrolle und Abschätzung des Einflusses solcher Tendenzen bietet sich wiederum an, mögliche verfälschende Einflüsse explizit zu modellieren. Das ist im Beispiel aus Abbildung 1 an-

hand der *s*-Parameter geschehen, die die Auswahl der angebotenen Antwortoptionen beschreiben, wenn die richtige Antwort durch deduktives Denken (mit Wahrscheinlichkeit 1-*r*) nicht gefunden werden kann. Die Parameter hängen von den Prämissen und der Schlussfolgerung ab, um systematisches Raten aufgrund oberflächlicher Merkmale von Prämissen und Schlussfolgerungen beschreiben zu können. Ähnliche Modelle haben Klauer et al. (2000) für das syllogistische Schlussfolgern und Gleißner und Erdfelder (1994) für die Wasonsche Auswahlaufgabe vorgeschlagen.

Indem Theorien des deduktiven Denkens auf diese oder ähnliche Weise um Hilfsannahmen über die unterschiedlichen Möglichkeiten der Prämissenin terpretation und über Antwortstrategien bei Unsicherheit ergänzt werden, gelingen strengere und fairere Tests dieser Theorien. Das explizite Berücksichtigen der Hilfsannahmen hat darüber hinaus den erwünschten Nebeneffekt, solche Hilfsannahmen, ihre theoretische Begründung und ihre Rolle am Zustandekommen beobachteter Effektmuster in den Blickpunkt zu rücken und der wissenschaftlichen Diskussion zugänglich zu machen.

Literatur

Batchelder, W.H., Riefer, D.M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Braine, M.D.S., O'Brien, D.P., Noveck, I.A., Samuels, M.C., Lea, R.B., Fisch, S.M., Yang, Y. (1995). Predicting intermediate and multiple conclusions in propositional logic inference problems: Further evidence for a mental logic. *Journal of Experimental Psychology: General, 124,* 263–292.

Braine, M.D.S., Reiser, B.J., Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In G.H. Bower (Hg.), *The psychology of learning and motivation: Advances in research and thinking* (Bd. 18, S. 313–371). San Diego, CA: Academic Press.

Braine, M.D.S., Rumain, B. (1983). Logical reasoning. In J.H. Flavell u. E.M. Markman (Hg.), *Handbook of child psychology. Vol. 3: Cognitive development* (S. 263–340). New York: Wiley.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1996). Grundlagen experimenteller Methoden. In E. Erdfelder, R. Mausfeld, T. Meiser u. G. Rudinger (Hg.), *Handbuch Quantitativer Methoden* (S. 37–46). Weinheim: Psychologie Verlags Union.

Bucciarelli, M., Johnson-Laird, P.N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23,* 247–303.

Chater, N., Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38,* 191–258.

Dickstein, L.S. (1975). Effects of instruction and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology, 104,* 376–384.

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann u. W. Tack

(Hg.), *Enzyklopädie der Psychologie. Band: Methodologische Grundlagen der Psychologie* (S. 604–648). Göttingen: Hogrefe.

Evans, J.St.B.T., Handley, S.J. (1999). The role of negation in conditional inference. *Quarterly Journal of Experimental Psychology, 52A,* 739–769.

Evans, J.St.B.T., Newstead, S.E., Byrne, R.M.J. (1993). *Human reasoning.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

Gleißner, U., Erdfelder, E. (1994). *Heuristische und algorithmische Strategien beim Wason Selection Task.* Vortrag gehalten auf der 36. Tagung experimentell arbeitender Psychologen, München, 28.–31.3.1994.

Janis, I.L., Frick, F. (1943). The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology, 33,* 73–77.

Johnson-Laird, P.N. (1983). *Mental models. Towards a cognitive science of language, inference, and consciousness.* Cambridge, UK: University Press.

Johnson-Laird, P.N., Bara, B.G. (1984). Syllogistic inference. *Cognition, 16,* 1–61.

Johnson-Laird, P.N., Byrne, R.M.J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109,* 646–678.

Johnson-Laird, P.N., Byrne, R.M.J., Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99,* 418–439.

Klauer, K.C., Meiser, T., Naumer, B. (2000). Training propositional reasoning. *Quarterly Journal of Experimental Psychology, 53A,* 868–895.

Klauer, K.C., Musch, J., Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107,* 852–884.

Klauer, K.C., Oberauer, K. (1995). Testing the mental model theory of propositional reasoning. *Quarterly Journal of Experimental Psychology, 48A,* 671–687.

Klauer, K.C., Stegmaier, R., Meiser, T. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking and Reasoning, 3,* 9–46.

Meiser, T., Klauer, K.C., Naumer, B. (2001). Propositional reasoning and working memory: The role of prior training and pragmatic contents. *Acta Psychologica, 106,* 303–327.

Oaksford, M. (2002). Predicting the results of reasoning experiments: Reply to Feeney and Handley (2000). *Quarterly Journal of Experimental Psychology, 55A,* 793–798.

Oaksford, M., Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631.

Oaksford, M., Chater, N., Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5,* 193–243.

Oberauer, K., Wilhelm, O., Rosas Diaz, R. (1999). Bayesian rationality for the Wason selection task? A test of optimal data selection theory. *Thinking and Reasoning, 5,* 115–144.

Osman, M., Laming, D. (2001). Misinterpretation of conditional statements in Wason's selection task. *Psychological Research, 65,* 128–144.

Rips, L.J. (1994). *The psychology of proof.* Cambridge, MA: MIT Press.

Roberts, M.J., Newstead, S.E., Griggs, R.A. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking and Reasoning, 7,* 173–204.

Wason, P.C. (1966). Reasoning. In B.M. Foss (Hg.), *New horizons in psychology 1* (S. 135–151). Harmondsworth, UK: Penguin.

Woodworth, R.S., Sells, S.B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18,* 451–460.

Joachim Funke[1]

Komplexes Problemlösen: Möglichkeiten deduktivistischen Vorgehens

Zusammenfassung

Dieser Beitrag befasst sich mit Möglichkeiten deduktivistischen Vorgehens im Bereich des Komplexen Problemlösens. Nach einer kurzen Einführung werden zunächst zwei Forschungsrichtungen in diesem Bereich – Suche nach interindividuellen Differenzen und Suche nach allgemeinen Gesetzmäßigkeiten – vorgestellt, um dann auf Mythen und Fakten über problemlösendes Denken und Komplexes Problemlösen einzugehen. Im Anschluss daran werden Bausteine für eine Theorie komplexen Problemlösens vorgestellt, die ein deduktivistisches Forschungsprogramm in diesem Bereich begründen könnten.

Einführung

Seit bald 30 Jahren existiert ein neuer Forschungsgegenstand der Problemlösepsychologie, der unter dem Label »Komplexes Problemlösen« den Umgang von Menschen mit komplexen Situationen untersucht. Schon der Titel der Pionierarbeit »Wie Menschen eine Welt verbessern wollten« (Dörner, 1975) suggeriert eine Unfähigkeit des menschlichen Umgangs mit Komplexität, die in vielen Folgearbeiten belegt werden sollte (Dörner, 1981, 1989b; Dörner u. Wearing, 1995). Diese und eine Reihe anderer interessanter Arbeiten zeigen, dass dieser Themenbereich eine substanzielle Erweiterung des traditionellen Gegenstandsbereichs »Problemlösen« darstellt (zum Überblick siehe z. B. Funke, 2003). Beim komplexen Problemlösen kommen (im Unterschied zu einfachem Problemlösen) maßgeblich die Merkmale der *Vernetztheit* und *Dynamik* hinzu, das heißt, die problemlösende

1 Für hilfreiche Kommentare zu einer Vorfassung bedanke ich mich bei Dipl.-Psych. Miriam Spering (Universität Gießen). Jochen Musch und Stephan Rückert (Universität Mannheim) haben wertvolle Hinweise auf Verbesserungsmöglichkeiten gegeben. Besonderer Dank gebührt meinem Mitherausgeber Edgar Erdfelder, der mir – wie schon oft – geholfen hat, meine unscharfen Gedanken präziser zu formulieren.

Person ist mit einer Vielzahl voneinander abhängiger Variablen konfrontiert, die sich – sowohl aufgrund von Eingriffen in das System wie auch aufgrund eigendynamischer Abläufe – über die Zeit hinweg verändern (vgl. Funke, 2001a, 2001b).

Die theoretische Durchdringung der ablaufenden Prozesse beim problemlösenden Umgang mit komplexen Umwelten ist nicht weit fortgeschritten, wenn man einmal von den Modellierungsansätzen der Bamberger Gruppe absieht (Dörner, 1999; Dörner et al., 2002; Schaub, 2001). Die seinerzeit formulierte These (Funke, 1984, S. 159, im Original kursiv) besitzt auch heute noch weitgehend Gültigkeit: »Forschung zum komplexen Problemlösen wird momentan theoriearm betrieben. Die Phase des explorierenden Vorgehens muß abgelöst werden durch eine Phase der Modellbildung und -testung.«

Der vorliegende Beitrag wird kurz auf zwei Forschungsrichtungen zum Komplexen Problemlösen eingehen, die sich als wichtig in der kurzen Geschichte dieses Forschungsgebiets erwiesen haben. Es wird der Wunschtraum einer deduktivistisch prüfbaren Theorie zum Komplexen Problemlösen skizziert, bevor Mythen und Fakten zum problemlösenden Denken benannt werden, die zu vorschnellen Einengungen des Forschungsgegenstands geführt haben könnten. Schließlich wird der Versuch unternommen, Bausteine für eine deduktivistisch prüfbare Theorie des Komplexen Problemlösens vorzustellen.

Zwei Forschungsrichtungen zum Komplexen Problemlösen

Buchner (1995) unterscheidet zwei Forschungsrichtungen zum Komplexen Problemlösen, die interessanterweise beide ihren Ursprung in Europa haben: Eine stammt von Broadbents Forschung über verschiedene Gedächtnissysteme aus den siebziger Jahren (Berry u. Broadbent, 1984, 1988; Broadbent, 1977; Broadbent u. Aston, 1978; Broadbent et al., 1986; Hayes u. Broadbent, 1988) und baut auf formalen Aufgabenanalysen auf, die andere entstammt Dörners etwa zur gleichen Zeit durchgeführten Forschungen über die Struktur intelligenten Verhaltens in komplexen Anforderungssituationen (Dörner, 1981; Dörner, Kreuzig et al., 1983; Dörner u. Wearing, 1995) und zielt ab auf die Suche nach interindividuellen Unterschieden. Die zuerst genannte Richtung (Broadbent) ist der experimentellen, theorietestenden Forschung verpflichtet und benutzt daher sehr einfache dynamische Aufgabenstellungen (z. B. ein simuliertes Transport-

system oder eine kleine Zuckerfabrik); die zweite Richtung (Dörner) stellt ein mehr exploratives Vorgehen dar, wie an der mehr als 2000 Variablen umfassenden simulierten Kleinstadt LOHHAUSEN zu erkennen ist. Beiden Richtungen gemeinsam ist der Rückgriff auf unerfahrene Versuchspersonen. Damit unterscheiden sich diese Zugänge deutlich von den in den USA präferierten Expertenstudien (vgl. Sternberg, 1995). Auf beide Zugänge soll kurz eingegangen werden.

Zugang 1 – Experimentalpsychologische Perspektive: Systematische Manipulation von Szenarien

Broadbents experimentelle Arbeiten führten zur Konzeption zweier verschiedener Lernmodi, eines expliziten und eines impliziten Lernmodus (z. B. Hayes u. Broadbent, 1988). Der explizite Modus bezieht sich auf das verbalisierbare Wissen, das eine Person über das zu explorierende System erwirbt und das in regelhafter Form vorliegt (Wenn-dann-Aussagen). Der implizite Modus stellt eine Art Kontingenzdetektion dar, der man sich weniger deutlich bewusst ist und die sich über Handlungen (z. B. Steuerungseingriffe) manifestiert, deren Gründe nur schwer anzugeben sind. Broadbent und Berry begründen diese beiden Modi aus Dissoziationsbefunden: Versuchspersonen können zwar ein System steuern (implizites Lernen), aber Fragen über die Struktur des Systems nur schlecht beantworten. Zu diesen Dissoziationsstudien sind in der Zwischenzeit eine Reihe von Nachfolgeuntersuchungen durchgeführt worden (für einen Überblick siehe Berry u. Broadbent, 1995). Auch Kritik an dieser Interpretation der Dissoziationsbefunde ist vorgetragen worden, die eine wesentlich einfachere Deutung nahe legt (Buchner et al., 1995): Danach ist die Dissoziation zwischen Wissen über ein System und der Steuerungsleistung dieses Systems bedingt durch die Tatsache, dass Personen mit guter Steuerleistung kaum Wissen über das gesamte Spektrum des Systems sammeln, da sie sich definitionsgemäß die meiste Zeit im Zielzustand oder zumindest nah am Ziel befinden.

Wesentliches Kennzeichen dieses Zugangs 1 ist eine experimentelle Manipulation des Reizmaterials und dessen Präsentationsbedingungen. Insbesondere die systematische Manipulation von Szenarien (bzw. Systemmerkmalen) ist zu einem Charakteristikum dieser Forschungsrichtung geworden (Funke, 1986, 2001a, 2001b; Vollmeyer u. Funke, 1999). Erst durch systematische Variation von Systemmerkmalen wie Anzahl der Vernetzungen, Grad der Eigendynamik oder Ausmaß an Zeitverzögerung konnte deren Einfluss aufgezeigt werden.

Zugang 2 – Korrelationsstatistische Perspektive: Suche nach interindividuellen Unterschieden

Dörners Arbeiten haben sowohl zu einer intensiven Diskussion über die Schwächen gegenwärtiger Intelligenztests als auch zu einer handlungstheoretischen Analyse des Verhaltens in komplexen Umwelten geführt (Dörner, 1986, 1996; Dörner u. Kreuzig, 1983; Dörner et al., 1999). In den frühen Arbeiten von Dörner und Mitarbeitern stand die Enttäuschung über die geringe prädiktive Potenz traditioneller Intelligenztests hinsichtlich des Problemlösens in alltäglichen Situationen im Vordergrund. Anstatt zur Untersuchung des Problemlösens Aufgaben zu verwenden, die man als eher akademisch bezeichnen könnte, schlug Dörner einen alternativen Zugang vor: Die Konstruktion komplexer alltäglicher Problemlösungen in Form simulierter Szenarien, mit denen Versuchspersonen unter kontrollierten Laborbedingungen zu interagieren hatten (siehe Brehmer u. Dörner, 1993; Brehmer et al., 1991).

Versuchspersonen, die mit diesen Szenarien umgingen, waren tatsächlich mit anderen Anforderungen konfrontiert, als sie etwa in herkömmlichen Intelligenztests gestellt werden: (a) die Komplexität der Situation und (b) die Vernetztheit zwischen einer großen Zahl von Variablen zwingen die handelnde Person zu einer Reduktion der großen Menge an Informationen und zur Antizipation von Nebenwirkungen; (c) die dynamische Art der Problemstellung verlangt sowohl die Vorhersage zukünftiger Entwicklungen (eine Art von Planung) als auch die langfristige Kontrolle von Entscheidungen; (d) die Intransparenz der Szenarios erfordert die systematische Sammlung von Informationen; (e) die Präsenz multipler Ziele (Polytelie) erfordert die sorgfältige Ausarbeitung von Prioritäten und ein Ausbalancieren widersprüchlicher, konfligierender Ziele.

Wesentliches Kennzeichen von Zugang 2 ist die Suche nach interindividuellen Unterschieden und Korrelaten von Erfolg oder Misserfolg. Bei konstant gehaltenen Systemen werden Verhaltensspielräume betrachtet und individuelle Trajektorien durch ein System sowie deren Resultate mit differenzialpsychologisch relevanten Konstrukten (Intelligenz, Persönlichkeitsmerkmale etc.) in Verbindung gebracht.

Dörners Vorschlag, computersimulierte Szenarien als alternative Messverfahren zur Erfassung einer »operativen Intelligenz« zu erproben (Dörner, 1986), haben seither viele Forscher aufgegriffen. Als Folge daraus sind viele Szenarien entstanden, die nicht nur in der Forschung, sondern auch in Anwendungskontexten zum Einsatz kamen (Überblicke bei Funke, 1995, 1998; Strauß u. Kleinmann, 1995). Dabei ist die übertriebene Euphorie der

ersten Jahre angesichts eines neuartigen Instruments durch gesunde Skepsis hinsichtlich dessen Messgüte und Aussagekraft ersetzt worden (Kersting, 1999).

Deduktivistisch prüfbare Theorie des Komplexen Problemlösens: Ein Wunschtraum?

In seinem Buch »Theorie und Planung psychologischer Experimente« liefert Bredenkamp (1980, S. 4) eine Definition des Begriffs »deduktivistisch« und schildert zugleich sein Verständnis experimenteller Forschung: »›Deduktivistisch‹ heißt vorerst, daß induktive Verallgemeinerungen der experimentellen Resultate vermieden werden sollen. ... Im Rahmen einer deduktivistischen Theorie des Experiments läßt sich die Durchführung ›künstlicher‹, kontrollierter Versuche nicht nur rechtfertigen, sondern sie ist sogar geboten.« Mit diesem Verständnis experimenteller Forschung ist ein Streitpunkt der Forschung zum Komplexen Problemlösen verbunden: Darf und kann man mit komplexen Systemen experimentell arbeiten?

Die Diskussion um die Möglichkeiten experimenteller Forschung mit komplexen Problemlösungsszenarien findet sich etwa in einer Arbeit von Dörner (1989a), in der er sich über die experimentelle Methode am Beispiel der Analyse von unbekannten KI-Schildkröten lustig macht: Die experimentelle Analyse bringe nichts, erst das »Verstehen« des Systems erlaube die richtige Modellbildung. In einem späteren Papier (Dörner, 1992) geht es um die Nachteile von Reduktions- und die Vorteile von Kondensationsverfahren bei der Systemanalyse. Schließlich stellen Brehmer und Dörner (1993) die (unangenehmen) »Meerengen des Labors« der (wunderbaren) »tiefblauen See der Feldforschung« gegenüber und erzeugen damit eine methodische Opposition, in der man sich dem einen ab- und dem anderen zuwenden soll. Anstatt hier einen der beiden Zugänge herauszuheben, scheint mir eine liberale Position – wie so oft – erstrebenswert.

Auf die Frage, welche der möglichen Zugangsweisen zur Erforschung komplexen Problemlösens zu bevorzugen sei, muss die Antwort wohl lauten: Idealerweise sollten alle Zugänge zum Gegenstand ausgeschöpft werden! Dieser pluralistische Standpunkt steht in Einklang mit den Vorschlägen von Sternberg und Grigorenko (2001), die – in anderem Kontext – für eine »unified psychology« eintreten und hinsichtlich der dafür einzusetzenden Methoden wie folgt argumentieren (S. 1072 f.): »The truth is that no method will provide a panacea: Different methods have different advan-

tages and disadvantages, and, by using multiple methods, one capitalizes on the strengths of the methods while helping to minimize the effects of their weaknesses.«

Warum stellt eine deduktivistisch prüfbare Theorie des Komplexen Problemlösens einen Wunschtraum dar, und das sogar in zweierlei Hinsicht? Zum einen aus dem einfachen Grund, dass bislang keine kohärente Theorie mit empirischem Gehalt zum Lösen komplexer Probleme existiert; zum anderen aber auch aus dem Grund heraus, dass eine deduktivistisch prüfbare Theorie wünschenswert wäre, um aus der phänomenorientierten Phase des »Schmetterlingsammelns« (Dörner, 1983) in die Phase theorieorientierten Testens von Annahmen über problemlösendes Denken zu wechseln.

Gerade die Anfangszeit der Forschung bestand im Wesentlichen aus einer Phase ungestümen Datensammelns, von dem man sich das Hervortreten salienter Muster erhofft hatte. So wurden etwa in dem klassischen Szenario LOHHAUSEN pro Versuchsperson über 100.000 Daten erhoben (Dörner, Kreuzig et al., 1983, S. 137), die anschließend mühsam (und nicht immer auf brauchbare Weise) kondensiert werden mussten. Die Autoren bedienten sich dabei der »Methode der theoretischen Konsistenz«, wonach aus einem Teil der Daten (der *Induktionsbasis*) per Induktion ein »theoretisches System« konstruiert wird, das anhand eines anderen Teils der Daten (der *Prüfbasis*) überprüft wird. An dieses Induktionssystem werden bestimmte Anforderungen gestellt, so beispielsweise die logische Unabhängigkeit zwischen Variablen der Induktions- und der Prüfbasis oder die vorherige Unbekanntheit des empirischen Zusammenhangs zwischen Induktions- und Prüfbasis.

So interessant dieser Plan auch aussehen mag, so wenig wurde er in der konkreten Forschungspraxis beachtet: Vielfach finden sich Beiträge, die so viele Korrelationen berichten, dass am Ende nur noch pauschal geprüft wird, ob die *Anzahl* gefundener Signifikanzen signifikant sei. So schreibt etwa – ein Beispiel unter vielen – Schaub (1990, S. 91) enttäuscht: »Von insgesamt mehr als hundert Korrelationen ... sind noch nicht einmal ein Dutzend signifikant.« Und am Ende wird die Induktion explizit eingeräumt (Dörner, Kreuzig et al., 1983, S. 445): »Die Tatsache, dass Theorie und Daten so wunderschön zusammenpassen, ... kann ... nicht allzu sehr verwundern; die Theorie folgte den Daten.«

Dass die Entwicklung einer deduktivistisch prüfbaren Theorie des Komplexen Problemlösens bislang nicht erfolgt ist, mag unter anderem mit Annahmen zu tun haben, die einer derartigen Entwicklung im Wege stehen und lediglich Theorien mit wenig empirischem Gehalt hervorbringen. Nachfolgend wird deswegen versucht, mögliche Mythen zu benennen und so weit wie möglich zu entkräften.

Mythen und Fakten über Komplexes Problemlösen

Was sind Grundannahmen, die bei der Erforschung problemlösenden Denkens mehr oder weniger implizit präsupponiert werden? Grundannahmen, die zu möglicherweise voreiligen Beschränkungen der Forschung geführt haben könnten, sollen im nachfolgenden Teil unter besonderer Berücksichtigung des Komplexen Problemlösens diskutiert werden.

(1) Denken findet im Kopf statt. Diese Vorannahme wird verschärft, wenn man das Wort »ausschließlich« hinzufügt und sich damit völlig auf eine innenbezogene Perspektive beschränkt.

Was ist daran problematisch? Geschieht Denken nicht durch Verwendung des Gehirns? Gibt es noch andere Organe? Sicherlich nicht – aber die Konzentration auf das Gehirn als den Ort des Geschehens darf nicht darüber hinwegtäuschen, dass wir in vielen Fällen zum Problemlösen auf die Interaktion mit der Umwelt angewiesen sind. Diese *Ökologie des Denkens* darf nicht übersehen werden (siehe dazu Graumann, 1997, 2002). Gerade komplexe Probleme bestehen häufig in einer Mensch-Maschine-Interaktion, in der das problemlösende *Handeln* das problemlösende *Denken* dominiert. Dies ist insbesondere an Arbeiten aus dem Bereich »naturalistic decision making« (z. B. Zsambok, 1997) abzulesen – ein Bereich, der hohe Affinität zum Ansatz des Komplexen Problemlösens aufweist, aber vor allem aus der Unzufriedenheit mit Paradigmen klassischer Entscheidungs- und Urteilsforschung gespeist wird.

Natürlich soll mit diesen Ausführungen keinesfalls impliziert werden, es sei ein Mythos, dass die dem Denken zugrunde liegenden neuronalen Aktivitäten im Kopf erfolgten – aber um von »Denken« und erst recht von »Problemlösen« zu sprechen, muss *mehr* dazu genommen werden. Man könnte das Argument in abgeschwächter Form auch so formulieren: Komplexes Problemlösen hat sich zu sehr auf innere Prozesse kapriziert und zu wenig auf die Ökologie des Denkens geachtet.

(2) Je schwieriger die Problemstellungen, umso komplexer die zu ihrer Bewältigung benötigten Denkprozesse.

Dies ist eine Annahme, die bereits Bühler (1908) als fragwürdig bezeichnete. Welche Prozesse bei welcher Art von Aufgabenstellung zum Einsatz kommen, ist vermutlich weniger von der Aufgabe als vielmehr von den beabsichtigten Operationen abhängig. Welche Werkzeuge zum Einsatz kommen, ist nicht allein durch die Art der Problemstellung bedingt – vorrangig

sind es die beabsichtigten Operationen, die man auszuführen gedenkt, welche über die Art der Werkzeuge entscheiden. Auch zur Beschreibung herausragender Problemlösungen wie der Ziehung der 137. Wurzel aus einer 1.000-stelligen Zahl durch einen Rechenkünstler braucht man keine neuen Konzepte, sondern kommt mit dem Standard-Repertoire kognitiver Erklärungen weiter (Bredenkamp, 1990).

(3) Denken hat wenig oder nichts mit Gefühlen zu tun, Kognition und Emotion sind scharf zu trennen.

Diese Fehlkonzeption hat zu einem unglücklichen Auseinanderdriften eng verwandter Bereiche in der Forschung geführt. Diese Tendenz ist allerdings schon früh erkannt worden (z. B. von Dörner, Reither u. Stäudel, 1983; Kuhl, 1983) und hat zu einer nicht geringen Zahl von Arbeiten geführt, die dieses Spannungsverhältnis klären (siehe z. B. Dörner, 1993, 1998; Dörner et al., 1988; Spies, 1990, 1999; Spies u. Hesse, 1986). Eine Separierung von Kognition, Emotion und Motivation sollte nicht vorgenommen werden, obwohl die Spezialisten der drei Richtungen nur wenig Bezug aufeinander nehmen. Globalere handlungstheoretische Modellvorstellungen (Dörner et al., 1988), aber auch ältere Vorstellungen der Informationsverarbeitungstheorien (Simon, 1979/1967) leisten hier die notwendige Integration. Auch Stimmungseinflüsse sind beim Komplexen Problemlösen inzwischen belegt (Spering et al., under revision). Otto et al. (2002) konnten zeigen, dass Maße der Emotionalen Intelligenz einen eigenständigen Prädiktionswert beim Komplexen Problemlösen besitzen, der etwa gleich hoch ist wie derjenige der Verarbeitungskapazität.

(4) Problemlösendes Denken ist eine Fähigkeit, die – wenn man sie besitzt – in allen möglichen Situationen zum Einsatz gebracht werden kann.

Die Annahme einer universellen Kompetenz, die in allen Lebensbereichen zum Einsatz kommen könne, konfligiert mit Erkenntnissen über die hochgradige Bereichsspezifität kognitiver Leistungen (Frensch u. Buchner, 1999). Auch wenn Studien wie die von Wittmann und Süß (1999) beeindruckende Konvergenz der Problemlösungsgüte in drei computersimulierten Szenarien (FSYS, LEARN und POWERPLANT) nachweisen, zeigen eine ganze Reihe von Studien derartige Korrelationen allenfalls auf Verhaltens-, nicht aber auf Leistungsebene (Kotkamp, 1999; Rigas u. Brehmer, 1999; Rigas et al. 2002; Strohschneider, 1986, 1991).

(5) Problemlösendes Denken erfolgt regelgeleitet.

Die Annahme, dass es sich um eine geordnete Abfolge von Schritten handelt, die im Vorhinein konstruiert und dann abgearbeitet werden, mag nur bestimmte Erscheinungsformen problemlösenden Denkens charakterisieren. Genauso gut sind Problemsituationen vorstellbar, in denen Akteure ein Geschehen »laufen lassen« und nur durch wenige kontrollierte Eingriffe in die erwünschte Richtung lenken. Der Ansatz situierter Kognition (z. B. Clark, 1997; Hutchins, 1995) macht deutlich, dass Denken nicht nur im Kopf, sondern auch in der Welt abläuft (so der Titel des Einführungsbeitrags zu einem entsprechenden Themenheft der *Cognitive Science* von Norman, 1993). Dies ist ja bereits im ersten Punkt erwähnt worden.

Die Annahme regelgeleiteter Prozesse findet sich als Standard in so genannten Produktionssystemen, wie sie im Rahmen kognitiver Modellierungen verwendet werden (z. B. mittels ACT-R von Anderson u. Lebiere, 1998). Demgegenüber stehen konnektionistische Systeme, die auf symbolische Repräsentationen völlig verzichten (zum Überblick siehe Medler, 1998). Auf den ersten Blick liegt es nahe, Problemlösungsprozesse mit Produktionsregeln zu modellieren, wie dies von Wallach (1998) gemacht wird. Aber es zeigt sich, dass auch komplexe Planungsprozesse mit Erfolg als konnektionistisches Netz dargestellt werden können (Schenck, 2001).

(6) Einfaches Problemlösen lässt sich experimentell untersuchen, Komplexes Problemlösen dagegen entzieht sich aufgrund seines Systemcharakters diesem Zugang.

Komplexe Systeme weisen vielfältige Abhängigkeiten innerhalb und zwischen den verschiedenen Subsystemen auf. Auf den ersten Blick scheint ein systematisches Ausprobieren hier nicht möglich (siehe die oben bereits erwähnte ironische Attacke auf die »dummen« Experimentalisten bei Dörner, 1989a). Dennoch ist es gerade die anspruchsvolle Kunst der Systemanalyse, auch bei komplexen Gebilden etwa Störungen zu lokalisieren und latente Funktionen zu identifizieren (Sonntag u. Schaper, 1997).

(7) Man braucht viele Versuchspersonen, um die kleinen Effekte nachzuweisen, die der Denkpsychologie wichtig sind.

Tatsächlich benötigt man viele Beobachtungen, wenn die erwartete Effektstärke klein ist. Aber bereits Bredenkamp (1972, S. 180 f.) räumt $N = 1$-Experimenten einen eigenständigen Platz ein, auch wenn er sofort einschränkt: »Selbstverständlich soll der Hinweis auf solche Experimente kein Plädoyer für das zukünftige Experimentieren mit nur einer einzigen Vp sein. Er soll nur zeigen, dass es genügend Experimente gibt, in denen das

zur Prüfung einer Hypothese relevante Datum an *jeder* Vp gewonnen wird.« Für die denkpsychologische Forschung bedeutet dies: Einzelfälle sind nicht nur zur Hypothesengenerierung zu verwenden (Kluwe, 1995), sondern haben auch in der Hypothesentestung ihren Platz. Die von Hussy und Jain (2002, S. 25) vorgenommene Gegenüberstellung von Experiment und Einzelfallanalyse ist daher zwar prinzipiell berechtigt, legt aber den Fehlschluss nahe, Einzelfälle könnten nicht ebenfalls interessante Gegenstände experimenteller Untersuchungen sein (eine experimentelle theorietestende Untersuchung aus dem Bereich Zeitschätzung mit $N = 1$ findet sich z. B. auch bei Funke, 1988).

(8) Problemlösendes Denken (insbesondere den Umgang mit komplexen Problemen) kann man nicht lernen.

Tatsächlich zeigen sich gerade im Kontext komplexer Anforderungen die Schwierigkeiten pädagogischer Maßnahmen. »Kritisches Denken« (Dauer, 1989; Halpern, 1989) kann man nicht einfach unterrichten. Hinsichtlich des Umgangs mit komplexen Problemen ist Dörner (1989b, S. 305) zuversichtlich, durch Verwendung des Zeitraffers, den computersimulierte Szenarien inhärent aufweisen, multiple Lernerfahrungen durch eine »Anforderungssymphonie« unterschiedlichster Szenarien möglich zu machen: »Spielen war immer eine wichtige Methode zur Vorbereitung auf den Ernstfall. Man sollte es in gezielter Weise verwenden. Wir haben dafür heute viel bessere Möglichkeiten als früher. Wir sollten sie nutzen!«

(9) Bildgebende Verfahren erlauben neue Einblicke in die Arbeit des Gehirns. Die bisherige kognitive Theoriebildung wird abgelöst durch neuropsychologische Konzepte.

Die Möglichkeiten bildgebender Verfahren sind in den letzten Jahren enorm gewachsen (Posner u. Raichle, 1994). Die Erwartungen sind geweckt, das Gehirn »live in action« zu beobachten. Werden wir ihm demnächst beim problemlösenden Denken zusehen können? Und falls ja, was werden wir sehen? Op de Beeck et al. (2001) haben in einer kritischen Übersicht vor falschen Erwartungen gewarnt und schreiben:

»... if one wants to construct a cognitive model of behavior, then a mere localization of these processes is not that important (contrary to its importance for neuropsychology). For a cognitive scientist, it is important which cognitive processes are involved in a task and how these are computed, but it does not matter where they are computed ...« (S. 344).

Die alte Diskussion um die richtige Analyseebene, die von Marr (1982) bereits im Kontext seiner Arbeiten zu KI-Modellen des Sehens aufgeworfen wurde und die in der Kontroverse um konnektionistische versus symbolistische Ansätze heftig entfacht wurde (vgl. Broadbent, 1985; McClelland, 1988; McClelland u. Rumelhart, 1985; McCloskey, 1991), taucht hier erneut auf. Bringen Erkenntnisse über neuronale Strukturen tatsächlich Licht in Theorien, die auf einer viel abstrakteren Analyseebene formuliert sind? Die Antwort wird auch hier im Fall der bildgebenden Verfahren ähnlich sein wie bei der Debatte um konnektionistische Systeme, also darauf hinauslaufen, dass eine Analyse konkreter Implementationen natürlich spannend ist, dass aber die letztendlich wichtigen Fragen abstrakter gestellt und deshalb nicht durch diese Erkenntnisse auf tieferer Ebene beantwortet werden können.

Bausteine für eine Theorie des Komplexen Problemlösens

Mit den nachfolgend dargestellten grundlegenden Annahmen werden Eckpunkte für eine deduktivistisch prüfbare Theorie des Komplexen Problemlösens skizziert, in der es um die Frage geht, wie Menschen komplexe Probleme erkennen, bearbeiten und lösen. Als komplexes Problem wird eine Nicht-Routinesituation bezeichnet, in der die problemlösende Person Ziele in einer partiell intransparenten, dynamischen, polytelischen Umgebung verfolgt.[2] Eine derartige Situation liegt etwa beim Bearbeiten von Dynamis-Szenarien vor (siehe dazu Funke, 2001a, 2001b).

(1) *Repräsentationsannahme:* Es wird angenommen, dass die problemlösende Person eine (kognitive) Repräsentation des zu bearbeitenden Systems erstellt.

Diese allgemeine Annahme lässt zunächst offen, in welchem Format eine derartige Repräsentation erfolgt und wie vollständig beziehungsweise exakt diese ausfällt. Der Zusatz »kognitiv« soll darauf hinweisen, dass es sich um eine *interne* Repräsentation handelt und nicht um ein extern vorliegendes Modell. Ob der Erwerb explizit oder implizit erfolgt, ob es bildliche, konkrete und/oder abstrakte Gedächtnisspuren gibt, sollte möglichst genau beschrieben werden. Die Annahme geht zurück auf die Theorie des Problemraums (Newell et al., 1958; Newell u. Simon, 1972), den die problemlösende

2 Diese Überlegungen knüpfen an Darlegungen zu Theorie-Bausteinen an, die ursprünglich bei Funke (1986, S. 107 f.) formuliert wurden.

Person repräsentiert und absucht. Dieser Raum sollte auch bei komplexen Szenarien möglichst präzise skizziert werden, wobei dies durchaus unvollständig bleiben kann (siehe Annahme 4), worauf bereits Kaplan und Simon (1990) hingewiesen haben. Wissenschaftstheoretisch gesehen handelt es sich bei dieser Annahme nicht einfach nur um eine prüfbare empirische Hypothese, sondern um eine Voraussetzung, deren Erfüllung für die weiteren Annahmen unterstellt wird.

(2) *Sequenzannahme:* Es wird angenommen, dass sich diese Repräsentation schrittweise entwickelt, und zwar beginnend mit den starken, salienten Systemeffekten und dann absteigend zu den schwächeren Effekten.

Bei dieser Annahme geht es um den Wissens*erwerb* als Prozess. Die unter (1) postulierte Repräsentation kommt nicht nach einem Alles-oder-Nichts-Prinzip zustande, ist nicht auf ein Mal vorhanden, sondern wird entsprechend der Stärke und Auffälligkeit der implementierten Effekte sukzessive aufgebaut. Gerade bei komplexeren Kausalmodellen, in denen eine Reihe einzelner Effekte in ein Gesamtmodell integriert werden müssen, kommt es zu einem stufenweisen Aufbau dieses Gesamtmodells. Dies ließe sich beispielsweise in Form eines Produktionssystems darstellen, bei dem zunehmend mehr Wenn-dann-Regeln hinzukommen. Ausgehend von den Kausalannahmen, die aufgrund der Semantik nahe liegen (z. B. Verabreichung von Dünger führt zu Pflanzenwachstum), über Kausaleffekte, die aufgrund ihrer Stärke leicht zu erkennen sind, geht es bis zu den schwachen oder wenig salienten Effekten, die erst am Ende in das Gesamtmodell integriert werden.

(3) *Ökonomieannahme:* Es wird angenommen, dass Problemlöser in derartigen Situationen versuchen, mit einem möglichst sparsamen Modell auszukommen. In solchen einfachen Modellen werden beispielsweise direkte Effekte den komplizierteren indirekten Effekten vorgezogen.

Diese Annahme soll zum Beispiel die Kompensationshypothese von Funke und Müller (1988) erklären, wonach Versuchspersonen faktisch vorhandene, komplizierte indirekte Effekte durch einfachere Haupteffekt-Modelle zu erklären versuchen. Solange sich die durch solche fehlerhaft gebildeten subjektiven Modelle verursachten Prädiktionsfehler in Grenzen halten, bleiben Versuchspersonen beim einfacheren falschen Modell. Diese Annahme passt übrigens auch zur Theorie deklarativer Vereinfachung (Klauer, 1993), nach der es bei Überlast der Verarbeitungskapazität nicht bei prozeduralen, sondern bei deklarativen Aspekten des Problems zu Vereinfachungen kommt.

(4) *Unvollständigkeitsannahme:* Es wird angenommen, dass Versuchspersonen kein vollständiges Systembild brauchen, um handlungsfähig zu sein. Je nach den geforderten Steuerungsleistungen und je nach Anspruchsniveau können auch unvollständige subjektive Modelle die Handlungsfähigkeit sicherstellen.

Die hier aufgestellte Vermutung über eine mögliche Unschärfe und Unvollständigkeit subjektiver Modelle hat mit der Tatsache zu tun, dass auch unvollständige Modelle nützlich sein können (vgl. Johnson-Laird, 1983, S. 474). So benötige ich als Autofahrer keineswegs ein vollständiges Wissen über die Eigenarten eines Verbrennungsmotors – ein paar wesentliche Aspekte (benötigt Treibstoff, erzeugt Hitze und Abgas, wird zum Antrieb benötigt, erzeugt über eine Lichtmaschine Strom etc.) reichen für den Alltag durchaus. Im Fall einer Störung oder dem Wunsch nach einer Konstruktionsoptimierung reicht dies allerdings nicht.

(5) *Kontextannahme:* Es wird angenommen, dass die problemlösende Person im Fall semantisch eingekleideter Systeme versucht, so weit wie möglich ihr vorhandenes Vorwissen zu nutzen. Bereits vor jeder Exploration des Systems bestehen dadurch erste Modellannahmen, die die nachfolgende Exploration und Steuerung leiten.

Diese Annahme greift Erkenntnisse der Schematheorie auf, wonach etwa Textverstehen durch die in Überschriften aktivierten Schemata geleitet wird (vgl. Bransford, 1979; Bransford u. Johnson, 1972). Allgemein ist davon auszugehen, dass jede Art von Kontext, in die ein System eingebettet wird, zugehöriges Vorwissen aktiviert. Selbst abstrakte Darbietungen eines Systems als »Maschine« (vgl. Beckmann, 1994) sind keinesfalls vorwissensneutral, da bereits die Bezeichnung Maschine ein bestimmtes Schema aktiviert (unbelebt, funktional, mit Ein- und Ausgabeelementen versehen etc.).

(6) *Gedächtnisannahme:* Es wird angenommen, dass für ein gebildetes, subjektives Modell prinzipiell alle bekannten Gedächtniseffekte kurz- und langfristiger Speicher gelten können, also Vergessen, Interferenz mit anderen Modellen, Übungseffekte und so weiter.

Mit dieser Annahme werden subjektive Modelle über dynamische Systeme auf die gleiche Ebene wie andere Gedächtnisobjekte gestellt. Für subjektive Modelle gelten die gleichen gedächtnispsychologischen Effekte wie für Wissen über bestimmte Gegenstandsbereiche – es gibt keinen Grund, diesen Modellen eine Sonderstellung hinsichtlich ihrer Lern- und Vergessenscharakteristika einzuräumen. Sowohl die Theorie des Arbeitsgedächtnisses

(Baddeley, 1986, 2001) als auch die speziell auf Problemlösungsprozesse zugeschnittene Theorie kognitiver Belastung (Sweller, 1988) sind einschlägig. Die Rolle der Verarbeitungskapazität als Leistungsprädiktor beim Komplexen Problemlösen ist vielfältig nachgewiesen (Hussy, 1991; Süß, 1999). Je genauer man sich beim subjektiven Modell auf bestimmte Repräsentationsformen festlegen kann, umso mehr empirischen Gehalt bekommen entsprechende Gedächtnisannahmen.

Konsequenzen. Aus dem eben skizzierten Annahmekomplex lassen sich bei entsprechender Konkretisierung eine Reihe von prüfbaren Hypothesen ableiten und in experimentellen Anordnungen testen. Zunächst sollen Aussagen aufgeführt werden, die bei Gültigkeit der Annahmen verboten sind. Solche Verbote sind aus kritisch-rationaler Sicht besonders interessant, da sich gezielt überprüfen lässt, ob die Verbote gelten – falls nicht, hat man damit wichtige Bestandteile der theoretischen Annahmen falsifiziert.

Folgende beispielhaft aufgeführte Beobachtungen sind nach dem eben beschriebenen Annahmengefüge verboten:

- Eine Versuchsperson bearbeitet ein System, ohne eine Repräsentation desselben zu erstellen (verletzt Repräsentationsannahme).
- Eine Versuchsperson repräsentiert zunächst schwache, wenig saliente und dann erst starke, saliente Effekte (verletzt Sequenzannahme).
- Eine Versuchsperson bildet kompliziertere Modelle als sie tatsächlich implementiert sind (verletzt Ökonomieannahme).
- Eine Versuchsperson ist erst handlungsfähig, wenn sie eine komplette Repräsentation des Systems erstellt hat (verletzt Unvollständigkeitsannahme).
- Eine Versuchsperson unterscheidet sich in ihrer Identifikations- und Steuerleistung nicht danach, ob das System abstrakt oder semantisch eingekleidet ist (verletzt Kontextannahme).
- Eine Versuchsperson erlebt keine Interferenz beim Bearbeiten zweier verschiedener Systeme (verletzt Gedächtnisannahme).

Für alle hier aufgeführten Verbote lassen sich Versuchsanordnungen herstellen, in denen zumindest prinzipiell die Verletzung der jeweiligen Annahme auftreten könnte. Mit der Ableitung statistischer Hypothesen aus den psychologischen Hypothesen wäre ein Schritt in Richtung deduktivistischen Vorgehens vollzogen. Richtig interessant würde es, wenn man überraschende, nichttriviale Vorhersagen aus entsprechend spezifizierten Annahmen ableiten könnte.

Abschluss

Die hier vorgestellten Überlegungen sollten deutlich machen, dass eine deduktivistisch prüfbare Theorie zum Komplexen Problemlösen nicht nur wünschenswert, sondern auch prinzipiell möglich ist. Neben dem Aufräumen mit einigen Mythen, die eine derartige Theoriebildung bremsen oder gar verhindern, zeigen die am Ende vorgestellten Bausteine einen gangbaren Weg auf, den ich für vielversprechend halte. Selbstverständlich sind Bausteine für eine Theorie noch nicht die Theorie selbst, und es mag unklar sein, wie der Weg von den Bausteinen zu einer prüfbaren Theorie führt. Aber man sollte mit kleinen Schritten beginnen, wie sie hier vollzogen wurden. Zusammen mit einem systematischen Forschungsprogramm, das sich auf die Suche nach »verbotenen« Beobachtungen macht, könnte damit die Basis für einen Gegenstandsbereich gelegt werden, der aus der Phase des Schmetterlingsammelns heraus- und in die Phase des Konstruierens und Testens empirisch gehaltvoller Theorien hineingeführt werden sollte. Ein deduktivistisch orientiertes Vorgehen scheint mir dabei möglich und sinnvoll zugleich.

Literatur

Anderson, J.R., Lebiere, C. (Hg.) (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Baddeley, A.D. (1986). *Working memory.* Oxford: Oxford University Press.

Baddeley, A.D. (2001). Is working memory still working? *American Psychologist, 56,* 851–864.

Beckmann, J.F. (1994). *Lernen und komplexes Problemlösen. Ein Beitrag zur Konstruktvalidierung von Lerntests.* Bonn: Holos.

Berry, D.C., Broadbent, D.E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology, 36A,* 209–231.

Berry, D.C., Broadbent, D.E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79,* 251–272.

Berry, D.C., Broadbent, D.E. (1995). Implicit learning in the control of complex systems. In P.A. Frensch u. J. Funke (Hg.), *Complex problem solving: The European perspective* (S. 131–150). Hillsdale, NJ: Erlbaum.

Bransford, J.D. (1979). *Human cognition: Learning, understanding, and remembering.* Belmont, CA: Wadsworth.

Bransford, J.D., Johnson, M.K. (1972). Contextual prerequisites for understanding: Some investigators of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717–726.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt a. M.: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1990). Kognitionspsychologische Untersuchungen eines Rechenkünstlers. In H. Feger (Hg.), *Wissenschaft und Verantwortung. Festschrift für Karl Josef Klauer* (S. 47–70). Göttingen: Hogrefe.

Brehmer, B., Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior, 9,* 171–184.

Brehmer, B., Leplat, J., Rasmussen, J. (1991). Use of simulation in the study of complex decision making. In J. Rasmussen, B. Brehmer u. J. Leplat (Hg.), *Distributed decision making: Cognitive models for cooperative work* (S. 373–386). New York: Wiley.

Broadbent, D.E. (1977). Levels, hierarchies, and the locus of control. *Quarterly Journal of Experimental Psychology, 29,* 181–201.

Broadbent, D.E. (1985). A question of levels: Comments on McClelland and Rumelhart. *Journal of Experimental Psychology: General, 114,* 189–192.

Broadbent, D.E., Aston, B. (1978). Human control of a simulated economic system. *Ergonomics, 21,* 1035–1043.

Broadbent, D.E., Fitzgerald, P., Broadbent, M.H.P. (1986). Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology, 77,* 33–50.

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P.A. Frensch u. J. Funke (Hg.), *Complex problem solving: The European perspective* (S. 27–63). Hillsdale, NJ: Erlbaum.

Buchner, A., Funke, J., Berry, D.C. (1995). Negative correlations between control performance and verbalizable knowledge: Indicators for implicit learning in process control tasks? *Quarterly Journal of Experimental Psychology, 48A,* 166–187.

Bühler, K. (1908). Antwort auf die von W. Wundt erhobenen Einwände gegen die Methode der Selbstbeobachtung an experimentell erzeugten Erlebnissen. *Archiv für die Gesamte Psychologie, 12,* 93–112.

Clark, A. (1997). *Being there. Putting brain, body, and world together again.* Cambridge, MA: MIT Press.

Dauer, F.W. (1989). *Critical thinking. An introduction to reasoning.* New York: Oxford University Press.

Dörner, D. (1975). Wie Menschen eine Welt verbessern wollten. *Bild der Wissenschaft, 12,* 48–53.

Dörner, D. (1981). Über die Schwierigkeiten menschlichen Umgangs mit Komplexität. *Psychologische Rundschau, 32,* 163–179.

Dörner, D. (1983). Empirische Psychologie und Alltagsrelevanz. In G. Jüttemann (Hg.), *Psychologie in der Veränderung. Perspektiven für eine gegenstandsangemessene Forschungspraxis* (S. 13–29). Weinheim: Beltz.

Dörner, D. (1986). Diagnostik der operativen Intelligenz. *Diagnostica, 32,* 290–308.

Dörner, D. (1989a). Die kleinen grünen Schildkröten und die Methoden der experimentellen Psychologie. *Sprache & Kognition, 8,* 86–97.

Dörner, D. (1989b). *Die Logik des Mißlingens. Strategisches Denken in komplexen Situationen.* Hamburg: Rowohlt.

Dörner, D. (1992). Über die Philosophie der Verwendung von Mikrowelten oder »Computerszenarios« in der psychologischen Forschung. In H. Gundlach (Hg.), *Psychologische Forschung und Methode: Das Versprechen des Experiments. Festschrift für Werner Traxel* (S. 53–87). Passau: Passavia-Universitäts-Verlag.

Dörner, D. (1993). Wissen, Emotionen und Handlungsregulation oder Die Vernunft der Gefühle. *Zeitschrift für Psychologie, 201,* 167–202.

Dörner, D. (1996). Verhalten und Handeln. In D. Dörner u. H. Selg (Hg.), *Psychologie.*

Eine Einführung in ihre Grundlagen und Anwendungsfelder (S. 100–114). Stuttgart: Kohlhammer.

Dörner, D. (1998). Emotionen, kognitive Prozesse und der Gebrauch von Wissen. In F. Klix u. H. Spada (Hg.), *Wissen* (S. 301–333). Göttingen: Hogrefe.

Dörner, D. (1999). *Bauplan für eine Seele.* Reinbek bei Hamburg: Rowohlt.

Dörner, D., Bartl, C., Detje, F., Gerdes, J., Halcour, D., Schaub, H. et al. (2002). *Die Mechanik des Seelenwagens. Eine neuronale Theorie der Handlungsregulation.* Bern: Huber.

Dörner, D., Kreuzig, H.W. (1983). Problemlösefähigkeit und Intelligenz. *Psychologische Rundschau, 34,* 185–192.

Dörner, D., Kreuzig, H.W., Reither, F., Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität.* Bern: Huber.

Dörner, D., Reither, F., Stäudel, T. (1983). Emotion und problemlösendes Denken. In H. Mandl u. G.L. Huber (Hg.), *Emotion und Kognition* (S. 61–81). München: Urban & Schwarzenberg.

Dörner, D., Schaub, H., Stäudel, T., Strohschneider, S. (1988). Ein System zur Handlungsregulation oder – Die Interaktion von Emotion, Kognition und Motivation. *Sprache & Kognition, 7,* 217–232.

Dörner, D., Schaub, H., Strohschneider, S. (1999). Komplexes Problemlösen – Königsweg der Theoretischen Psychologie? *Psychologische Rundschau, 50,* 198–205.

Dörner, D., Wearing, A.J. (1995). Complex problem solving: Toward a (§computersimulated) theory. In P.A. Frensch u. J. Funke (Hg.), *Complex problem solving: The European perspective* (S. 65–99). Hillsdale, NJ: Erlbaum.

Frensch, P.A., Buchner, A. (1999). Domain-generality versus domain-specificity: A binary concept and its impact on psychological theory and research. In R.J. Sternberg (Hg.), *The nature of cognition* (S. 137–172). Cambridge, MA: MIT Press.

Funke, J. (1984). Diagnose der westdeutschen Problemlöseforschung in Form einiger Thesen. *Sprache & Kognition, 3,* 159–172.

Funke, J. (1986). *Komplexes Problemlösen – Bestandsaufnahme und Perspektiven.* Heidelberg: Springer-Verlag.

Funke, J. (1988). »Changes« oder »Effort«? Eine Überprüfung von zwei Theorien zur Zeitschätzung. *Zeitschrift für Experimentelle und Angewandte Psychologie, 35,* 218–241.

Funke, J. (1995). Experimental research on complex problem solving. In P.A. Frensch u. J. Funke (Hg.), *Complex problem solving: The European perspective* (S. 243–268). Hillsdale, NJ: Erlbaum.

Funke, J. (1998). Computer-based testing and training with scenarios from complex problem-solving research: Advantages and disadvantages. *International Journal of Selection and Assessment, 6,* 90–96.

Funke, J. (2001a). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning, 7,* 69–89.

Funke, J. (2001b). Neue Verfahren zur Erfassung intelligenten Umgangs mit komplexen und dynamischen Anforderungen. In E. Stern u. J. Guthke (Hg.), *Perspektiven der Intelligenzforschung. Ein Lehrbuch für Fortgeschrittene* (S. 89–107). Lengerich: Pabst Science Publishers.

Funke, J. (2003). *Problemlösendes Denken.* Stuttgart: Kohlhammer.

Funke, J., Müller, H. (1988). Eingreifen und Prognostizieren als Determinanten von Systemidentifikation und Systemsteuerung. *Sprache & Kognition, 7,* 176–186.

Graumann, C.F. (1997). Zur Ökologie des Gedächtnisses. In G. Lüer u. U. Lass (Hg.), *Erinnern und Behalten. Wege zur Erforschung des menschlichen Gedächtnisses* (S. 269–286). Göttingen: Vandenhoeck & Ruprecht.

Graumann, C.F. (2002). The phenomenological approach to people-environment studies. In R.B. Bechtel u. A. Churchman (Hg.), *Handbook of Environmental Psychology* (S. 95–113). New York: Wiley.

Halpern, D.F. (1989). *Thought and knowledge. An introduction to critical thinking.* Hillsdale, NJ: Erlbaum.

Hayes, N.A., Broadbent, D.E. (1988). Two modes of learning for interactive tasks. *Cognition, 28,* 249–276.

Hussy, W. (1991). Komplexes Problemlösen und Verarbeitungskapazität. *Sprache & Kognition, 10,* 208–220.

Hussy, W., Jain, A. (2002). *Experimentelle Hypothesenprüfung in der Psychologie.* Göttingen: Hogrefe.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Johnson-Laird, P.N. (1983). *Mental models. Towards a cognitive science of language, inference, and consciousness.* Cambridge: Cambridge University Press.

Kaplan, C.A., Simon, H.A. (1990). In search of insight. *Cognitive Psychology, 22,* 374–419.

Kersting, M. (1999). *Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien? Zur Kriteriumsvalidität von Problemlöseszenarien und Intelligenztests.* Göttingen: Hogrefe.

Klauer, K.C. (1993). *Belastung und Entlastung beim Problemlösen. Eine Theorie des deklarativen Vereinfachens.* Göttingen: Hogrefe.

Kluwe, R.H. (1995). Single case studies and models of complex problem solving. In P.A. Frensch u. J. Funke (Hg.), *Complex problem solving: The European perspective* (S. 269–291). Hillsdale, NJ: Erlbaum.

Kotkamp, U. (1999). *Elementares und komplexes Problemlösen: Über Invarianzeigenschaften von Denkprozessen.* Lengerich: Pabst Science Publishers.

Kuhl, J. (1983). Emotion, Kognition und Motivation. II: Die funktionale Bedeutung der Emotionen für das problemlösende Denken und für das konkrete Handeln. *Sprache & Kognition, 2,* 228–253.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information.* New York: W.H. Freeman.

McClelland, J.L. (1988). Connectionist models and psychological evidence. *Journal of Memory and Language, 27,* 107–123.

McClelland, J.L., Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science, 2,* 387–395.

Medler, D.A. (1998). A brief history of connectionism. *Neural Computing Surveys, 1,* 61–101.

Newell, A., Shaw, J.C., Simon, H.A. (1958). Elements of a theory of human problem-solving. *Psychological Review, 65,* 151–166.

Newell, A., Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Norman, D.A. (1993). Cognition in the head and in the world: An introduction to the special issue on situated action. *Cognitive Science, 17,* 1–6.

op de Beeck, H., Wagemans, J., Vogels, R. (2001). Can neuroimaging really tell us what the human brain is doing? The relevance of indirect measures of population activity. *Acta Psychologica, 107,* 323–351.

Otto, J.H., Döring-Seipel, E., Lantermann, E.-D. (2002). Zur Bedeutung von subjektiven, emotionalen Intelligenzkomponenten für das komplexe Problemlösen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23,* 417–433.

Posner, M.I., Raichle, M.E. (1994). *Images of mind.* New York: W.H. Freeman.

Rigas, G., Brehmer, B. (1999). Mental processes in intelligence tests and dynamic decision making tasks. In P. Juslin u. H. Montgomery (Hg.), *Judgement and decision making: Neo-Brunswikean and process-tracing approaches* (S. 45–65). Hillsdale, NJ: Erlbaum.

Rigas, G., Carling, E., Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence, 30,* 463–480.

Schaub, H. (1990). Die Situationsspezifität des Problemlöseverhaltens. *Zeitschrift für Psychologie, 198,* 83–96.

Schaub, H. (2001). *Persönlichkeit und Problemlösen. Persönlichkeitsfaktoren als Parameter eines informationsverarbeitenden Systems.* Weinheim: Beltz Psychologie Verlags Union.

Schenck, W. (2001). *A connectionist approach to human planning.* Retrieved 11.3.2003, from http://www.ub.uni-heidelberg.de/archiv/1428.

Simon, H.A. (1979/1967). Motivational and emotional controls of cognition. In H.A. Simon (Hg.), *Models of thought* (S. 29–38). New Haven: Yale University Press.

Sonntag, K., Schaper, N. (Hg.). (1997). *Störungsmanagement und Diagnosekompetenz. Leistungskritisches Denken und Handeln in komplexen technischen Systemen.* Zürich: vdf Hochschulverlag.

Spering, M., Wagener, D., Funke, J. (under revision). Control beliefs moderate emotion influences on complex problem solving. *Cognition and Emotion.*

Spies, K. (1990). Einfluß von Emotionen auf die Ziel- und Handlungsauswahl. *Zeitschrift für Experimentelle und Angewandte Psychologie, 37,* 124–152.

Spies, K. (1999). Negative Stimmung, kognitive Verarbeitungskapazität und ablenkende Wirkung der Bearbeitung kognitiver Aufgaben. In W. Hacker u. M. Rinck (Hg.), *Zukunft gestalten. Bericht über den 41. Kongreß der Deutschen Gesellschaft für Psychologie in Dresden 1998* (S. 527–533). Lengerich: Pabst Science Publishers.

Spies, K., Hesse, F.W. (1986). Interaktion von Emotion und Kognition. *Psychologische Rundschau, 37,* 75–90.

Sternberg, R.J. (1995). Expertise in complex problem solving: A comparison of alternative conceptions. In P.A. Frensch u. J. Funke (Hg.), *Complex problem solving: The European perspective* (S. 295–321). Hillsdale, NJ: Erlbaum.

Sternberg, R.J., Grigorenko, E.L. (2001). Unified Psychology. *American Psychologist, 56,* 1069–1079.

Strauß, B., Kleinmann, M. (Hg.). (1995). *Computersimulierte Szenarien in der Personalarbeit.* Göttingen: Verlag für Angewandte Psychologie.

Strohschneider, S. (1986). Zur Stabilität und Validität von Handeln in komplexen Realitätsbereichen. *Sprache & Kognition, 5,* 42–48.

Strohschneider, S. (1991). Problemlösen und Intelligenz: Über die Effekte der Konkretisierung komplexer Probleme. *Diagnostica, 37,* 353–371.

Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen – Perspektiven für eine Kooperation zwischen differentiell-psychometrischer und kognitionspsychologischer Forschung. *Psychologische Rundschau, 50,* 220–228.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285.

Vollmeyer, R., Funke, J. (1999). Personen- und Aufgabenmerkmale beim komplexen Problemlösen. *Psychologische Rundschau, 50,* 213–219.

Wallach, D. (1998). *Komplexe Regelungsprozesse. Eine kognitionswissenschaftliche Analyse.* Wiesbaden: Deutscher Universitäts-Verlag.

Wittmann, W.W., Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge and complex problem solving: Performances via Brunswik-symmetry. In P.L. Ackerman, P.C. Kyllonen u. R.D. Roberts (Hg.), *Learning and indi-*

vidual differences: Process, trait, and content (S. 77–108). Washington: American Psychological Association.

Zsambok, C.E. (1997). Naturalistic decision making: Where are we now? In C.E. Zsambok u. G.A. Klein (Hg.), *Naturalistic decision making* (S. 3–16). Mahwah, NJ: Erlbaum.

Edeltrud Marx

Methoden der Erhebung von Versprecherdaten in der Sprachproduktionsforschung: Probleme und Lösungsansätze

»... if errors do not reflect the nature
of spontaneous speech processes, nothing does«
(Cutler u. Fay, 1978).

Zusammenfassung

Analysen von Versprecherdaten spielen nach wie vor eine wichtige Rolle bei der Erforschung von Sprachproduktionsprozessen. Im folgenden Beitrag werden Vor- und Nachteile dieser Methode diskutiert und Möglichkeiten erörtert, ihre Objektivität, Reliabilität und Validität zu erhöhen.

Einleitung

Sprechen ist eine unserer faszinierendsten und wichtigsten kognitiven Leistungen. Seit über hundert Jahren werden die damit verbundenen Prozesse zwischen der Entstehung eines Gedankens und seiner Artikulation untersucht. Die Analyse von Versprecher- oder Sprechfehlerdaten spielt dabei eine zentrale Rolle. Sie gilt nach wie vor als eine wichtige Methode bei der Erforschung von Sprachproduktionsprozessen, insbesondere lexikalischen Prozessen (Cutler, 1982, 1988; Stemberger, 1992; Levelt, 1999). Im Folgenden werden methodische Probleme diskutiert, die bei der Erhebung und Auswertung von Sprechfehlern auftreten. Dabei werden Ergebnisse aus Sprachwahrnehmungsexperimenten, Sprechfehleranalysen sowie Sprechfehlerinduktionsexperimenten berücksichtigt und Möglichkeiten erörtert, die Objektivität, Reliabilität und Validität von Versprecherdaten zu erhöhen.

Sprechfehlerdaten sind objektiv, wenn Art und Anzahl von Versprechern pro Zeitstichprobe nicht zwischen verschiedenen Beobachtern variieren. Ein Beispiel für objektiv erhobene Versprecherdaten ist die Sammlung von

Meringer und Mayer (1895), die nur solche Versprecher enthält, die von mindestens zwei Beobachtern gleichzeitig gehört und notiert worden waren. Ein Versprecherkorpus enthält – in signalentdeckungstheoretischer Terminologie ausgedrückt – allerdings nur Treffer (d. h., produzierte Versprecher werden vom Hörer wahrgenommen) und falsche Alarme (korrekte Äußerungen werden fälschlich als Versprecher gewertet), nicht jedoch die Anzahl der Verpasser (produzierte Versprecher werden nicht als solche wahrgenommen) und der korrekten Zurückweisungen (korrekte Äußerungen werden als solche erkannt). Damit fehlt die Grundlage für ein geeignetes Messmodell, das eine Beziehung zwischen der Verteilung tatsächlich produzierter und wahrgenommener Versprecher herstellen könnte.

Um der Tatsache Rechnung zu tragen, dass nicht überprüft werden kann, ob die Verteilung im Korpus Voraussetzungen für parametrische Tests erfüllt, empfehlen sich bei der Analyse spontaner Versprecher nonparametrische Verfahren, die wie der Vorzeichentest auswerterunabhängige Aspekte der Daten (Richtung des Unterschiedes zwischen experimentellen Bedingungen, nicht aber dessen Größe) analysieren.

Die meisten Phänomene, die bei spontanen Versprechern beobachtet werden, lassen sich experimentell replizieren, was nach Stemberger (1992) die Annahme einer hohen Reliabilität spontaner Daten begründet.

Bekanntermaßen ist Reliabilität zwar eine notwendige, nicht aber hinreichende Bedingung für die Validität von Versprecherdaten. Die Frage nach der Konstruktvalidität zielt damit darauf ab, ob Merkmale bestimmter Versprecher genau die latenten Sprachproduktionsprozesse oder -strukturen widerspiegeln, die ihnen zugeschrieben werden. Zugrunde liegende Konstrukte sind die Planungsprozesse, die der Artikulation einer Äußerung vorangehen. In der Versprecherforschung wird implizit davon ausgegangen, dass die Prozesse, die sich aus Versprechern ableiten lassen, die sind, die auch bei fehlerfreier Sprachproduktion stattfinden. Diese Annahme wird von Schriefers (2003) zumindest zur Diskussion gestellt.

Im Wesentlichen verfolgt die Versprecherforschung zwei Ziele: erstens die Erforschung von Strukturen und Prozessen innerhalb des Sprachverarbeitungssystems. Zweitens die Erforschung des Sprachverarbeitungssystems innerhalb des allgemeinen kognitiven Systems. Nahezu alles, was wir heute über diese Fragen wissen, wurde zunächst aus Versprechern abgeleitet.

Sprechfehler oder Versprecher werden als unbeabsichtigte Äußerungen definiert und in Abgrenzung von Sprachstörungen pathologischen Ursprungs als »Störungen des normalen Sprechens« aufgefasst (Pechmann, 1994). Von eindeutigen Fällen abgesehen, ist es nicht immer einfach zu ent-

scheiden, was ein Sprecher zu sagen beabsichtigte. Am validesten sind dabei seine eigenen Angaben oder Nachweise aus Selbstkorrekturen.

Sprechfehler sind keine Regelfehler. Sie betreffen nicht die sprachliche Kompetenz, sondern die Performanz: Sofern ein Sprecher seine Äußerung kontrolliert, bemerkt er, dass er sich versprochen hat, was er bei Regelfehlern, die aufgrund mangelnder Kompetenz entstehen, nicht bemerkt.

Versprecher üben eine so große und anhaltende Faszination aus, dass sie immer wieder Gegenstand von Forschungsaktivitäten werden. Sie sind häufig der Ausgangspunkt oder Anlass, eine bestimmte Frage in den Mittelpunkt der Aufmerksamkeit zu rücken. So hat Berg (1992) bei seinen Analysen spontaner deutscher Versprecher festgestellt, dass intendiertes und fehlerhaftes Nomen überzufällig häufig dasselbe Genus aufweisen. Das ist auch im folgenden Beispiel aus seiner Sammlung der Fall: »Ich weiß, dass die *Regel* _ *Rede* war vom 11. Mai.« Diese Entdeckung hat bis heute eine Fülle von Untersuchungen zur Genusverarbeitung in der Sprachproduktion nach sich gezogen. Vor allem aber haben die unterschiedlichen statistischen Verteilungen in den Sprechfehlerkorpora von Garrett (z. B. 1980) versus Dell und Reich (1981) eine Flut von chronometrischen Studien zur Untersuchung lexikalischer Zeitverläufe ausgelöst. Sie haben gezeigt, dass eine Theorie der Sprachproduktion nicht allein auf Befunden aus gestörten Produktionsprozessen basieren sollte (vgl. Meyer, 1992).

Das Analysieren von Versprechern umfasst mittlerweile nicht nur gesammelte spontane, sondern auch experimentell erzeugte Sprechfehler sowie Daten aus rechnergesteuerten Simulationen.

Erhebung und Kategorisierung natürlicher Versprecher

Als Versprecherforscher braucht man nicht mehr als ein gutes Gefühl für Sprache, einen Bleistift und Papier; mit diesen Dingen ausgestattet, kann man eine enorme Spannbreite von Fragen untersuchen (Bock, 1996). Allerdings braucht man auch Geduld, denn die Fehlerrate beim Sprechen ist erstaunlich gering: Levelt (1989) schätzt sie auf 1:1000. Dementsprechend braucht es Zeit, bis ein Versprecherkorpus den Umfang aufweist, der erforderlich ist, um Analysen durchführen zu können. Wir wissen heute, dass Versprecher nicht zufällig entstehen, sondern Gesetzmäßigkeiten folgen. Die Basisannahme der Sprechfehlerforschung ist, dass diese Gesetzmäßigkeiten auch dem störungsfreien Sprechen zugrunde liegen. Vertauschungsfehler wie »Der dickste Bauer hat die dümmsten Kartoffeln« anstatt »Der

dümmste Bauer hat die dicksten Kartoffeln« hinterlassen charakteristische Spuren an der sprachlichen Oberfläche, die Rückschlüsse auf zugrunde liegende Prozesse ermöglichen. Im Beispiel werden zwei Wörter (dumm und dick) miteinander vertauscht. Wie bei nahezu allen Wortvertauschungsfehlern beobachtbar, gehören die beiden Wörter derselben grammatischen Wortklasse (Adjektiv) an. Dieser so genannte Wortklasseneffekt hat Sprachproduktionstheoretiker wie Garrett (z. B. 1980) zu der Annahme veranlasst, dass die Vertauschung von Wörtern in einer Prozessphase geschieht, in der die syntaktische Information von Wörtern verarbeitet wird. An dem Versprecher fällt weiterhin auf, dass die Numerusinformation nicht mit den vertauschten Adjektiven sozusagen gewandert ist, denn der Versprecher heißt nicht: »Der dicksten Bauer hat die dümmste Kartoffeln.« Stattdessen verharren Singular- und Pluralinformation in der syntaktischen Struktur bei ihren entsprechenden Nomen. Das vertauschte Wort wird seiner neuen Umgebung (Singular- versus Pluralnomen) grammatisch angepasst (d. h. es akkommodiert). Auch diese Beobachtung spricht für die Annahme, dass während der Vertauschung syntaktische Information verarbeitet wurde.

Bock (1996) erläutert die Logik von Versprecheranalysen an dem Sprechfehlerbeispiel »He called her yesterday« anstatt »She called him yesterday«. Hierbei handelt es sich nicht um eine Vertauschung von Pronomen, denn der Fehler ist nicht: »Him called she yesterday«. Demzufolge muss der Fehler entstanden sein, bevor den Pronomen ihre entsprechenden Kasus zugewiesen und die entsprechenden Wortformen der Pronomen festgelegt wurden. Da dies bei Pronomenvertauschungen generell der Fall ist, ist anzunehmen, dass es im Produktionsprozess einen Zeitpunkt gibt, an dem bereits Subjekte und Objekte von Sätzen feststehen, bevor die entsprechenden Wörter, die sie ausdrücken sollen, überhaupt abgerufen sind. Diese Merkmale an der sprachlichen Oberfläche erlauben, sofern sie immer oder so gut wie immer auftreten, derartige Rückschlüsse auf den Zeitpunkt der Fehlerentstehung sowie auf die Informationsart, die vor und nach der Fehlerentstehung verarbeitet wurde. Dass Rückschlüsse dieser Art nicht immer unproblematisch sind, wird in den folgenden Ausführungen deutlich.

Effekte bei der Fehlerentdeckung und -sammlung

Sprechfehlersammler sind nicht perfekt. Vielmehr kommt es beim Sammeln von Versprechern unausweichlich zu einer Konfundierung von Fehlern und der Wahrnehmung von Fehlern (Cutler, 1982). Einige der Effekte, die bei der Fehlerentdeckung auftreten, können aus Sprachwahrnehmungs-

experimenten abgeleitet werden. So beobachtete etwa Warren (1970), dass Versuchspersonen das Fehlen eines Phonems in der Wortmitte überhörten, wenn dieses Wort im Satzkontext präsentiert und mit einem Störgeräusch überlagert wurde. Bei den entsprechenden experimentellen Items war ein Phonem in der Wortmitte gelöscht und mit einem Hustenton überspielt worden. In dem Wort *legislatures* zum Beispiel, das in einen Satzkontext eingebettet war, fehlte das erste /s/: legi*latures. Obwohl es ihnen physikalisch gar nicht vorgespielt worden war, gaben alle Versuchspersonen an, das /s/ gehört zu haben. Dieser Effekt (als phonemischer Restaurationseffekt bezeichnet) legt nahe, dass wir Fehler in der Wortmitte wahrscheinlich häufig überhören und sie automatisch sozusagen restaurieren.

Aus Shadowing- und Fehlerentdeckungsexperimenten von Marslen-Wilson und Welsh (1978) lässt sich ableiten, dass die Silbenposition im Wort ein wichtiger Faktor bei der Fehlerentdeckung und Korrektur ist. In Shadowing-Experimenten sprechen Versuchspersonen einen akustisch präsentierten Text fortlaufend so schnell wie möglich nach. Die Texte in den Experimenten enthielten Wörter, die durch Phonemfehler verzerrt waren, wie *travedy* oder *trachedy* anstatt *tragedy*. Diese Fehler wurden von 49 Prozent der Versuchspersonen während des Nachsprechens korrigiert, ohne dass sich dabei ihr Redefluss verzögerte. Die Korrekturen waren jedoch abhängig von der Silbenposition, in der sich das fehlerhafte Phonem befand: Sie wurden häufiger in finaler Position als an Wortanfängen vorgenommen.

In den Shadowing-Experimenten von Cohen (1980) hatten die Versuchspersonen, anders als in den Experimenten von Marslen-Wilson und Welsh (1978), die ausdrückliche Instruktion, beim Nachsprechen keine Fehler zu korrigieren. Die Ergebnisse zeigten, dass flüssige Restaurationen von Phonemen häufiger (60 %) vorgenommen wurden als Restaurationen von Wörtern (15 %). Demzufolge werden Phonemfehler seltener entdeckt als Wortfehler. Andererseits haben Nootebooms (1980) Analysen von Meringers (1908) Sammlung ergeben, dass beim spontanen Sprechen Wortfehler nahezu genauso häufig wie phonologische Fehler korrigiert werden. Allerdings muss dabei berücksichtigt werden, dass hier Daten unterschiedlicher Qualität vorliegen. Möglicherweise sind Fehlerentdeckungs- und Korrekturprozesse für Sprecher und für Hörer auch nicht identisch (Cutler, 1982; Fowler, 1982).

Cutler (1982) weist im Zusammenhang mit Fehlerentdeckungsprozessen auf die überragende Rolle von Anfangssegmenten bei der Erkennung eines Wortes hin. Sie macht darauf aufmerksam, dass der Befund, dass Sprechfehlersammlungen sehr viel mehr Beispiele für phonetische Fehler in Initialpositionen als in Finalpositionen von Wörtern enthalten, ein Ar-

tefakt sein kann, wenn man bedenkt, dass Fehler an Wortanfängen offensichtlich häufiger wahrgenommen werden als in der Mitte oder am Wortende.

Dell und Reich (1981) haben deshalb ihre Analysen auf genau diese Fehler beschränkt, das heißt, sie haben in ihre Untersuchungen Wortfehler und nur solche Lautfehler einbezogen, bei denen sich die Anfangskonsonanten verändert hatten, um Wahrnehmungseffekten aus dem Weg zu gehen. Um die Reliabilität ihrer Daten zu erhöhen, ließen sie circa 200 Studenten, die keinerlei theoretisches Hintergrundwissen über Sprechfehler besaßen, jeweils über einen Monat Sprechfehler sammeln. Systematische Einflüsse, die Forscher häufig bei der Wahrnehmung von Sprechfehlern leiten, sollten auf diese Weise verringert werden. Um diesen Effekt, der sich auch bei Laien einstellt, weiter zu minimieren, haben Dell und Reich dann bei ihren Analysen nur Fehler berücksichtigt, die besonders leicht zu entdecken sind.

Sammeln online versus offline

Beim Sammeln von Versprechern sind zwei Vorgehensweisen möglich: entweder die des unmittelbaren Mitschreibens oder die Arbeit mit Tonträgern. Beim Vergleich von Fehlern, die online erhoben wurden, und Bandaufzeichnungen zeigen sich zuweilen unterschiedliche Verteilungen (vgl. Bock, 1996). Der Vorteil von Bandaufzeichnungen liegt darin, dass innerhalb von Zeitstichproben alle Fehler registriert und Hörfehler durch mehrmaliges Abspielen minimiert werden können. Darüber hinaus sind die Daten wiederholt überprüfbar und der sprachliche Kontext, in dem der Fehler entsteht, kann ausreichend berücksichtigt werden.

Ferber (1991) hat die Arbeit mit Tonträgern als Offline-Methode mit der Online-Methode des unmittelbaren Mitschreibens bezüglich ihrer Reliabilität verglichen. Die Resultate waren die folgenden: Selbst bei mehrfachem Abhören von Bandaufnahmen durch unterschiedliche Beobachter traten Wahrnehmungsfehler auf, und es wurden unterschiedliche Kategorisierungen vorgenommen. Bei der online-Methode des unmittelbaren Mitschreibens wurde nur 1/3 der tatsächlichen Fehler registriert. Bei wiederholten Offline-Analysen mit Tonträgern wurden aber auch nicht alle Fehler entdeckt, die mit der Online-Methode registriert worden waren.

Vor- und Nachteile beinhalten beide Methoden, wie beispielsweise Aufwand von Technik und Zeit, rechtliche Grundlagen bei der Arbeit mit Tonträgern oder Überprüfbarkeit durch Wiederholbarkeit. In experimentellen Situationen sind heute Bandaufzeichnungen und anschließende Transkrip-

tionen durch mindestens zwei unabhängige Beurteiler üblich. Beim unmittelbaren Mitschreiben spontaner Fehlerdaten nun ist vorteilhaft, dass Sprecher zur Entstehung des Versprechers befragt werden können. Wahrnehmungsbedingte Mängel lassen sich dabei verringern, indem nach dem Vorbild von Meringer und Mayer (1895) nur solche Fehler in die Sammlung aufgenommen werden, die von mehreren gehört und aufgezeichnet wurden.

Effekte linguistischer Verteilungen

Nicht nur wahrnehmungsbedingte Verteilungen beeinflussen die Datenstruktur spontaner Versprecher, sondern auch linguistische Verteilungsmuster. Im Deutschen sind beispielsweise Nomen die größte und gleichzeitig am häufigsten benutzte Wortklasse. So sollte sich ein Sammler deutscher Sprechfehler nicht wundern, wenn Nomenfehler in seiner Sammlung am häufigsten auftreten.

Heute kann zwar in vielen Fällen die Verteilung linguistischer Einheiten aus aktuellen Datenbanken abgefragt werden, ebenso oft sind allerdings die Basiswahrscheinlichkeiten für das Auftreten bestimmter Phänomene beim Sprechen noch unbekannt, sodass aufgrund fehlender Daten keine Basiswahrscheinlichkeiten berechnet werden können. Für Stemberger (1992) ist das Fehlen von Basiswahrscheinlichkeiten für das Auftreten verschiedener Fehlerarten neben den wahrnehmungsbedingten Effekten eines der Hauptprobleme von Versprecherdaten. Er ist dennoch überzeugt, dass natürliche Sprechfehlerdaten reliabel und replizierbar sind, und dass die Reliabilität natürlicher Sprechfehlerdaten empirisch überprüft werden kann. Bei einem Vergleich natürlicher und experimenteller Sprechfehlerstudien kommt er zu dem Ergebnis, dass sie in den meisten Fällen dieselben Ergebnisse aufweisen, oder wenn Unterschiede vorhanden sind, diese auf Aufgabendifferenzen zurückgeführt werden können. Er empfiehlt, natürliche mit experimentellen Sprechfehlerdaten zu kombinieren. Natürliche Versprecherdaten können als reliabel und brauchbar zur Konstruktion von Sprachproduktionsmodellen gesehen werden, wenn sie sich unter kontrollierten experimentellen Bedingungen als replizierbar erweisen. Eine Kombination natürlicher mit experimentell erzeugten Fehlerdaten empfiehlt auch Cutler (1982, 1988). Dabei favorisiert sie die Arbeit mit Tonträgern, um Wahrnehmungsfehler bei der Aufnahme spontaner Fehlerdaten zu minimieren. Die Auswertung der Verteilungsmuster von spontanen Sprechfehlern schließlich erfordere die Berücksichtigung wahrnehmungsbedingter Verteilungen.

Erhobene Daten seien als umso zuverlässiger zu bewerten, je eher ihre Verteilung der wahrnehmungsbedingten Verteilung zuwiderlaufe.

Versprecherforscher versuchen auf zweierlei Weise Zufallswahrscheinlichkeiten für einige Fehlerarten zu berechnen, um statistische Verteilungen in ihren Korpora zu kontrollieren. Eine Methode leitet die Zufallswahrscheinlichkeit eines Fehlers von den Wörtern im Fehlerkorpus selbst ab (Dell u. Reich, 1981), eine andere basiert auf Schätzungen der Frequenz des Auftretens bestimmter Merkmale in der Sprache (z. B. Garrett, 1976).

Kategorisierungsprobleme

Leider ist auch nicht jeder Sprechfehler perfekt (Cutler, 1988). Er ist nicht immer eindeutig bezüglich seiner Kategorie, seiner linguistischen Ebene oder bezüglich seiner Fehlerquelle zu klassifizieren. Betrachten wir folgendes Beispiel. Bei dem Versprecher »Dann findet die Feier natürlich in der *Falle* des Flughafens statt« (anstatt Halle) kann nicht eindeutig entschieden werden, ob die Fehlereinheit ein Phonem (das »F« in Falle) oder ein Wort (Falle) ist. Diese Frage ist aber entscheidend, denn Phoneme und Wörter werden auf unterschiedlichen Ebenen verarbeitet. Von daher hat die Entscheidung, ob es sich um einen Phonem- oder Wortfehler handelt, Konsequenzen bezüglich weiterer Schlussfolgerungen. Diese Ambiguität trifft für alle Substitutionsfehler zu, bei denen sich intendiertes und fehlerhaftes Nomen nur in einem Phonem unterscheiden, wie Tasse und Kasse. Auch Stemberger (1992) weist auf die Probleme bei der Kategorisierung von Fehlern hin. Diese waren bereits von Meringer und Mayer (1895) sowie von Wundt (1904) thematisiert worden. Cutler (1988) vertritt im Umgang mit Kategorisierungsproblemen eine klare Haltung: Fehler, die bezüglich ihrer Kategorie, Ebene und Fehlerquelle nicht eindeutig sind, müssen von der Berechnung ausgeschlossen werden.

Ein Beispiel für Validitätsprobleme

Versprecher sind eine Datenquelle, die in der Sprachwissenschaft und Psychologie gleichermaßen benutzt wird, ohne dass sich ihre Erhebungs- oder Kategorisierungsmethoden systematisch voneinander unterscheiden. Ein wesentlicher Unterschied liegt jedoch in der Zielsetzung beider Disziplinen: Während Linguisten versuchen, aus Versprechern linguistische Einheiten sowie Regeln und Strukturmerkmale von Sprache abzuleiten, sind Psycho-

logen vornehmlich an der Architektur des Sprachverarbeitungssystems und an den Prozessverläufen interessiert, in denen linguistische Einheiten verarbeitet werden.

Bei der viel zitierten Auseinandersetzung zwischen dem Linguisten Meringer und dem Psychoanalytiker Freud Anfang des 20. Jahrhunderts ging es aus heutiger Sicht um Validitätsfragen. Meringer und Mayer (1895) sowie Meringer (1908) hatten die ersten systematischen Versprechersammlungen angelegt und sprachwissenschaftlich ausgewertet. Bei der Erhebung ihrer Versprecher waren sie sehr sorgfältig vorgegangen: Jeder Versprecher war von mindestens zwei Hörern wahrgenommen worden. Neben dem sprachlichen Kontext und vielen anderen Informationen hatten sie auch Angaben des Sprechers zu seinen Gedanken und möglichen Ursachen des Versprechers protokolliert. Meringer und Mayers Arbeit entspricht hohen methodischen Standards und hat bis heute eine Vielzahl von Folgeuntersuchungen motiviert. Bei ihren Analysen haben die beiden Forscher bereits mehr als die Hälfte der Gesetzmäßigkeiten entdeckt, die heute in der Versprecherforschung als gesichert gelten (Cutler u. Fay, 1978), beispielsweise die Aufrechterhaltung der Silbenposition bei Phonemfehlern oder den Wortklasseneffekt bei Wortfehlern.

Während Meringers Interesse ein rein sprachwissenschaftliches war und er vor allem linguistische Merkmale von Sprechfehlern untersuchte, hat Freud Versprecher im Rahmen seines topografischen Modells des psychischen Apparates erklärt. Dabei wählte er selektiv (z. B. 1901) nur wenige Versprecher aus der immensen Sammlung Meringer und Mayers aus und benutzte sie zur Unterstützung seiner eigenen Hypothesen. Nach Freud sind Versprecher, ebenso wie das Vergessen und Vergreifen, Fehlleistungen, denen Interferenzen zwischen verschiedenen Absichten zugrunde liegen. Sprachliche Fehlleistungen werden als Kompromisse zwischen einer zurückgedrängten oder verdrängten Intention auf der einen und einer bewussten Absicht auf der anderen Seite aufgefasst. Weil Versprecher in diesem Sinne Bedeutungsträger sind, findet Freud ihre Ursachen auch nicht in der Struktur der Sprache, sondern vielmehr in sinnhaften seelischen Akten, die sich lediglich der Sprachstruktur bedienen, um an die Oberfläche zu gelangen, um zur Kenntnis genommen zu werden.

Ohne Meringers und Freuds theoretische Positionen ausführlicher darstellen zu wollen (vgl. hierzu Marx, 2000, 2001), wird am folgenden Versprecherbeispiel deutlich werden, dass die Prozesse, die Meringer aus Versprechern ableitet, von denen, auf die Freud schließt, grundverschieden sind.

Bei dem von Meringer und Mayer protokollierten Versprecher »dann

aber sind Thatsachen zum Vorschwein gekommen« (Meringer u. Mayer, 1895, S. 62), hatte der Sprecher bestätigt, dass er zwar an Schweinereien gedacht, diese aber nicht hatte thematisieren wollen. Während Meringer und Mayer die Ursache dieses Versprechers einzig darin sehen, dass die Wörter Schwein und Schein sich ähneln, besteht das im Sinne Freuds Faszinierende darin, dass diese formale Ähnlichkeit einer tief erliegenden oder zurückgedrängten Absicht ermöglicht, in Gestalt eines Phonems an die Oberfläche zu dringen und dabei zum Bedeutungsträger des komplexen Motivs zu werden. Nach Freuds Auffassung wird dabei die Ähnlichkeit von Wortvorstellungen zum Anlass genommen, um eine Kompromissvorstellung zu schaffen. So gesehen steht die Verwendung eines strukturell ähnlichen Worts unter der Kontrolle des zurückgedrängten Motivs.

Meringer hat in der Folgezeit Freuds Methode, seine Versprecher zu analysieren, zurückgewiesen und ihre Validität in Frage gestellt:

> »Wenn ein Versprechen etwas schafft, was an Obszönes anklingt, so ist nach Freud ein entsprechender **Trieb** daran schuld. Die Fehler, die das beweisen sollen, stammen aus V. u. V. (Versprechen und Verlesen, Anmerkung der Autorin), S. 21 *Eischeißweibchen* für *Eiweißscheibchen* (eine Vertauschung), S. 37 *Lokuskapitäl* für *Lotuskapitäl* (Anticipation), S. 96 *a Popo, Fritz* für *á propos Fritz!* (R-Unterdrückung). Mit welchem Recht sucht hier Freud etwas Besonderes, wenn in unzähligen gleichartigen Fällen nichts Anstößiges zum Vorschein kommt?« (Meringer, 1923, S. 132 f.).

Nur wenige Sprechfehler, die nach Meringer denselben Regeln folgen, wie Tausende von anderen Fehlern seiner Sammlung, benutze Freud für spekulative Einfälle: »Natürlich kommt (Freud S. 83) wieder eine saubere Erklärung aus dem Sexualleben zum Vorschwein« (a. a. O., S. 132).

Nach Cutler (1982) ist der Disput der beiden Forscher ein Beispiel für das Vermengen von Ursachen und Mechanismen, die sich in Sprechfehlern widerspiegeln. Beide machten nach ihrer Auffassung den Fehler anzunehmen, dass ihre jeweiligen Erklärungen auf derselben Ebene liegen. Freuds Erklärungen beziehen sich auf psychische Ursachen von Versprechern. Diese können laut Cutler bei der Erforschung von Mechanismen innerhalb des Formulierungssystems, wie sie von Meringer durchgeführt wurde, ignoriert werden.

Experimentelles Induzieren von Sprechfehlern

Einen Ausweg aus dem Zufalls- und Konfundierungsdilemma natürlicher Sprechfehler bietet das experimentelle Induzieren von Versprechern. Im Experiment kann die Zahl der tatsächlichen Fehler mit der Zahl der möglichen Fehler unter variierenden Bedingungen verglichen werden.

Vor fast dreißig Jahren haben Baars et al. (u. a. 1975) damit begonnen, Techniken zu entwickeln, um Gesetzmäßigkeiten bei der Produktion von Sprechfehlern unter kontrollierten experimentellen Bedingungen zu studieren. Diese Techniken basieren auf der Annahme, dass beim Sprechen häufig mehrere alternative Pläne entwickelt werden, von denen bei störungsfreiem Verlauf schließlich einer favorisiert und ausgeführt wird. Manchmal kommt es zu Verschmelzungen von zwei interferierenden Produktionsplänen, die dann in der sprachlichen Oberflächenstruktur als Sprechfehler erscheinen.

Angesichts der Kreativität unserer Sprache, die uns ermöglicht, einen und denselben Gedanken auf vielfältige Weise in eine sprachliche Form zu bringen, ist nach Baars (1980) anzunehmen, dass innerhalb des Sprachproduktionssystems Entscheidungen über verschiedene linguistische Pläne getroffen werden, die allesamt die intendierte Bedeutung ausdrücken könnten. Als Ergebnis des Entscheidungsprozesses wird schließlich einer der beiden Pläne präferiert und realisiert. Offenbar erzeugt das Sprachproduktionssystem häufig alternative Pläne, die sich bedeutungsmäßig oder stilistisch lediglich in Nuancen unterscheiden.

Auf der Basis dieser Annahme haben Baars und Motley (1974) sowie Motley und Baars (1975) und MacKay (1971) mehr als ein Dutzend Techniken entwickelt, um Sprechfehler experimentell zu erzeugen (vgl. dazu z. B. Baars, 1992). Für den Erfolg der Methode sind zwei Bedingungen notwendig. Erstens muss die Versuchsperson zwei alternative Produktionspläne entwickeln und zweitens muss es ihr an Zeit fehlen, einen der beiden Pläne zu verwerfen (Baars u. MacKay, 1978). Das gleichzeitige Entwickeln alternativer Produktionspläne führt nach Auffassung der Autoren (Baars, 1992) zumeist zu einer kurzfristigen Überlastung des Systems, sodass für Kontrollprozesse weniger Kapazitäten vorhanden sind und Fehler umso wahrscheinlicher auftreten.

Zu den wichtigsten Methoden der drei Forscher, von denen hier nur eine beschrieben wird, gehören antizipatorische Bias-Techniken (anticipatory bias techniques) sowie Reihenfolge-Konflikt-Techniken (order-competition techniques), mit deren Hilfe Fehler auf der Phonem-, Wort- und Phrasenebene erzeugt werden können (Motley et al., 1983). Im deutschsprachigen Raum wurden erstmals in den neunziger Jahren Sprechfehlerexperi-

mente durchgeführt, um extralinguistischer Einflüsse auf die Entstehung von Versprechern zu untersuchen (Bröder u. Bredenkamp, 1996; Schüttauf et al., 1997; Bredenkamp u. Dilger, 1998).

Die Phonemprimingmethode (phonemic biasing method) zur Produktion von Phonemvertauschungen (Baars u. Motley, 1974) oder lexikalischen Abruffehlern (Kimble u. Perlmutter, 1970; Reason, 1992) ist die wahrscheinlich bekannteste Methode zur Erzeugung von Versprechern (Baars, 1992), bekannt geworden unter dem Begriff Slip-Technik (»Spoonerisms of Laboratory-Induced Predisposition«). Mit ihr gelang es den Forschern, bei Versuchspersonen, denen Wortpaare präsentiert wurden, Erwartungen bezüglich der Anfangskonsonanten der Wörter zu erzeugen und zu manipulieren. Dabei wird die antizipatorische Bereitschaft, bestimmte Wortpaare zu produzieren, systematisch variiert. Das geschieht folgendermaßen: Den Versuchspersonen werden tachistoskopisch oder auf dem Bildschirm eines Computers nacheinander für circa eine Sekunde Wortpaare wie *give book, go back, get boot* präsentiert. Diese Wortpaare (Induktoren) müssen fortlaufend still gelesen und kurz abrufbereit im Gedächtnis behalten werden. Bei den Versuchspersonen wird so systematisch die Erwartung aufgebaut, dass auch das nächste Wortpaar das Anfangskonsonantenmuster /g/. . ./b/. . . aufweist und eine Voraktivierung genau dieses Musters auf der phonologischen Ebene bewirkt (Priming). Wenn jedoch ein kritisches Wortpaar mit umgekehrten Anfangskonsonanten, wie *bad* goof präsentiert und unmittelbar danach eine Aufforderung zum lauten Wiederholen gegeben wird, kann in 10–30 Prozent der Fälle beobachtet werden, dass die Versuchspersonen nicht *bad* goof (kritisches Wortpaar), sondern *gad* boof (Zielwortpaar) produzieren. Wichtig bei dieser Technik ist, dass ein Sprachproduktionsprozess und nicht ein Leseprozess gemessen wird. Daher dürfen in der Wiederholungsphase keine Items auf dem Bildschirm sichtbar sein. Dennoch lässt sich der Einwand, dass es bei den Experimenten zu einer Konfundierung von Verstehens- und Produktionsprozessen kommen kann, nicht ganz von der Hand weisen.

Die experimentell erzeugten Phonemkollisionen resultieren jedoch nicht nur in kompletten Vertauschungen der Anfangsphoneme (Spoonerismen), sondern auch in Antizipationen (*gad goof*) oder in Perseverationen (*bad boof*). Wichtig für das Gelingen der Induktionstechnik ist, dass die Versuchspersonen jederzeit darauf gefasst sind, das zuletzt wahrgenommene Wortpaar aus dem Gedächtnis abrufen und produzieren zu müssen, ohne vorhersagen zu können, wann die Aufforderung dazu erfolgt.

Dass der Kontext bei der Art und Auftretenshäufigkeit von Fehlern eine nicht unwesentliche Rolle spielt, zeigt eine berühmt gewordene experimen-

telle Untersuchung von Motley (1980), von der hier nur zwei Teilergebnisse berichtet werden; sie dokumentiert das Zusammenwirken von situativem Kontext und Phonempriming eindrucksvoll. Die in den Experimenten entstehenden Fehler spiegeln die experimentellen Bedingungen direkt wider. Versuchspersonen, die aufgrund von Instruktion und Experimentalbedingung damit rechnen, einen leichten Elektroschock zu bekommen, produzieren mehr Versprecher, die sich auf die Schockbedingung beziehen, wie etwa *bad shock* anstatt *shad bock*, als Versuchspersonen einer Kontrollgruppe. Männliche Versuchspersonen, die unter einer »Sexualitätsbedingung« untersucht wurden, das heißt von einer attraktiven, provozierend gekleideten Versuchsleiterin, produzierten mehr sexualitätsbezogene Versprecher, wie *foxi girl* anstatt *goxi firl*. Beide Fehlerarten wurden in neutralen Versuchsbedingungen seltener produziert.

Baars et al. (1975) haben aus ihren Beobachtungen geschlossen, dass der Artikulation einer Äußerung eine prä-artikulatorische Kontrollkomponente (Editor) vorgeschaltet ist, welche linguistische und extralinguistische Faktoren in die Überprüfung einbezieht. In der betreffenden Untersuchung bestätigte sich die Hypothese, dass sich die Häufigkeit von Phonemvertauschungen als eine Funktion des lexikalischen Status des entstehenden Fehlers verändert (lexical-bias). Die Wahrscheinlichkeit für eine (experimentell induzierte) Phonemvertauschung ist höher, wenn sie zu einem Wort führt, als wenn sie nicht zu einem Wort führen würde (Lexical-bias-Effekt). Beispielsweise führt das Wortpaar *darn bore* häufiger zu *barn door* (= lexikalisches Item) als *dart board* zu *bart doard* (= Nichtwort). Aber auch dieser Effekt ist von der Versuchsbedingung abhängig, denn wenn nur Nichtwörter im Experiment präsentiert werden, tritt der Lexical-bias-Effekt seltener auf. Sobald echte Wörter präsentiert werden, erhöht sich der Lexical-bias-Effekt. Die drei Forscher haben die Ergebnisse dahingehend interpretiert, dass Sprecher ihren Sprachfluss prä-artikulatorisch unter anderem bezüglich des lexikalischen Status der zu artikulierenden Phonemstränge überwachen. Dabei werden Nichtwörter häufiger entdeckt und intern korrigiert (vgl. dagegen Dell u. Reich, 1981).

Seit 1971 haben Tausende von Versuchspersonen an den Sprechfehlerexperimenten der drei Forscher teilgenommen. Als wesentliches Ergebnis kann festgehalten werden, dass sich die experimentellen Sprechfehlermerkmale mit den Merkmalen natürlicher Daten decken. Vorteilhaft beim Experimentieren ist, dass hier immer Tonträger eingesetzt werden können und die Daten dadurch überprüfbar sind. Die im Experiment beobachtbaren Fehler folgen denselben Prinzipien wie natürlich beobachtbare (Baars, 1992; Stemberger, 1992). Im Alltag wie im Experiment sind Betrof-

fene sehr überrascht über ihre Versprecher. Fordert man aber jemanden auf, aus einem Wortpaar ganz bewusst einen Versprecher zu produzieren, produziert dieser häufig künstlich einen Versprecher, der etwa gegen phonologische Regeln verstößt und im Alltag sowie im Experiment als unbeabsichtigter Fehler nie produziert würde.

Fromkin (1980) hatte sich gegenüber Sprechfehlerexperimenten skeptisch geäußert und, vor allem im Hinblick auf den oben beschriebenen Lexical-bias-Effekt, den ihre natürlichen Fehlerdaten nach ihrer Auffassung nicht aufwiesen, deren ökologische Validität angezweifelt. Diesem Einwand begegnet Fowler (1982) mit methodisch-statistischen Argumenten, die sich gegen das Fehlen einer Nullhypothese in Fromkins Daten richten: »In fact, however, this is an example where the experimentalists and statisticians can educate the arm-chair collectors on evaluation of data. A proportion of nonword errors, like 60 %, cannot be evaluated unless it is compared to a chance value ...« (Fowler, 1982, S. 278).

Auch die Tatsache, dass die Versprecherinduktionstechnik sehr effektiv ist, spricht dafür, dass es den Forschern gelungen ist, genau die Eigenschaften des Produktionssystems anzusprechen, die an der Verursachung von Fehlern beteiligt sind. Darüber hinaus sind die Induktionstechniken in den letzten 20 Jahren auch im Zuge technischer Weiterentwicklungen kontinuierlich verbessert und aufgrund theoretischer Überlegungen sowie experimenteller Erfahrungen verändert und durch neue Methoden ersetzt worden. Ein Beispiel dafür ist etwa die von Baars (u. a. 1980) vorgestellte Phrasenumstellungstechnik zur Erzeugung von Wortvertauschungen, die in jüngster Zeit durch die Frage-Antwort-Technik ersetzt wurde, da diese nach Auffassung der Forscher natürliches Sprechen noch besser zu simulieren vermag (Baars, 1992).

Bei allen methodisch-statistischen Vorteilen einer experimentellen Situation, in der die an der Entstehung von Sprechfehlern beteiligten Faktoren systematisch variiert, das heißt kontrolliert werden können, darf nicht vergessen werden, dass sich auch hier die Validitätsfrage stellt. Schriefers (2003) weist auf folgendes Problem hin. Wenn aus einem Versprecher auf zugrunde liegende Verarbeitungsprozesse geschlossen wird, bleibt eine zentrale Frage dabei unbeantwortet: Spiegeln die Spuren, die der Versprecher hinterlässt, normale Produktionsverläufe wider oder sind sie lediglich charakteristisch für den gestörten Verlauf, der den Versprecher verursacht hat? Speziell bei der oben beschriebenen Slip-Technik werden Items produziert, nachdem unmittelbar zuvor Wahrnehmungsprozesse stattgefunden haben, die auf Konfusionen phonemischer Sequenzierungsprozesse abzielen. Ob und wie dieser Unterschied zwischen experimenteller und natürli-

cher Sprachproduktion die experimentellen Daten beeinflusst, ist nicht bekannt. Die derzeit vorliegenden Ergebnisse jedenfalls decken sich mit den in natürlichen Situationen erhobenen Daten.

Resümee

In den vorangegangenen Ausführungen wurden Vor- und Nachteile der Gewinnung natürlicher und experimenteller Sprechfehlerdaten diskutiert. Dabei hat sich herausgestellt, dass bei der Arbeit mit natürlichen Sprechfehlerdaten, auch unter Zuhilfenahme von Tonträgern, eine Konfundierung von Fehlern und der Wahrnehmung von Fehlern nicht ausgeschlossen werden kann.

Die Gesetzmäßigkeiten, die bei der Fehlerentdeckung zum Tragen kommen, können aus Sprachwahrnehmungsexperimenten abgeleitet werden. Bei der Erhebung und Auswertung der Daten werden reliable und valide Ergebnisse erzielt, sofern wahrnehmungsbedingte und linguistische Verteilungen Berücksichtigung finden, ambige Fehler von den Berechnungen ausgeschlossen werden und vor allem statistische Hypothesen formuliert werden können. Ob Daten reliabel und replizierbar sind, kann empirisch überprüft werden durch Testen abgeleiteter Hypothesen unter kontrollierten experimentellen Bedingungen.

Experimentelles Induzieren von Sprechfehlern über einen Zeitraum von circa 30 Jahren hat ergeben, dass natürliche und experimentelle Daten dieselben Gesetzmäßigkeiten aufweisen. Dennoch ist bei der Datenerhebung eine Kombination beider Methoden empfehlenswert, da jede mit gewissen Mängeln behaftet ist und sich naturgemäß experimentelle von natürlichen Bedingungen unterscheiden. So besteht der Sinn von Experimenten gerade darin, die Häufigkeit von Fehlern, die im Alltag zu selten auftreten, um an ihnen systematisch Gesetzmäßigkeiten untersuchen zu können, gezielt zu erhöhen. Nicht nur eine Kombination von Erhebungsmethoden, sondern auch das Analysieren desselben Datenmaterials mit mehreren unabhängigen Untersuchungsverfahren kann die Validität der Messung erhöhen. Ein Beispiel für ein solches Vorgehen ist ein frühes Experiment von Cohen (1966), der das Stimulusmaterial in seinen Versuchen auf der Basis von ihm gesammelter natürlicher Sprechfehler erstellte und die in den Experimenten entstandenen Lesefehler mit den Ergebnissen der natürlichen Sprechfehlerdaten verglich. Auch wenn es sich hierbei um das Induzieren von Le-

sefehlern handelt, kann das Experiment als ein direkter Vorläufer der Sprechfehlerinduktionsexperimente gesehen werden.

Ein anderes Beispiel für das gleichzeitige Einsetzen unabhängiger Methoden ist eine Untersuchung von MacKay (1980), der aus seiner Sammlung natürlicher Sprechfehler in Fehlern interagierende Wörter von unabhängigen Ratern bezüglich ihrer semantischen Ähnlichkeit einschätzen ließ. In Übereinstimmung mit seinen Beobachtungen auf der Basis der natürlichen Fehlerdaten hatte MacKay auf diese Weise einen unabhängigen Nachweis für seine Annahme erbracht, dass der Entstehung von Blends und semantischen Wortsubstitutionen offenbar unterschiedliche Prozesse zugrunde liegen.

Zusammenfassend legen die Ausführungen der vorangegangenen Kapitel nahe, natürliche, experimentelle und aus Simulationen gewonnene Daten zu kombinieren, um die Stärken jeder Methode zu nutzen und Schwächen auszugleichen (Fowler, 1982). So hat etwa Dell (1986) auf der Basis natürlicher Fehlerdaten Hypothesen abgeleitet, ein Sprachproduktionsmodell entworfen, seine Hypothesen in Experimenten getestet und schließlich mithilfe experimenteller und natürlicher Fehlerdaten Simulationen entwickelt.

Festzuhalten bleibt noch, dass die Aussagekraft von Sprechfehlerdaten natürlich begrenzt ist. Sie liefern häufig die Basis für weitere Experimente. Aus der Untersuchung von Sprechfehlermerkmalen entstehen Hypothesen und Modelle. Sie erlauben Rückschlüsse auf Planungsvorgänge und vermögen zeitliche Strukturen nachzuzeichnen (vgl. Meyer, 1992). Das Testen von Zeitverläufen innerhalb der Mikroprozesse, die häufig wenige Millisekunden beinhalten, kann jedoch im zweiten Schritt nur in chronometrischen Studien erfolgen.

Literatur

Baars, B.J. (1980). On eliciting predictable speech errors in the laboratory. In V.A. Fromkin (Hg.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (S. 307–318). New York: Academic Press.

Baars, B.J. (1992). A dozen competing-plans techniques for inducing predictable slips in speech and action. In B.J. Baars (Hg.), *Experimental slips and human error: Exploring the architecture of volition* (S. 129–150). New York: Plenum.

Baars, B.J., Motley, M.T. (1974). Spoonerisms: Experimental elicitation of human speech errors. *Catalog of Selected Documents in Psychology, 3,* 28–47.

Baars, B.J., Motley, M.T., MacKay. D. (1975). Output editing for lexical status from artificial elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 14,* 382–391.

Baars, B.J., MacKay, D. G. (1978). Experimentally eliciting phonetic and sentential speech errors: methods, implications, and work in progress. *Language in Society, 7,* 105–109.

Berg, Th. (1992). Prelexical and postlexical features in language production. *Applied Psycholinguistics, 13,* 199–235.

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review, 3,* 395–421.

Bredenkamp, J., Dilger, St. (1998). Erkundungsexperimente über den Zusammenhang von Versprechern und Arbeitsgedächtnis. *Zeitschrift für Experimentelle Psychologie, 45,* 72–79.

Bröder, A., Bredenkamp, J. (1996). SLIP-Technik, Prozessdissoziationsmodell und multinomiale Modellierung: neue Werkzeuge zum experimentellen Nachweis »Freudscher Versprecher«? *Zeitschrift für Experimentelle Psychologie, 43,* 175–202.

Cohen, A. (1966). Errors of speech and their implication for understanding the strategy of the language users. *Zeitschrift für Phonetik, 21,* 177–181.

Cohen, A. (1980). Correcting of speech errors in a shadowing task. In V.A. Fromkin (Hg.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (S. 157–163). New York: Academic Press.

Cutler, A. (1982). The reliability of speech error data. In A. Cutler (Hg.), *Slips of the tongue and language production* (S. 7–28). Amsterdam: de Gruyter.

Cutler, A. (1988). The perfect speech error. In L.M. Hyman u. Ch.N. Li (Hg.), *Language, speech and mind: studies in honour of Victoria A. Fromkin* (S. 209–223). New York: Routledge.

Cutler, A., Fay, D. (1978) Introduction. In A. Cutler u. D. Fay (Hg.), *Re-issue of R. Meringer & C. Mayer's Versprechen und Verlesen.* Amsterdam: Benjamins.

Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93,* 283–321.

Dell, G.S., Reich, P.A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20,* 611–629.

Ferber, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research, 20,* 105–122.

Fowler, C. (1982). Review of V.A. Fromkin (Hg.), Errors in linguistic performance: Slips of the tongue, ear, pen and hand. In A. Cutler (Hg.), *Slips of the tongue and language production* (S. 265–286). Amsterdam: Mouton Publishers.

Freud, S. (1901/1993). *Zur Psychopathologie des Alltagslebens.* Frankfurt a. M.: Fischer.

Fromkin, V.A. (1980). Introduction. In V.A. Fromkin (Hg.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (S. 1–12). New York: Academic Press.

Garrett, M.F. (1976). Syntactic processes in sentence production. In R.J. Wales u. E. Walker (Hg.), *New approaches to language mechanisms* (S. 231–256). Amsterdam: North Holland.

Garrett, M.F. (1980). Levels of processing in sentence production. In B. Butterworth (Hg.), *Language production* (Bd. 1, S. 177–220). London: Academic Press.

Kimble, G.A., Perlmutter, D. (1970). The problem of volition. *Psychological Review, 77,* 361–384.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Levelt, W.J.M. (1999). Models of word production. *Trends in Cognitive Science, 3,* 223–232.

MacKay, D.G. (1971). Stress pre-entry in motor systems. *American Journal of Psychology, 84,* 21–43.

MacKay, D.H. (1980). Speech errors: Retrospect and prospect. In V.A. Fromkin (Hg.), *Errors in linguistic performance. Slips of the tongue, ear, pen, and hand* (S. 319–332). New York: Academic Press.

Marslen-Wilson, W.D., Welsh, A. (1978). Processing Interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

Marx, E. (2000). *Versprecher und Genusverarbeitung. Analysen natürlicher und experimentell erzeugter Sprechfehler.* Wiesbaden: Deutscher Universitätsverlag.

Marx, E. (2001). Gewisst wu – gewusst wie! Was die Versprecherforschung über Sprachproduktion weiß. *Psychologische Rundschau, 52,* 195–204.

Meringer, R. (1908). *Aus dem Leben der Sprache. Versprechen. Kindersprache. Nachahmungstrieb.* Berlin: B. Behr's Verlag.

Meringer, R. (1923). Die täglichen Fehler im Sprechen, Lesen und Handeln (zu Freuds Psychopathologie des Alltagslebens). *Wörter und Sachen 8,* 122–141.

Meringer, R., Mayer, C. (1895). *Versprechen und Verlesen. Eine psychologisch-linguistische Studie.* Stuttgart: Göschen'sche Verlagshandlung.

Meyer, A.S. (1992). Investigation of phonological encoding through speech error analysis: Achievements, limitations, and alternatives. *Cognition, 42,* 181–212.

Motley, M.T. (1980). Verification of »Freudian slips« and semantic prearticulatory editing via laboratory-induced spoonerisms. In V.A. Fromkin (Hg.), *Errors in linguistic performance. Slips of the tongue, ear, pen, and hand* (S. 133–147). New York: Academic Press.

Motley, M.T., Baars, B.J. (1975). Encoding sensitivities to phonological markedness and traditional probability: Evidence from spoonerisms. *Human Communication Research, 2,* 353–361.

Motley, M.T., Baars, B.J., Camden, C.T. (1983). Experimental verbal slip studies: A review and an editing model of language encoding. *Communication Monographs, 50,* June 1983.

Nooteboom, S.G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V.A. Fromkin (Hg.), *Errors in linguistic performance. Slips of the tongue, ear, pen, and hand* (S. 87–95). New York: Academic Press.

Pechmann, T. (1994). *Sprachproduktion. Zur Generierung komplexer Nominalphrasen.* Opladen: Westdeutscher Verlag.

Reason, J.T. (1992). Cognitive underspecification: It's variety and consequences. In B.J. Baars (Hg.), *Experimental slips and human error: Exploring the architecture of volition* (S. 72–92). New York: Plenum.

Schriefers, H. (2003). Methodologische Probleme. In Th. Herrmann u. J. Grabowski (Hg.), *Psychologie der Sprachproduktion. Enzyklopädie der Psychologie, Themenbereich C: Theorie und Forschung, Serie III: Sprache* (Bd. 1, S. 3–26). Göttingen: Hogrefe.

Schüttauf, K., Bredenkamp, J., Specht, K.E. (1997). Induzierte »Freudsche Versprecher« und zwangsneurotischer Konflikt. *Sprache und Kognition, 16,* 3–13.

Stemberger, J.P. (1992). The reliability and replicability of naturalistic speech error data. A comparison with experimentally induced errors. In B.J. Baars (Hg.), *Experimental slips and human error: Exploring the architecture of volition* (S. 195–215). New York: Plenum.

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science, 167,* 392–393.

Wundt, W. (1904). *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythos und Sitte. Erster Band. Die Sprache* (2. Aufl.; 1. Aufl. 1900). Leipzig: Verlag Wilhelm Engelmann.

Teil IV:
Über Jürgen Bredenkamp

Über Jürgen Bredenkamp

Kurzbiografie von Jürgen Bredenkamp

1939	geboren als Sohn der Eheleute Hans Bredenkamp und Anneliese geb. Behrmann in Hamburg
1945–1947	Volksschule Eldingen Kreis Celle
1947–1949	Volksschule Sprötze Kreis Harburg
1949–1953	Mittelschule Buchholz Kreis Harburg
1953–1958	Förderzweig des Gymnasiums für Jungen St. Georg, Hamburg
1958	Abitur, Studium der Jurisprudenz an der Universität Hamburg
1958–1963	Studium der Psychologie in Hamburg bei C. Bondy, P. R. Hofstätter und G. A. Lienert
1963	Dipl.-Psych.
1964	Promotion zum Dr. phil. in Hamburg (Doktorvater: G. A. Lienert)
1964–1965	Verwalter einer wissenschaftlichen Assistentenstelle bei C. F. Graumann, Universität Heidelberg
1965–1971	wissenschaftlicher Assistent am Psychologischen Institut der Universität Heidelberg bei C. F. Graumann
1968	Eheschließung mit Karin geb. Spies
1971	Habilitation im Fach Psychologie an der Universität Heidelberg
1971- 1972	Universitätsdozent in Heidelberg
1972	H3-Professor für psychologische Methodenlehre an der Universität Bonn. Geburt der Tochter Silke; Ruf auf einen Lehrstuhl für Psychologie an der Universität Göttingen und Annahme des Rufes
1972–1980	Universität Göttingen
1975	Geburt der Tochter Birthe; Ruf auf einen Lehrstuhl für Psychologie in Bochum (abgelehnt)
1980	Ruf auf einen Lehrstuhl für Psychologie an der Universität Trier (angenommen)
1980–1984	Universität Trier

1984	Ruf auf einen Lehrstuhl für Psychologie an der Universität Bonn (angenommen)
Seit 1984	Universität Bonn
1984–1992	Mitglied des Fachausschusses Psychologie der DFG
1990–1992	Präsident der Deutschen Gesellschaft für Psychologie

Veröffentlichungen von Jürgen Bredenkamp
(Chronologisch geordnet)

Bredenkamp, J. (1966). Eine Analyse der Flimmerverschmelzungsfrequenz als Ermüdungsindikator. *Zeitschrift für Experimentelle und Angewandte Psychologie, 13,* 199–221.

Bredenkamp, J. (1968). F-Tests zur Prüfung von Trends und Trendunterschieden. *Zeitschrift für Experimentelle und Angewandte Psychologie, 15,* 239–272.

Wambsganss, E., Bredenkamp, J. (1968). Eine experimentalpsychologische Untersuchung über die Wirkung von Haloperidol in niedriger Dosierung bei alleiniger Applikation und in Verbindung mit Alkohol. *Arzneimittelforschung (Drug Res.), 18,* 238–243.

Bredenkamp, J. (1969). Experiment und Feldexperiment. In C.F. Graumann (Hg.), *Handbuch der Psychologie, Band 7, 1. Halbband: Theorien und Methoden der Sozialpsychologie* (S. 332–374). Göttingen: Hogrefe.

Bredenkamp, J. (1969). Über die Anwendung von Signifikanztests bei Theorie-testenden Experimenten. *Psychologische Beiträge, 11,* 275–285.

Bredenkamp, J. (1969). Über ein Problem bei der Prüfung psychologischer Hypothesen. *Psychologische Beiträge, 11,* 595–598.

Bredenkamp, J. (1970). Über Maße der praktischen Signifikanz. *Zeitschrift für Psychologie, 177,* 310–318.

Bredenkamp, J., Feger, H. (1970). Kriterien für die Entscheidung über die Aufnahme empirischer Arbeiten in die Zeitschrift für Sozialpsychologie. *Zeitschrift für Sozialpsychologie, 1,* 43–47.

Bredenkamp, K., Bredenkamp, J. (1970). Eine Untersuchung zum funktionalen Reiz beim seriellen Lernen. *Archiv für Psychologie, 122,* 249–258.

Bredenkamp, J. (1971). Bemerkungen zu der Trendanalyse nach Ferguson. *Zeitschrift für Experimentelle und Angewandte Psychologie, 18,* 386–391.

Bredenkamp, J. (1971). Zwei Anmerkungen zu Münch und Schmid: Konventionalismus und empirische Forschungspraxis. *Zeitschrift für Sozialpsychologie, 2,* 273–274.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Akademische Verlagsgesellschaft.

Bredenkamp, J., Wördehoff, W. (1972). Zur Invarianzhypothese beim verbalen Lernen. *Archiv für Psychologie, 124,* 110–116.

Bredenkamp, J., Wirsching, W. (1972). Weitere Untersuchungen der Invarianzhypothese. *Archiv für Psychologie, 124,* 117–122.

Bredenkamp, K., Bredenkamp, J. (1972). Lernen und Gedächtnis. In K. E. Rogge (Hg.), *Steckbrief der Psychologie* (S. 76–91). Heidelberg: Quelle & Meyer.

Bredenkamp, J., Graumann, C.F. (1973). Möglichkeiten und Grenzen mathematischen

Verfahrens in den Verhaltenswissenschaften. In H.G. Gadamer u. Vogler, P. (Hg.), *Neue Anthropologie* (S. 51–93). Stuttgart: Thieme. (Brasilianische Ausgabe 1977).

Bredenkamp, J. (1973). Der systematische Kulturvergleich (Rezension zu Eckensberger: Methodenprobleme der kulturvergleichenden Psychologie). *Zeitschrift für Sozialpsychologie, 4,* 153–156.

Bredenkamp, K., Bredenkamp, J. (1974). Was ist Lernen? In F.E. Weinert, C.F. Graumann, H. Heckhausen u. M. Hofer (Hg.), *Funk-Kolleg Pädagogische Psychologie* (Band 2, S. 605–630). Frankfurt a. M.: Fischer.

Bredenkamp, K., Bredenkamp, J. (1974). Die Bedingungen des Erlernens, Behaltens und Vergessens von sprachlichem Material. In F.E. Weinert, C.F. Graumann, H. Heckhausen u. M. Hofer (Hg.), *Funk-Kolleg Pädagogische Psychologie* (Band 2, S. 631–656). Frankfurt a. M.: Fischer.

Bredenkamp, J. (1974). Nonparametrische Prüfung von Wechselwirkungen. *Psychologische Beiträge, 16,* 398–416.

Bredenkamp, J. (1975). Varianzanalytische und regressionsanalytische Verfahren in der Curriculumevaluation. In K. Frey (Hg.), *Curriculum-Handbuch* (Band 2, S. 786–823). München: Piper.

Bredenkamp, J. (1975). Die modifizierte Invarianzhypothese und die Hypothese einer konstanten Langzeit-Gedächtnisspanne. *Psychologische Beiträge, 17,* 483–496.

Bredenkamp, J. (unsigniert) (1976). Lernen. In *Brockhaus* (Ergänzungsband I-Z, S. 156–157). Wiesbaden: Brockhaus.

Bredenkamp, K., Bredenkamp, J. (1976). Was ist Lernen? In F.E. Weinert, C.F. Graumann, H. Heckhausen u. M. Hofer (Hg.), *Pädagogische Psychologie. Teil V: Studienbegleitbriefe zum Funk-Kolleg Pädagogische Psychologie* (S. 1–20). Weinheim: Beltz.

Bredenkamp, K., Bredenkamp, J. (1976). Bedingungen des Erlernens, Behaltens und Vergessens von sprachlichem Material. In F.E. Weinert, C.F. Graumann, H. Heckhausen u. M. Hofer (Hg.), *Pädagogische Psychologie. Teil V: Studienbegleitbriefe zum Funk-Kolleg Pädagogische Psychologie* (S. 21–36). Weinheim: Beltz.

Bredenkamp, J., Wippich, W. (1977). *Lern- und Gedächtnispsychologie* (Band 1 und 2). Stuttgart: Kohlhammer.

Baschek, I.-L., Bredenkamp, J., Oehrle, B., Wippich, W. (1977). Bestimmung der Bildhaftigkeit (I), Konkretheit (C) und der Bedeutungshaltigkeit (m') von 800 Substantiven. *Zeitschrift für Experimentelle und Angewandte Psychologie, 24,* 353–396.

Bredenkamp, J. (1977). Gedächtnis. In Th. Herrmann, P.R. Hofstätter, H.P. Huber u. F.E. Weinert (Hg.), *Handbuch psychologischer Grundbegriffe* (S. 143–155). München: Kösel.

Wippich, W., Bredenkamp, J. (1977). Bestimmung der Bildhaftigkeit (I), Konkretheit (C) und der Bedeutungshaltigkeit (m') von 498 Verben und 400 Adjektiven. *Zeitschrift für Experimentelle und Angewandte Psychologie, 24,* 671–680.

Bredenkamp, J. (1979). Das Problem der externen Validität pädagogisch-psychologischer Untersuchungen. In J. Brandtstätter, G. Reinert u. K.A. Schneewind (Hg.), *Pädagogische Psychologie: Probleme und Perspektiven* (S. 267–290). Stuttgart: Klett-Cotta.

Wippich, W., Bredenkamp, J. (1979). *Bildhaftigkeit und Lernen.* Darmstadt: Steinkopff.

Bredenkamp, J., Hager, W. (1979). Experimentelle Befunde zur modifizierten Invarianzhypothese und zur Hypothese einer konstanten Langzeitgedächtnisspanne. *Psychologische Beiträge, 21,* 382–400.

Bredenkamp, J. (1980). Untersuchungen zum Kontexteinfluß auf das Wahrnehmungsurteil. In W. Lauterbach u. V. Sarris (Hg.), *Beiträge zur psychologischen Bezugssystemforschung* (S. 207–234). Bern: Huber.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1980). Vier Artikel über »Extinktion«, »Gedächtnis«, »Lernen«, »Verstärkung« im Fachlexikon der sozialen Arbeit (Hrsg. vom Deutschen Verein für öffentliche und private Fürsorge anläßlich seines 100jährigen Bestehens). Eigenverlegung des Deutschen Vereins [2002, 5. Auflage].

Bredenkamp, J. (1981). Gedächtnis. In H. Schiefele u. A. Krapp (Hg.), *Handlexikon zur Pädagogischen Psychologie* (S. 129–134). München: Ehrenwirth.

Bredenkamp, J. (1981). Untersuchungen zu einem Verfahren der nonparametrischen Wechselwirkung. In W. Janke (Hg.), *Beiträge zur Methodik in der differentiellen, diagnostischen und klinischen Psychologie. Festschrift zum 60. Geburtstag von G.A. Lienert* (S. 2–10). Königstein/Ts.: Anton Hain.

Bredenkamp, J. (1982). Psychophysikalische Analysen zur Erklärung des Entstehens und Verschwindens von Wahrnehmungstäuschungen. *Trierer Psychologische Berichte, Band 9, Heft 1.*

Bredenkamp, J. (1982). Verfahren zur Ermittlung des Typs einer statistischen Wechselwirkung. *Psychologische Beiträge, 24,* 56–75.

Bredenkamp, J. (1982). Psychophysikalische Analysen von Wahrnehmungstäuschungen. *Psychologische Beiträge, 24,* 322–325.

Bredenkamp, J. (1983). Übersicht. In J. Bredenkamp u. H. Feger (Hg.), *Hypothesenprüfung* (= Band 5 der Serie »Forschungsmethoden der Psychologie« im Rahmen der Enzyklopädie der Psychologie, S. 1–23). Göttingen: Hogrefe.

Bredenkamp, J., Feger, H. (Hg.). (1983). *Hypothesenprüfung* (= Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie I: Forschungsmethoden der Psychologie, Band 5). Göttingen: Hogrefe.

Bredenkamp, J., Feger, H. (Hg.). (1983). *Strukturierung und Reduzierung von Daten* (= Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie I: Forschungsmethoden der Psychologie, Band 4). Göttingen: Hogrefe.

Feger, H., Bredenkamp, J. (Hg.). (1983). *Datenerhebung* (= Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie I: Forschungsmethoden der Psychologie, Band 2). Göttingen: Hogrefe.

Feger, H., Bredenkamp, J. (Hg.). (1983). *Messen und Testen* (= Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie I: Forschungsmethoden der Psychologie, Band 3). Göttingen: Hogrefe.

Bredenkamp, J. (1984). Weitere kritische Anmerkungen zur Anwendung der Informationsintegrationstheorie in der Entwicklungspsychologie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 16,* 134–145.

Bredenkamp, J. (1984). Theoretische und experimentelle Analysen dreier Wahrnehmungstäuschungen. *Zeitschrift für Psychologie, 192,* 47–61.

Bredenkamp, J. (1984). Anmerkungen und Korrekturen zu Hager & Westermann: Entscheidung über statistische und wissenschaftliche Hypothesen: Probleme bei mehrfachen Signifikanztests zur Prüfung einer wissenschaftlichen Hypothese. *Zeitschrift für Sozialpsychologie, 15,* 224–229.

Bredenkamp, J., Gigerenzer, G. (1984). Einführung: Einige Gedanken zur Kontextabhängigkeit der Wahrnehmung und des Urteils. In G. Gigerenzer u. J. Bredenkamp (Hg.), Psychophysik heute: Aktuelle Probleme und Ergebnisse II. *Psychologische Beiträge, 26,* 89–157.

Bredenkamp, J. (1984). Psychophysikalische Analysen von Wahrnehmungstäuschungen: Theorie und weitere Befunde. In G. Gigerenzer u. J. Bredenkamp (Hg.), Psychophysik heute: Aktuelle Probleme und Ergebnisse II. *Psychologische Beiträge, 26,* 102–105.

Bredenkamp, J. (1984). Ausbildung in den Grundlagen. In W. Michaelis u. E. Stephan (Hg.), *Ausbildungsreform Psychologie. Argumente – Standpunkte – Forderungen* (S. 96–106). Göttingen: Hogrefe.

Bredenkamp, J. (1984). Vertraute und angenommene Größe als Signal der Entfernungs- und Tiefenwahrnehmung. *Archiv für Psychologie, 136,* 89–105.

Erdfelder, E., Bredenkamp, J. (1984). Kritik mehrfaktorieller Rangvarianzanalysen. *Psychologische Beiträge, 26,* 263–282.

Bredenkamp, J. (1984). Theoretische und experimentelle Analysen einiger Wahrnehmungstäuschungen. *Archiv für Psychologie, 136,* 281–291.

Sarris, V., Bredenkamp, J. (1984). Aufgaben, Ergebnisse und Probleme einer Psychophysik heute. In J. Bredenkamp u. V. Sarris (Hg.), Psychophysik heute: Aktuelle Probleme und Ergebnisse III. *Psychologische Beiträge, 26,* 616–671.

Bredenkamp, J., Erdfelder, E. (1985). Multivariate Varianzanalyse nach dem V-Kriterium. *Psychologische Beiträge, 27,* 127–154.

Bredenkamp, J. (1986). Untersuchungen zur Tauglichkeit der Antwortzeit als Basis für die Unterscheidung verschiedener Begriffsarten. *Zeitschrift für Psychologie, 194,* 177–189.

Bredenkamp, J. (1986). Dürfen wir psychologische Hypothesen statistisch prüfen? *Berichte aus dem Bonner Psychologischen Institut, Band 12, Heft 2.*

Bredenkamp, J., Wippich, W. (1986). Anmerkungen zu W. Hager: Methodische und empirische Analysen zur Interaktion von semantischen Eigenschaften mit der Lernart und der Verarbeitungstiefe. *Zeitschrift für Experimentelle und Angewandte Psychologie, 33,* 542–544.

Bredenkamp, J. (1986). Die Unterscheidung verschiedener Begriffsarten unter besonderer Berücksichtigung einer dreiwertigen Logik. *Sprache & Kognition, 5,* 155–162.

Bredenkamp, J. (1986). Untersuchungen zur Unterscheidung verschiedener Begriffsarten. In M. Amelang (Hg.), *Bericht über den 35. Kongreß der Deutschen Gesellschaft für Psychologie in Heidelberg 1986* (Band 1, S. 200). Göttingen: Hogrefe.

Bredenkamp, J., Sarris, V. (1987). Psychophysics today: Contemporary Research in the Federal Republic of Germany. *The German Journal of Psychology, 11,* 179–197.

Bredenkamp, J., Klein, K.-M., von Hayn, S., Vaterrodt, B. (1988). Gedächtnispsychologische Untersuchungen eines Rechenkünstlers. *Sprache & Kognition, 7,* 69–83.

Erdfelder, E., Bredenkamp, J. (1988). Psychophysik der Wahrnehmungstäuschungen. In H.J. Biersack, O. Breidbach, D.B. Linke u. K. Sandhoff (Hg.), *Interdisziplinäre Hirn- und Bewußtseinsforschung an der Universität Bonn. Referate eines Symposiums vom 6. Februar 1988 in Bonn* (S. 31–34). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Bredenkamp, J. (1988). Experimentelle Untersuchungen eines Rechenkünstlers. In W. Schönpflug (Hg.), *Bericht über den 36. Kongreß der Deutschen Gesellschaft für Psychologie in Berlin 1988* (Band 1, S. 25). Göttingen: Hogrefe.

Vaterrodt, B., Bredenkamp, J. (1989). Werden skript-typische Ereignisse enkodiert? *Sprache & Kognition, 8,* 126–132.

Bredenkamp, J. (1990). Kognitionspsychologische Untersuchungen eines Rechenkünstlers. In H. Feger (Hg.), *Wissenschaft und Verantwortung. Festschrift für Karl Josef Klauer* (S. 47–70). Göttingen: Hogrefe.

Bredenkamp, J., Vaterrodt, B. (1992). Direkte und indirekte Gedächtnisprüfung skriptbezogener Informationen. *Sprache & Kognition, 11,* 14–26.

Mecklenbräuker, S., Wippich, W., Bredenkamp, J. (1992). *Bildhaftigkeit und Metakognitionen.* Göttingen: Hogrefe.

Bredenkamp, J. (1993). Zur Lage der Psychologie in den neuen Bundesländern. *Psychologische Rundschau, 44,* 1–10.

Bredenkamp, J. (1993) Die Verknüpfung verschiedener Invarianzhypothesen im Bereich der Gedächtnispsychologie. *Zeitschrift für Experimentelle und Angewandte Psychologie, 40,* 368–385.

Erdfelder, E., Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann u. W. Tack (Hg.), *Enzyklopädie der Psychologie. Serie I: Forschungsmethoden der Psychologie, Band 1: Methodologische Grundlagen der Psychologie* (S. 604–648). Göttingen: Hogrefe.

Bredenkamp, J. (1995). Connection of various invariance hypotheses in the field of psychology of memory. Russian Academy of Science. *Journal of Psychology, 16,* 74–91.

Grube-Unglaub, S., Bredenkamp, J., Fischer, V., Funke, J. (1995). Investigations into the representation of scripts. *Swiss Journal of Psychology, 54,* 289–302.

Grube-Unglaub, S., Bredenkamp, J., Vaterrodt-Plünnecke, B., Fischer, V. (1995). Indirekte Gedächtnisprüfung skriptbezogener Informationen. *Zeitschrift für Experimentelle Psychologie, 42,* 1–24.

Bredenkamp, J. (1996). Grundlagen experimenteller Methoden. In E. Erdfelder, R. Mausfeld, T. Meiser u. G. Rudinger (Hg.), *Handbuch Quantitativer Methoden* (S. 37–46). Weinheim: Psychologie Verlags Union.

Bredenkamp, J., Erdfelder, E. (1996). Methoden der Gedächtnispsychologie. In D. Albert u. K.H. Stapf (Hg.), *Enzyklopädie der Psychologie. Themenbereich C: Theorie und Forschung, Serie II: Kognition, Band 4: Gedächtnis* (S. 1–94). Göttingen: Hogrefe.

Bredenkamp, J., Funke, J. (1996). A historical change: Psychology in eastern Germany after unification. *World Psychology, 2,* 103–115.

Bröder, A., Bredenkamp, J. (1996). SLIP-Technik, Prozeßdissoziationsmodell und multinomiale Modellierung: Neue Werkzeuge zum experimentellen Nachweis »Freudscher Versprecher«. *Zeitschrift für Experimentelle Psychologie, 43,* 175–202.

Vaterrodt-Plünnecke, B., Krüger, T., Gerdes, H., Bredenkamp, J. (1996). Prozeß-Dissoziations-Prozedur: Prüfbare Meßmodelle zur Erfassung von kontrollierten, automatischen und Antworttendenz-Prozessen. *Zeitschrift für Experimentelle Psychologie, 43,* 483–519.

Bredenkamp, J. (1997). Lernen. In W. Kempf, J. Straub u. H. Werbik (Hg.), *Einführung in die Psychologie* (S. 280–296). München: DTV.

Hormuth, S., Bredenkamp, J. (1997). Training and publication in German psychology. *World Psychology, 3,* 279–288

Krüger, T., Vaterrodt-Plünnecke, B., Bredenkamp, J. (1997). Prozeß-Dissoziations-Modelle: Kritik, Erweiterung und Überprüfung. In H. Mandl (Hg.), *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996* (S. 284–289). Göttingen: Hogrefe.

Schüttauf, K., Bredenkamp, J., Specht, E.K. (1997). Induzierte »Freudsche Versprecher« und zwangsneurotischer Konflikt. *Sprache & Kognition, 16,* 3–13.

Bredenkamp, J. (1998). Die Erfassung unbewußter Verarbeitungsprozesse. In K.C. Klauer u. H. Westmeyer (Hg.), *Psychologische Methoden und soziale Prozesse* (S. 30–46). Lengerich: Pabst.

Bredenkamp, J. (1998). *Lernen, Erinnern, Vergessen.* München: Beck [Bulgarische Ausgabe 2002].

Bredenkamp, J., Dilger, S. (1998). Erkundungsexperimente über den Zusammenhang von Versprechern und Arbeitsgedächtnis. *Zeitschrift für Experimentelle Psychologie, 45,* 72–79.

Bredenkamp, J., Klein, K.M. (1998). Experimental tests of a model connecting three invariance hypotheses on learning and memory processes. *Zeitschrift für Psychologie, 206,* 107–124.

Erdfelder, E., Bredenkamp, J. (1998). Recognition of script-typical versus script-atypical information: Effects of cognitive elaboration. *Memory & Cognition, 26,* 922–938.

Dilger, S., Bredenkamp, J. (2000). Arbeitsgedächtnis und Versprecher. *Sprache & Kognition, 19,* 23–30.

Ufer, N., Bredenkamp, J., Jacoby, G.E. (2000). Induzierte »Freudsche« Versprecher und Bulimie. *Zeitschrift für Klinische Psychologie und Psychotherapie, 29,* 180–186.

Bredenkamp, J. (2001). Laboratory Experiment: Methodology. In N.J. Smelser u. P.B. Baltes (Hg.), *International encyclopedia of the social & behavioral sciences* (S. 8226–8232). Oxford: Elsevier/Pergamon.

Bredenkamp, J., Hamm, S. (2001). Further experimental tests of invariance hypotheses on learning and memory processes. *Zeitschrift für Psychologie, 209,* 227–244.

Bredenkamp, J., Vaterrodt-Plünnecke, B. (2001). Kontrollierte und automatische Gedächtnisprozesse bei PTSD-Patienten und Gesunden. In R. Dohrenbusch u. F. Kaspers (Hg.), *Fortschritte der klinischen Psychologie und Verhaltensmedizin* (S. 89–97). Lengerich: Pabst.

Vaterrodt-Plünnecke, B., Krüger, T., Bredenkamp, J. (2002). Process-Dissociation Procedure: a testable model for consideration assumptions about the stochastic relation between consciously controlled and automatic processes. *Experimental Psychology, 49,* 3–26.

Bredenkamp, J. (2003). Cavanagh's Hypothesis within the context of other invariance hypotheses: Theory and data (S. 319–329) In C. Kaernbach, E. Schröger u. H. Müller (Hg.), *Psychophysics beyond sensation.* Mahwah, NJ: Erlbaum.

Dilger, S., Bredenkamp, J. (im Druck). Freudsche Versprecher in der kognitiven Psychologie. In C. Habel u. T. Pechmann (Hg.), *Sprachproduktion.* Opladen: Westdeutscher Verlag.

Hamm, S., Bredenkamp, J. (im Druck). Working memory and slips of the tongue. In T. Pechmann u. C. Habel (Hg.), *Current research on language production in Germany.* New York: Mouton.

Bredenkamp, J. (2003). Kommunikation und Versprecher. In D. Weimer u. M. Galliker (Hg.), *Sprache und Kommunikation. Ansätze und Perspektiven* (S. 119–121). Heidelberg: Asanger.

Hamm, S., Junglas, K., Bredenkamp, J. (im Druck). Die zentrale Exekutive als präartikulatorische Kontrollinstanz. *Zeitschrift für Psychologie, 212.*

(Diese Liste wurde zuletzt am 14.2.2004 aktualisiert.)

Liste der Promovenden von Jürgen Bredenkamp

(bei den mit (2) markierten Namen als Zweitgutachter)

Name	Jahr	Titel der Dissertation	Fakultät
Wippich, Werner (2)	1973	Syntaktische Komplexität und Reproduktion. Experimentelle Untersuchungen zur Reproduktion von Texten	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Hünecke, Peter	1974	Entwurf eines kognitiven Modells zur Bewegungs- und Richtungskonstanz in der Raumwahrnehmung	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Kuhr, Armin	1975	Auswirkungen eines Lehrertrainingskurses auf das Verhalten der Lehrer im Unterricht und auf die kognitiven Leistungen und die Emotionalität ihrer Schüler	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät

Name	Jahr	Titel der Dissertation	Fakultät
Hohmann, Annemarie	1976	Untersuchungen zur visuellen Neurophysiologie und Psychophysik	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Ehrenstein, Walter H. (2)	1977	Zur Psychophysik richtungsspezifischer Nachwirkungen der akustischen Bewegungswahrnehmung	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Engelhardt, Wolfgang	1977	Die Bedeutung der statistischen Interaktion in der empirischen Psychologie	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Kozielski, Peter M.	1977	Entwicklung eines aktiven Wortschatztests für drei- bis sechsjährige Kinder (AWST 3–6)	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Oetting, Manfred	1977	Die Regulation menschlicher Nahrungsaufnahme unter intern-externer Reizdiskrepanz. Ein experimenteller Ansatz zur Außenreizabhängigkeit bei Idealgewichtigen, manifest und latent Adipösen	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Wichmann, Karin	1977	Aktivation und Aufmerksamkeit im auditiven und visuellen CEP	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Baschek, Ilse	1978	Zur Abhängigkeit der Lernleistung von Informationsgehalt des Lernstoffes und der Darbietungszeit. Eine experimentelle Untersuchung zur simultanen Gültigkeit der modifizierten Invarianzhypothese und der Hypothese einer konstanten Langzeitgedächtnisspanne	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Hager, Willi	1978	Zum Lernen von Texten. Eine experimentelle Untersuchung zum Einfluß der Darbietungszeit und des Informationsgehalts auf die Leistung beim sinnzentrierten Lernen von Prosatexten	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Kiese, Christiane	1978	Validitätsuntersuchungen am aktiven Wortschatztest für drei- bis sechsjährige Kinder von Kozielski 1977	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Liebeck, Heinz (2)	1978	Untersuchungen zum Phänomen der Raumsymbolik in Psychotherapie und Psychodiagnostik. Eine empirische Absage an einen Mythos	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Schuri, Uwe (2)	1978	Zur Spontanaktivität, Reaktivität und Plastizität vegetativer Funktionen komatöser Patienten mit exogenen Intoxikationen	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät

Name	Jahr	Titel der Dissertation	Fakultät
Wilhelm, Hans (2)	1978	Beziehungen zwischen weiblichen Sexualhormonen und hirnelektrischer Spontanaktivität und dem Aspekt der Aktivierung	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Westermann, Rainer	1979	Eine wissenschafts- und messtheoretisch begründete Ableitung von Verfahren zur Messung von Einstellungen auf Intervallskalenniveau	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Mecklenbräuker, Silvia	1980	Eine Untersuchung zur Entwicklung konstruktiver Gedächtnisoperationen	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Oldenbürger, Hartmut	1981	Methodenheuristische Überlegungen und Untersuchungen zur »Erhebung« und Repräsentation kognitiver Strukturen	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Meyer, Gerhard	1982	Geldspielautomaten mit Gewinnmöglichkeit – Objekte pathologischen Glückspiels?	Uni Göttingen, Mathematisch-Naturwissenschaftliche Fakultät
Funke, Joachim	1984	Komplexes Problemlösen: kritische Bestandsaufnahme und weiterführende Perspektiven	Uni Trier, FB I Psychologie
Erdfelder, Edgar	1986	Die Entwicklung psychometrischer Intelligenz über die Lebensspanne. Aspekte eines allgemeinpsychologischen Zugangs	Uni Trier, FB I Psychologie
Willmes, Klaus	1987	Beiträge zu Theorie und Anwendung von Permutationstest in der uni- und multivariaten Datenanalyse	Uni Trier, FB I Psychologie
König-Fuchs, Claudia	1988	Therapeutischer Erfolg und Misserfolg – Kausalattributionen von Therapeuten, Klienten und Supervisoren	Uni Trier, FB I Psychologie
Hinnersmann, Herwig	1989	Theorie und Training deduktiven Urteilens	Uni Bonn, Philosophische Fakultät
Schrameier, Angela	1989	Wortbedeutung im Gedächtnis	Uni Bonn, Philosophische Fakultät
Grothgar, Barbara (2)	1990	Quasi-experimentelle Überprüfung der Veränderung psychischer und physiologischer Reaktionen von Acne vulgaris Patienten im Verlauf von dermatologisch induzierten Hautbildverbesserungen	Uni Bonn, Philosophische Fakultät
Hermanns, Norbert (2)	1991	Kognitive Wirkfaktoren auf Juckreiz und Hautreagibilität bei atopischer Dermatitis	Uni Bonn, Philosophische Fakultät
Vaterrodt, Bianca	1991	Skripts und Gedächtnis	Uni Bonn, Philosophische Fakultät

Name	Jahr	Titel der Dissertation	Fakultät
Fischer, Volkhard	1992	Ausländerstereotype und Gedächtnis	Uni Bonn, Philosophische Fakultät
Buchner, Axel	1993	Implizites Lernen: Probleme und Perspektiven	Uni Bonn, Philosophische Fakultät
Müller, Horst	1993	Komplexes Problemlösen: Reliabilität und Wissen	Uni Bonn, Philosophische Fakultät
Meyer, Matthias	1994	Serielle versus parallele Verarbeitung visueller Information. Eine experimentelle Untersuchung am Beispiel eines Konzentrationstests	Uni Bonn, Philosophische Fakultät
Klein, Klaus-Martin	1995	Experimentelle Untersuchungen zu zwei Invarianzhypothesen des Kurzzeitgedächtnisses	Uni Bonn, Philosophische Fakultät
Gerdes, Heike (2)	1997	Lernen mit Text und Hypertext	Uni Trier, FB I Psychologie
Bröder, Arndt (2)	1999	Take the Best – Ignore the Rest. Untersuchungen zur psychologischen Realität einer begrenzt rationalen Entscheidungsheuristik	Uni Heidelberg, Fakultät für Sozial- und Verhaltenswissenschaften
Krüger, Thomas (2)	1999	Die Erfassung bewußter und unbewußter Gedächtnisprozesse	Uni Heidelberg, Fakultät für Sozial- und Verhaltenswissenschaften
Dilger, Stefan	2000	Arbeitsgedächtnis und Versprecher	Uni Bonn, Philosophische Fakultät
Wegener, Ingo (2)	2000	Soziale Kategorisierung im situativen Kontext: Kognitive Flexibilität bei der Personenwahrnehmung	Uni Bonn, Philosophische Fakultät
Hamm, Silke	2001	Synthese von Kurz- und Langzeitgedächtnisspanne: Experimentelle Überprüfung eines mathematischen Modells	Uni Bonn, Philosophische Fakultät
Kaspers, Frank-A. (2)	2001	Endokrinologische Korrelate körperlicher Beschwerden bei chronischem Stress	Uni Bonn, Philosophische Fakultät
Massen, Christina	2001	Exekutive Kontrolle und sakkadische Augenbewegungen: Inhibitionsmechanismen in der Antisakkadenaufgabe	Uni Bonn, Philosophische Fakultät
Ufer, Nicole (2)	2001	Induzierte »Freudsche Versprecher« und Eßstörungen	Uni Heidelberg, Fakultät für Sozial- und Verhaltenswissenschaften

Die Autorinnen und Autoren

Brandt, Martin, geboren 1966, Dipl.-Psych. 1993 in Bonn, Dr. rer. nat. 2001 in Trier. Akademischer Rat an der Universität Mannheim, Lehrstuhl für Allgemeine und Differentielle Psychologie. Arbeitsschwerpunkte: Experimentelle Gedächtnispsychologie und formale Modelle in der kognitiven Psychologie.
Lehrstuhl für Psychologie III, Universität Mannheim, Schloss Ehrenhof Ost 239, D-68131 Mannheim. E-Mail: brandt@psychologie. uni-mannheim.de; http://www.psychologie.uni-mannheim.de/psycho3/ personen.html

Bröder, Arndt, geboren 1968, Dipl.-Psych. 1995 in Bonn, Dr. phil. 1999 in Heidelberg; Wissenschaftlicher Assistent in der Abteilung für Allgemeine Psychologie, Psychologisches Institut der Universität Bonn. Arbeitsschwerpunkte: experimentelle Erforschung begrenzt rationaler Entscheidungsstrategien, Kontextgedächtnis und Gedächtnistäuschungen.
Psychologisches Institut der Universität Bonn, Römerstraße 164, D-53117 Bonn. E-Mail: broeder@uni-bonn.de; http://www.psychologie.uni-bonn.de/allgm/neu/mitarb/aktuell/broede_a/broede_a.htm

Buchner, Axel, geboren 1961, Dipl.-Psych. 1988 in Trier, Promotion 1992 in Bonn, Habilitation 1998 in Trier, seit 2000 Professor für Allgemeine Psychologie und Arbeitspsychologie in Düsseldorf. Arbeitsschwerpunkte: Experimentelle Aufmerksamkeits-, Lern- und Gedächtnisforschung, Ergonomie.
Institut für Experimentelle Psychologie, Heinrich-Heine-Universität, D-40225 Düsseldorf. E-Mail: axel.buchner@uni-duesseldorf.de; http://www.psycho.uni-duesseldorf.de/aap/leute/buchner/

Erdfelder, Edgar, geboren 1953, Dipl.-Psych. 1980 in Göttingen, Dr. rer. nat. 1986 in Trier, Habilitation 2000 in Bonn, von 2001 bis 2002 Professor für Psychologische Methodenlehre in Gießen, seit 2002 Professor für Allgemei-

ne und Differentielle Psychologie in Mannheim. Arbeitsschwerpunkte: Experimentelle Gedächtnispsychologie und formale Modellbildung.
Lehrstuhl für Psychologie III, Universität Mannheim, Schloss Ehrenhof Ost 255, D-68131 Mannheim. E-Mail: erdfelder@psychologie.uni-mannheim.de; http://www.psychologie.uni-mannheim.de/psycho3/personen.html

Faul, Franz, geboren 1960, Dipl.-Psych. 1993 in Bonn, Dr. rer. nat. 1997 in Kiel. Wissenschaftlicher Angestellter an der Universität Kiel, Lehrstuhl für Allgemeine Psychologie II. Arbeitsschwerpunkt: Visuelle Wahrnehmung.
Institut für Psychologie, Universität Kiel, Olshausenstr. 62, D-24098 Kiel. E-Mail: ffaul@psychologie.uni-kiel.de; http://www.psychologie.uni-kiel.de/lehrstuehle.shtml

Funke, Joachim, geboren 1953, Dipl.-Psych. 1980 in Trier, Dr. rer. nat. 1984 in Trier, Habilitation 1990 in Bonn, Professor am Psychologischen Institut in Heidelberg. Arbeitsschwerpunkte: Gedächtnis, Denken und Problemlösen.
Psychologisches Institut der Universität Heidelberg, Hauptstr. 47–51, D-69117 Heidelberg. E-Mail: joachim.funke@psychologie.uni-heidelberg. de; http://funke.uni-hd.de/

Gadenne, Volker, geboren 1948. Studium der Psychologie, Philosophie und Wissenschaftstheorie in Mannheim. Promotion 1978 und Habilitation 1983 in Mannheim. 1989–94 Professor für Sozialwissenschaftliche Methodenlehre an der Universität der Bundeswehr München. Seit 1994 Professor für Philosophie und Wissenschaftstheorie an der Johannes-Kepler-Universität Linz. Arbeitsschwerpunkte: Wissenschaftstheorie, Philosophie des Geistes.
Institut für Philosophie und Wissenschaftstheorie, Johannes-Kepler-Universität Linz, Freistädterstr. 313, A-4040 Linz. E-Mail: Volker. Gadenne@jku.at; http://www.iwp.uni-linz.ac.at/gadenne/

Hager, Willi, geboren 1950, Dipl.-Psych. 1977 in Göttingen, Dr. rer. nat. 1978 in Göttingen, Habilitation 1990 in Göttingen, Professor am Georg-Elias-Müller Institut für Psychologie in Göttingen. Arbeitsschwerpunkte: Evaluationsforschung, Statistik und Versuchsplanung.
Georg-Elias-Müller-Institut für Psychologie, Goßlerstr. 14, D-37073 Göttingen. E-Mail: whager@gwdg. de; http://www.psych.uni-goettingen.de/abt/1/hager/index.shtml

Hamm, Silke, geboren 1974, Dipl.-Psych. 1998 in Bonn, Dr. phil. 2001 in Bonn, Wissenschaftliche Assistentin am Psychologischen Institut in Bonn. Arbeitsschwerpunkte: Gedächtnis und Sprache.

Psychologisches Institut der Universität Bonn, Römerstr. 164, D-53117 Bonn. E-Mail: s.hamm@uni-bonn.de; http://www.psychologie.uni-bonn.de/allgm/neu/mitarb/ aktuell/hamm_s/hamm_s.htm

Klauer, Karl Christoph, geboren 1961, Dipl.-Math. 1984 in Aachen, Dipl.-Psych. 1985 in Hamburg, Dr. phil. 1988 in Hamburg, Habilitation 1992 in Berlin, Professor am Psychologischen Institut in Bonn. Arbeitsschwerpunkte: Soziale Kognition, Denken, Arbeitsgedächtnis und Methoden.
Psychologisches Institut der Universität Bonn, Römerstr. 164, D-53117 Bonn. E-Mail: christoph.klauer@uni-bonn.de; http://www.psychologie.uni-bonn. de/sozial/staff/klauer/klauer.htm

Marx, Edeltrud, geboren 1953, Dipl.-Psych. 1981 in Münster, Dr. phil. 1985 in Münster, Habilitation 1999 in Köln, Professorin für Psychologie an der Katholischen Fachhochschule NRW in Köln. Arbeitsschwerpunkte: Sprache und Intelligenz.
Katholische Fachhochschule NRW, Abteilung Köln, Wörthstraße 10, D-50668 Köln. E-Mail: e.marx@kfhnw.de; http://www. kfhnw.de/personen/ sp_auto_1513.php?abteilung = koeln

Massen, Cristina, geboren 1969, Dipl.-Psych. 1996 in Bonn, Dr. phil. 2001 in Bonn, Wissenschaftliche Assistentin in der Abteilung für Allgemeine Psychologie, Psychologisches Institut der Universität Bonn. Arbeitsschwerpunkte: Kontrolle sakkadischer Augenbewegungen, visuell-räumliche Aufmerksamkeit und Gedächtnisprozesse.
Psychologisches Institut der Universität Bonn, Römerstraße 164, D-53117 Bonn. E-Mail: massen@uni-bonn.de; http://www.psychologie.uni-bonn.de/allgm/neu/mitarb/aktuell/ massen_c/massen_c.htm

Siemer, Matthias, geboren 1964, Dipl.-Psych. 1990 in Berlin, Dr. phil. 1996 in Berlin, Wissenschaftlicher Assistent am Institut für Psychologie in Greifswald. Arbeitsschwerpunkte: Emotionen, Methoden, Wissenschaftstheorie.
Institut für Psychologie der Ernst-Moritz-Arndt-Universität, Franz-Mehring-Str. 47, D-17487 Greifswald. E-Mail: masiemer@uni-greifswald.de; http://www.psychologie.uni-greifswald.de/allg/siemer.htm

Steyer, Rolf, geboren 1950, Diplom im Fach Psychologie 1977 an der Universität Göttingen, Promotion 1982 an der Universität Frankfurt am Main, Habilitation 1989 an der Universität Trier. Von 1994 bis 1995 Wissenschaftlicher Leiter am Zentrum für Methoden, Umfragen und Analysen (ZUMA) in Mannheim, von 1995 bis 1996 Professor für Methodik und Diagnostik

der Psychologie an der Universität Magdeburg; seit 1996 Inhaber des Lehrstuhls für Methodenlehre und Evaluationsforschung am Institut für Psychologie der Universität Jena, seit September 2002 Prorektor der Friedrich-Schiller-Universität Jena, seit Herbst 2002 Präsident des Zentrums für Humanwissenschaftliche Forschung. Arbeitsschwerpunkte: Theorie kausaler stochastischer Abhängigkeit, Latent-State-Trait-Modelle, Stochastische Messmodelle, Stimmungsmessung, Veränderungsmessung.
Friedrich-Schiller-Universität Jena, Institut für Psychologie, Am Steiger 3/Haus 1, 07743 Jena. E-Mail: rolf.steyer@ uni-jena.de; http://www.uni-jena.de/svw/metheval/

Vaterrodt-Plünnecke, Bianca, geboren 1964, Dipl.-Psych. 1989 in Bonn, Dr. phil. 1992 in Bonn. Arbeitsschwerpunkte: Lern- und Gedächtnispsychologie, Unbewusste Informationsverarbeitung, Gedächtnistraining, Pädagogische Diagnostik.
Psychologisches Institut der Universität Bonn, Römerstraße 164, 53117 Bonn. E-Mail: Bianca.Vaterrodt@uni-bonn.de; http://www. psychologie. uni-bonn.de/allgm/

Westermann, Rainer, geboren 1950, Dipl.-Psych. 1976; Dr. rer. nat. 1979; Habilitation 1986 in Göttingen; Professor für Allgemeine Psychologie und Forschungsmethodik an der Universität Greifswald, seit 2003 Rektor der Universität Greifswald. Arbeitsschwerpunkte: Experimentalmethodik, Wissenschaftsphilosophie, Evaluation, Handlungsregulation.
Ernst-Moritz-Arndt-Universität Greifswald, Institut für Psychologie, Franz-Mehring-Straße 47, D-17487 Greifswald. E-Mail: westerma@uni-greifswald.de; http://www.psychologie.uni-greifswald.de/allg/westerma.htm

Willmes, Klaus, geboren 1950, Dipl.-Math. 1974 in Aachen, Dipl.-Psych. 1979 in Aachen; Dr. rer. nat. 1987 in Trier; Habilitation 1994 in Bielefeld; Professor für Neuropsychologie an der RWTH Aachen. Arbeitsschwerpunkte: experimentelle (Neuro-)Kognitionsforschung zur Zahlenverarbeitung und zur Aufmerksamkeit einschließlich funktioneller Bildgebung, neuropsychologische Therapieforschung, psychometrische und statistische Methoden in der Neuropsychologie.
Lehr- und Forschungsgebiet Neuropsychologie – Neurologische Klinik, Universitätsklinikum RWTH Aachen, Pauwelsstraße 30, D-52074 Aachen. E-Mail: willmes@neuropsych.rwth- aachen.de

Psychologie bei V&R

Gerd Lüer / Uta Lass (Hg.)
Erinnern und Behalten
Wege zur Erforschung des menschlichen Gedächtnisses

Abhandlungen der Akademie der Wissenschaften zu Göttingen, Math.-Phys. Klasse III/47. 1997. 378 Seiten mit 7 Farbtafeln, 63 Abbildungen und 6 Tabellen, Leinen
ISBN 3-525-82124-7

Dieser Band informiert über vier unterschiedliche Richtungen zur Erforschung des menschlichen Gedächtnisses: Psychobiologie, Neuropsychologie, Kognitionspsychologie und ökologische Psychologie. Für jeden der vier Bereiche dokumentiert ein erster Beitrag den aktuellen Stand der Forschung. Daran schließen sich Beiträge über Einzeluntersuchungen an, die in den Kontext des Übersichtsartikels eingeordnet werden können. Sie demonstrieren am Beispiel einer konkreten Forschungsfrage die punktuell vertiefte Behandlung eines Themas. Den fachspezifischen Beiträgen ist ein Kapitel vorangestellt, das einem Pionier der gedächtnispsychologischen Forschung, Georg Elias Müller, gewidmet ist, sowie eine Darstellung der Evolution geistiger Prozesse.

Annette Kämmerer / Joachim Funke (Hg.)
Seelenlandschaften
Streifzüge durch die Psychologie
98 persönliche Positionen

2004. Ca. 220 Seiten mit einigen Abbildungen, kartoniert
ISBN 3-525-46206-9

Eine bunt gemischte Auswahl von 98 Professorinnen und Professoren für Psychologie aus den deutschsprachigen Ländern haben auf knappem Raum etwas ganz Persönliches dargestellt – etwas, das im Alltagsgeschäft der Psychologie keinen Platz findet, mit dem aber etwas angesprochen wird oder in dem etwas anklingt, was den speziellen Reiz der Psychologie oder des Psychologischen für sie ausmacht. Das Buch ist ein buntes Kaleidoskop geworden, es sind im wahrsten Sinn Streifzüge durch Seelenlandschaften.

V&R
Vandenhoeck & Ruprecht