14. Kapitel

# Hypothesenprüfung

***Edgar Erdfelder und Jürgen Bredenkamp***

Hypothesenprüfungen stehen im Dienste des übergeordneten Zieles wissenschaftlichen Arbeitens, beobachtbare Ereignisse und Gesetzmäßigkeiten stichhaltig ***erklären*** zu können. Wissenschaftliche Ereignis- oder Gesetzeserklärungen setzen in der Psychologie wie in anderen Erfahrungswissenschaften Hypothesen oder Theorien voraus, die u. a. dem empirischen Adäquatheitskriterium genügen müssen und sich somit - wie man sagt - „bewährt" haben (vgl. Westmeyer, 1973; Groeben & Westmeyer, 1981[2]; Gadenne, Kapitel 7 und 9 dieses Bandes). Die methodologischen Probleme, mit denen sich das vorliegende Kapitel auseinandersetzt, resultieren aus der Tatsache, daß durch empirische Untersuchungen im Regelfall nicht ohne weiteres festgestellt werden kann, ob Hypothesen und Theorien wahr oder falsch sind. Dies hat zum Teil ***logische,*** zum Teil aber auch ***forschungspraktische*** Gründe. Zu den logischen Gründen zählt etwa die Unmöglichkeit der Verifikation allgemeiner Hypothesen (Popper, 1982[7]), die als Bestandteile des Explanans wissenschaftlicher Erklärungsargumente von zentraler Bedeutung sind, aber auch die Tatsache, daß statistische Hypothesen, die in der Psychologie eine wichtige Rolle spielen, mit allen denkbaren empirischen Ereignissen prinzipiell kompatibel sind. In forschungspraktischer Hinsicht ergibt sich nicht selten das Problem, daß auch über deterministische und singuläre Hypothesen ohne Allgemeinheitsgrad empirisch nicht eindeutig entschieden werden kann, weil die Untersuchungsdaten fehlerbelastet sind oder eine direkte Hypothesenprüfung an einzelnen Individuen Verletzungen von Annahmen zur Folge hätte, die Bestandteile der zu prüfenden Hypothesen sind. Alle genannten Probleme und die daraus ableitbaren Konsequenzen für die Forschungspraxis werden in den folgenden Abschnitten anhand von Beispielen verdeutlicht und präzisiert.

Obwohl der modalen psychologischen Hypothese (PH) weder Verifizierbarkeit noch Falsifizierbarkeit im Sinne Poppers (1982[7]) zugesprochen werden kann, liegt kein Grund zur Resignation vor. Die Methodologie der Prüfung von PHn muß so ausgearbeitet sein, daß sie Regeln in die Hand gibt, deren Anwendung psychologische Untersuchungen so informativ macht, daß sich

Forscher bei einer bestimmten Befundlage ***entschließen*** können, die Theorie fallen zu lassen bzw. durch eine bessere Theorie zu ersetzen. Diese Position entspricht nach Lakatos (1970) nicht einem „dogmatischen", sondern einem ,,methodologischen Falsifikationismus". Der dogmatische Falsifikationismus geht von einer unfehlbaren empirischen Basis aus: Wenn eine Theorie falsifiziert wird, ist ihre Falschheit bewiesen. Im Rahmen des methodologischen Falsifikationismus kann eine empirisch nicht bewährte Theorie dagegen u.U. wahr sein. Demzufolge ist nicht nur das Risiko eines fälschlichen Bewährungsurteils (mangelnde „Strenge" einer Untersuchung), sondern auch das Risiko eines fälschlichen Nichtbewährungsurteils (mangelnde ,,Fairneß" einer Untersuchung) bei der methodologischen Analyse hypothesenprüfender empirischer Untersuchungen im Blick zu behalten (vgl. Gadenne, 1976, S.70; Hager, 1987, 1992; Westermann, 1987).

In diesem Kapitel wird eine Methodologie der Überprüfung von PHn vorgestellt, die den genannten Forderungen Rechnung trägt. Sie kann „deduktivistisch" in dem Sinne genannt werden, daß sie auf induktive Elemente, wie sie etwa in der Methodologie Campbell und Stanleys (1963) auszumachen sind (Gadenne, 1976, 1984; vgl. auch Erdfelder, Kapitel 2 dieses Bandes), soweit wie möglich verzichtet (zum Begriff der Induktion vgl. Westermann & Gerjets, Kapitel 10 dieses Bandes). Sie kann weiterhin als Methodologie der ***statistischen*** Prüfung von PHn bezeichnet werden, da sie davon ausgeht, daß - wie in der psychologischen Forschungspraxis üblich - statistische Hypothesen (SHn) als Prüfinstanzen für PHn fungieren. Mit dieser Schwerpunktsetzung wollen wir keineswegs suggerieren, daß die gegenwärtig dominierende Forschungspraxis die allein mögliche oder allein sinnvolle Form empirischer Psychologie darstellt. Empirische Prüfungen von PHn müssen nicht notwendig auf statistische Hypothesentests hinauslaufen. Andere, deskriptivstatistische oder auch nichtstatistische Wege der Überprüfung sind denkbar und u.U. auch sinnvoll. Dennoch scheint uns der Versuch einer rationalen Rekonstruktion der statistischen Prüfung von PHn allein deshalb von besonderer Bedeutung zu sein, weil dieses Vorgehen offenbar von der großen Mehrzahl empirisch arbeitender Psychologen präferiert wird. Dies allein rechtfertigt die in diesem Kapitel gewählte Perspektive.

Im ersten Abschnitt werden die wichtigsten Aspekte einer deduktivistischen Methodologie empirischer Untersuchungen in der Psychologie kurz skizziert, wobei insbesondere die Funktion der Inferenzstatistik im Überprüfungsprozeß in den Mittelpunkt geruckt wird. Dieser Abschnitt stützt sich wesentlich auf Arbeiten von Bredenkamp (1969, 1972, 1980), Gadenne (1976, 1984) und Hager und Westermann (1983b, 1983c). Der zweite Abschnitt ist der Explikation der Begriffe „Strenge", ,,Fairneß" und ,,Validität" einer Hypothesenprüfung sowie den methodologischen Folgerungen gewidmet, die sich aus der Forderung nach möglichst validen Untersuchungen ergeben. Zentrale Be-

standteile der deduktivistischen Methodologie werden in diesem Abschnitt formal analysiert und begründet. Der dritte Abschnitt gibt einige Empfehlungen zur statistischen Entscheidungsstrategie, die möglichst strenge und faire Prüfungen gewährleisten soll. Der vierte Abschnitt schließlich hat speziell das Theorie-Empirie-Überbrückungsproblem zum Gegenstand, d.h. die Frage, wie in geeigneter Form eine Beziehung zwischen theoretischen Begriffen in PHn und beobachtbaren Tatbeständen hergestellt werden kann.

## *1. Grundzüge einer deduktivistischen Theorie hypothesentestender Untersuchungen*

Empirische Untersuchungen in der Psychologie zeichnen sich fast ausnahmslos dadurch aus, daß substanzwissenschaftliche Fragestellungen mittels statistischer Hypothesentests beantwortet werden. Bemerkenswert ist dieses Vorgehen vor allem deshalb, weil die interessierende PH und die letztlich getestete SH im Regelfall nicht identisch sind. PHn beziehen sich typischerweise auf psychische Vorgänge, die im einzelnen Individuum ablaufen. Beispiele hierfür sind die allgemeinpsychologischen Hypothesen „Imaginale Verarbeitungsprozesse führen im Vergleich zu sprachlichen Verarbeitungsprozessen zu überlegenen Gedächtnisleistungen“, „Erwachsene Personen bewältigen konzeptuelle Probleme durch das Testen von Hypothesen“, „Die Empfindungsstärke ist eine logarithmische Funktion der Reizstärke”, um nur einige Beispiele zu nennen. Auf einzelne Individuen bezogene Hypothesen finden sich nicht nur in der Allgemeinen Psychologie, sondern z.B. auch in der Sozialpsychologie: „Wer öffentlich unter Druck eine Meinung äußert, die seiner eigenen Meinung widerspricht, tendiert dazu, seine Meinung in Richtung der Äußerung abzuändern.“ Statistisch getestet werden aber in der Allgemeinen Psychologie wie in der Sozialpsychologie häufig Populationsaussagen der Art, daß die Mittelwerte aus zwei oder mehr Populationen, in denen die interessierende abhängige Variable (AV) jeweils normal und mit gleichen Varianzen verteilt ist, identisch bzw. nicht identisch sind. Wenn dieses Vorgehen nicht schlicht am eigentlichen Zweck der Untersuchung, nämlich eine vorgegebene PH testen zu wollen, vorbei gehen soll, muß es eine Verknüpfung zwischen PH und SH geben, die so gestaltet ist, daß der statistische Test aussagekräftig in bezug auf die PH ist.

### 1.1 Warum werden psychologische Hypothesen statistisch geprüft?

Anhand zweier Beispiele soll zunächst untersucht werden, warum PHn auf dem - wie es scheint - „Umweg“ über die Statistik geprüft werden. Das erste Beispiel bezieht sich auf die Invarianzhypothese ***(total time hypothesis,*** vgl.

Baddeley, 1990) des verbalen Lernens, nach der die Gesamtlernzeit bis zur Beherrschung eines Stoffes eine individuelle Konstante ist, unabhängig davon, in wieviele Lernversuche diese Zeit aufgeteilt wird. Es muß demnach für alle Individuen $T_i = Y_{ij} \cdot t_j$ gelten, wobei $T_i$ die Gesamtlernzeit des Individuums $i$, $t_j$ die pro Lernversuch gewährte Bearbeitungszeit und $Y_{ij}$ die benötigte Anzahl der Lernversuche von Individuum $i$ bei Bearbeitungszeit $t_j$ pro Lernversuch ist. Verglichen werden üblicherweise die mittleren Gesamtlernzeiten von Versuchspersonen (Vpn), die den verschiedenen Versuchsbedingungen (unterschiedliche Werte für $t_j$) zufällig zugewiesen wurden (Bugelski, 1962). Unter diesen Umständen folgt aus der Gültigkeit der Invarianzhypothese für alle Individuen die statistische Hypothese $E(T) = E(Y_j) \cdot t_j$, wobei die Gesamtlernzeit $T$ und die Anzahl der Lernversuche $Y_j$ bei Bearbeitungszeit $t_j$ nunmehr als Zufallsvariablen (ZVn) aufzufassen sind. Von zentraler Bedeutung für die Implikationsbeziehung ist, daß mit den verschiedenen Werten für $t_j$ nicht andere Variablen wie z.B. die Lernfähigkeit der Vpn korreliert sind (vgl. Abschnitt 2.4). Derartige Konfundierungen sollen durch die Randomisierung verhindert werden.

Die Anwendung der Statistik ist hier offenbar sinnvoll. Die o. g. SH entspricht der Nullhypothese ($H_1$) einer Varianzanalyse, wenn vorausgesetzt wird, daß die Gesamtlernzeiten innerhalb aller Bedingungen normal und mit gleichen Varianzen verteilt sind. Wird durch den statistischen Test die $H_0$ nicht zurückgewiesen, so liegt ein Ergebnis vor, das der Invarianzhypothese nicht widerspricht. Muß dagegen $H_0$ zurückgewiesen werden, widerspricht der Ausgang des Tests der PH.

Dennoch ist zu fragen, warum die Invarianzhypothese ***auf diese*** Weise geprüft wird. Denkbar wäre auch ein Vorgehen, bei dem jede Vp unter allen Versuchsbedingungen lernt und sodann geprüft wird, ob die Invarianzhypothese für jede Vp bestätigt werden kann. Ginge man so vor, müßte dennoch auch in diesem Fall Statistik angewendet werden. Wegen der Fehlerbehaftetheit der Messungen sind nicht alle Gesamtlernzeitwerte eines Individuums unter konstanten Lernbedingungen identisch, und der statistische Test hätte zu prüfen, ob trotz dieser Abweichungen ein Ergebnis vorliegt, das $H_0$ entspricht. Allerdings müßte in einem derartigen Experiment mit wiederholten Messungen aus ersichtlichen Gründen jede Vp je Versuchsbedingung unterschiedliche Lernstoffe bewältigen. Um Aufwärmeffekte zu vermeiden, müßte die Untersuchung jeder Vp auf mehrere Tage verteilt werden, was nicht ausschließt, daß sie Strategien der Auseinandersetzung mit der Art des Lernstoffes erwerben, die das spätere Lernen schneller ablaufen lassen. Derartige Aufgaben- und Strategievariablen wären für jedes Individuum mit den interessierenden Versuchsbedingungen vermischt, so daß die Invarianzhypothese je Individuum nicht ohne Ausschaltung dieser Störvariablen geprüft werden könnte. Erreicht werden könnte in einer Untersuchung, in der jede Vp unter allen Bedingungen

$t_j$. untersucht wird, daß ***im Durchschnitt*** Aufgaben- und Strategievariablen nicht mit $t_j$ korrelieren. In diesem Fall ließe sich wieder eine auf interindividuell aggregierte Daten bezogene SH prüfen. Auch die intraindividuelle Variation der Bedingungen schließt also in diesem Beispiel - und in vielen anderen auch - die Überprüfung der PH am einzelnen Individuum aus. Zu viele Störvariablen beeinflussen den Ausgang dieser Prüfung. Das gilt nicht für die Prüfung einer Populationsaussage, die, wie gesagt, sinnvoll ist, wenn sie mit der PH deduktiv Verknüpfbar ist.

Betrachten wir noch ein zweites Beispiel. Die PH sei: „Wer unter Druck eine Meinung äußert, die seiner eigenen Meinung widerspricht, tendiert dazu, seine Meinung in Richtung der Äußerung zu verändern." Diese Hypothese gehört zur Theorie der kognitiven Dissonanz. In ihr treten theoretische Terme auf, deren Operationalisierung in einem Experiment von Festinger und Carlsmith (1959) versucht wurde. Darauf ist hier nicht einzugehen (vgl. dazu Abschnitt 4). Geprüft werden könnte diese PH an jedem Individuum, wenn folgendermaßen vorgegangen wird: Zuerst ist ohne Druckanwendung die Meinung zu einem Sachverhalt zu erkunden. Danach muß die Vp unter Druckanwendung dazu gebracht werden, eine ihrer eigenen Auffassung widersprechende Meinung zu äußern, bevor geprüft wird, ob sich die tatsächliche Auffassung geändert hat. Die Gefahr, daß ein Versuchsteilnehmer die Hypothese durchschaut und willentlich Ergebnisse produziert, die ihr entsprechen oder widersprechen, ist groß. Deshalb verbot sich für Festinger und Carlsmith (1959) die wiederholte Erfassung von Meinungen jeder Vp. Sie wiesen die Vpn zufällig den Versuchsbedingungen „Druck" und „kein Druck" zu und erhoben zu einem Zeitpunkt, als die Vpn glaubten, das Experiment sei schon beendet, nur einmal die Meinung zu dem Experiment. Verglichen wurden die durchschnittlichen Meinungen der Versuchsgruppen. Das ist sinnvoll, wenn aus der PH folgt, daß im Durchschnitt die unter Druck stehenden Vpn eine als ausgesprochen monoton empfundene Tätigkeit nicht als so langweilig beurteilen wie die Vpn der Kontrollgruppe. Auch hier wurde eine auf Individuen bezogene PH auf dem Umweg über die Prüfung einer Populationsaussage getestet, um den Einfluß von Störvariablen zu kontrollieren.

Derartige Beispiele ließen sich beliebig fortsetzen (vgl. Bredenkamp 1972, 1980). Das Testen von statistischen Populationsaussagen, denen auf Individuen bezogene PHn vorgeordnet sind, ist typisch für weite Bereiche der Psychologie. Sinnvoll ist dieses Vorgehen aber nur dann, wenn eine Verknüpfung zwischen beiden Aussagen konstruiert werden kann. Im Falle der Invarianzhypothese war dies die implikative Verknüpfung von PH und $H_o$: $\mu_1 = \mu_2 = \ldots = \mu_m$. Im Fall der kognitiven Dissonanzhypothese war dies die Implikationsbeziehung PH $\Rightarrow$ H,: $\mu_1 < \mu_2$. Selbstverständlich muß diese implikative Beziehung im Einzelfall bewiesen werden (vgl. Abschnitt 2.4), und außerdem müssen bestimmte statistische Annahmen gelten, wenn man als sta-

tistisches Prüfverfahren die Varianzanalyse bzw. den t-Test verwendet. Wenn wir von diesen Annahmen zunächst einmal absehen, läßt sich sagen, daß bei Gültigkeit der implikativen Verknüpfung ein Ergebnis, das für die Verneinung der implizierten SH spricht, notwendig, aber nicht hinreichend für die Falsifikation der PH ist'. Eine notwendige, aber nicht hinreichende Bedingung für die Falsifikation liegt vor, weil im Rahmen eines sophistizierten methodologischen Falsifikationismus eine Hypothese erst falsifiziert werden sollte, wenn eine bessere zur Verfügung steht (Lakatos, 1970).

## 1.2 Die Notwendigkeit der simultanen Kontrolle von $\alpha$ und ß

Wenn eine implikative Beziehung zwischen PH und SH vorliegt, ist weiterhin zu gewährleisten, daß neben der statistischen Fehlerwahrscheinlichkeit $\alpha$ auch die Fehlerwahrscheinlichkeit ß kontrolliert wird. Bekanntlich ist $\alpha$ die Wahrscheinlichkeit für eine falsche Verwerfung von $H_0$, ß die für eine falsche Akzeptanz von $H_0$. Die Forderung auch der Kontrolle von ß, die sich - wie in Abschnitt 2 gezeigt wird - aus der Verknüpfung von PH und SH ergibt, zeigt die Relevanz der Neyman-Pearson-Testtheorie für die psychologische Forschung auf (vgl. Ostmann & Wutke, Kapitel 16 dieses Bandes). Sie erlaubt, vor der Durchführung einer Untersuchung unter Annahme bestimmter Verteilungsvoraussetzungen den Stichprobenumfang so zu bestimmen, daß eine Abweichung bestimmter Größe von $H_0$ mit der Wahrscheinlichkeit $1-\beta$ entdeckt werden kann[2]. Auf bestimmte Probleme und Lösungsvorschläge im Zusammenhang mit der Verfolgung von Entscheidungsstrategien im Rahmen der Neyman-Pearson-Testtheorie gehen wir in Abschnitt 3 ein (ausführlich dazu Bredenkamp 1972, 1980; Hager & Westermann, 1983c; Hager, 1987; Ostmann & Wutke, Kapitel 16 dieses Bandes). In Abschnitt 2 wird der Bezug der Wahrscheinlichkeiten $\alpha$ und ß zur Strenge und Fairneß einer Untersuchung dargestellt.

Die implikative PH-SH-Verknüpfung und die Durchführung eines Signifikanztests, der neben $\alpha$ auch ß kontrolliert, können die Falsifikation einer PH selbstverständlich niemals logisch erzwingen. Bei kleinen Fehlerwahrscheinlichkeiten ist aber die Untersuchung so aussagekräftig, daß man sich u. U. zu einer Falsifikation ***entschließen*** kann. Die Strategie ist also dem methodologischen Falsifikationismus zuzuordnen. Lakatos (1970, S. 109) betont ausdrücklich, daß „the Neyman-Pearson approach rests completely on methodological falsificationism". Daß in der Psychologie die nach Neyman-Pear-

1 Die Alternative zur implizierten SH hat hier eine ähnliche Funktion wie die falsifizierenden Basissätze in der Falsifikationstheorie von Popper (1982').

2 Die Sequentialstatistik nach Wald (z.B. Wald, 1948) ist in manchen Fällen eine ökonomische Alternative zu Entscheidungsstrategien, die von einer Festlegung des Stichprobenumfangs a priori ausgehen.

sonstrategien testbaren Populationsaussagen häufig nicht die interessierenden PHn sind, liegt nicht an der fehlenden Falsifizierbarkeit im Sinne des dogmatischen Falsifikationismus, sondern ausschließlich daran, daß PHn im Gegensatz zu Populationsaussagen auf individuelle Prozesse abzielen. Die auf Individuen bezogenen PHn wurden von Bredenkamp (1972) als deterministische oder statistische Allaussagen interpretiert, die sich auf eine offene Population von Vpn beziehen. Derartige Aussagen sind notwendig, um dem Erklärungsanspruch genügen zu können (vgl. Groeben & Westmeyer, 1981[2]). Zugleich wirft diese Interpretation aber auch das Problem der Rechtfertigung von Verteilungsannahmen statistischer Tests auf. Hierüber sagen individuenbezogene PHn gewöhnlich nichts aus. Der nächste Abschnitt zeigt, wie dieses Problem im Rahmen der experimentalpsychologischen Forschung gelöst werden kann.

## 1.3 Die Bedeutung von Randomisierungstests

Verteilungsannahmen werden hinfällig, wenn statt der üblichen parametrischen Signifikanztests sog. Randomisierungstests Verwendung finden. Wir erläutern das Prinzip des Randomisierungstests, das von R. A. Fisher stammt und für die experimentalpsychologische Forschung von Edgington (1969) propagiert wurde, anhand einer varianzanalytischen Fragestellung, wie sie z.B. der Prüfung der Invarianzhypothese (vgl. Abschnitt 1.1) zugrunde liegt. Sei $n_j$ $(j = 1,..., 3)$ die Anzahl der unter der j-ten Versuchsbedingung erhobenen Meßwerte. Es sind dann $(n_1 + n_2 + n_3)!/(n_1!\ n_2!\ n_3!)$ Aufteilungen dieser Werte auf die drei Versuchsbedingungen möglich, die unter $H_0$ alle gleich wahrscheinlich sind. Für jede Aufteilung kann die bekannte varianzanalytische *F*-Statistik berechnet werden. Fällt die F-Statistik für die empirisch tatsächlich eingetretene Aufteilung der Meßwerte unter die 100 . $\alpha$ %-Aufteilungen mit dem größten F-Wert, wird $H_0$ zurückgewiesen; andernfalls wird $H_0$ beibehalten.

Das Verfahren macht keine Verteilungsvoraussetzungen. Es prüft, ob für die ***Vpn des Experiments*** $H_0$ zurückzuweisen ist; Verallgemeinerungen auf andere Personen, die nicht am Experiment teilgenommen haben, sind nicht möglich und werden auch nicht angestrebt. In der deduktivistischen Methodologie sind alle Personen, die zum Geltungsbereich der PH gehören, gleichermaßen repräsentativ. Wenn mit mehr als einer Vp experimentiert wird, dann nicht, um die Repräsentativität einer Untersuchung zu erhöhen, sondern um PHn strenger überprüfen zu können.

Der Randomisierungstest paßt ideal zur deduktivistischen Methodologie. Allerdings ist eine Kontrolle des ß-Fehlers vor der Durchführung eines Experiments nur unter problematischen Zusatzannahmen und nur mit sehr großem Rechenaufwand möglich (vgl. Gabriel & Hsu, 1983). Da jedoch häufig die

Stichprobenverteilung der Statistik eines Randomisierungstests mit der eines parametrischen Tests approximativ übereinstimmt, hat Bredenkamp (1980) letzteren als rechnerische Vereinfachung eines Randomisierungstests interpretiert. Diese Interpretation, die auf empirischen Robustheitsstudien beruht (Bredenkamp, 1980; Willmes, 1987), verbindet die Vorteile eines Randomisierungstests mit denen der Neyman-Pearson-Theorie für parametrische Tests. Will man sich darauf nicht verlassen, kann man auch tatsächlich Randomisierungstests durchführen und den Stichprobenumfang nach Erwägungen festlegen, die in Abschnitt 3.4 behandelt werden.

## 1.4 Das Theorie-Empirie-Überbrückungsproblem

In PHn werden universelle theoretische Begriffe wie ,,imaginale Verarbeitung", „Gedächtnisleistung" , ,,Empfindungsstärke" usw. verwendet. Dies ist notwendig, wenn dem Erklärungsanspruch Rechnung getragen werden soll, da Gesetzeshypothesen in ihrem Geltungsanspruch über raumzeitliche Beschränkungen hinausgehen müssen. Obwohl die Grenze zwischen Beobachtungs- und Theoriensprache nicht scharf ist, ist die Unterscheidung theoretischer und empirischer Begriffe nützlich in dem Sinne, daß bestimmte Probleme der Operationalisierung angesprochen werden können (vgl. auch Gadenne, Kapitel 7 dieses Bandes). Wir gehen hier nur auf eines dieser Probleme ein, nämlich inwieweit die empirischen Variablen die in der psychologischen Hypothese benannten Vorgänge repräsentieren. U. E. ist das die schwierigste Frage überhaupt, die im hypothetisch-deduktiven Forschungsprozeß auftritt. Die Operationalisierung setzt bestimmte Hilfshypothesen voraus, damit aus Theorie, Anfangsbedingungen und Hilfsannahmen eine Prognose abgeleitet werden kann, die empirisch zu prüfen ist. Oftmals dürfte unklar sein, welche Hilfshypothesen zu präferieren sind. Ob wir z.B. über Instruktionen oder über die Auswahl bestimmter Lernmaterialien imaginale Verarbeitungsprozesse besser induzieren, ist ungewiß. Für diesen Fall wurde von einigen Autoren die ***konzeptuelle Replikation*** vorgeschlagen. Eine theoretische Variable konzeptuell zu replizieren heißt, sie unterschiedlich zu operationalisieren. Dies kann in ein- und demselben Versuch (intraexperimentelle Replikation) - es liegen dann multifaktorielle Versuche vor, wenn sich die Replikation auf die ***W*** bezieht, oder multivariate Versuche, wenn sie sich auf die AV bezieht - oder in verschiedenen Versuchen geschehen (interexperimentelle Replikation). Wenn in einem Versuch mehrere Indikatoren für die interessierenden Variablen vorliegen, kann man sie als fehlerbelastete Indikatorvariablen von latenten Variablen ansehen und entsprechende statistische Analysen durchführen (vgl. Abschnitt 4). Allerdings sind dann auch im experimentellen Kontext Randomisierungstests nicht mehr durchführbar, und der Einsatz von statistischen

Verfahren der Modellgeltungsprüfung ist neu zu rechtfertigen (vgl. dazu Abschnitt 4.4).

Bei interexperimenteller Replikation der interessierenden theoretischen Größe(n) kann der Bewährungsgrad der zu testenden PH unter Zuhilfenahme des Bayes--schen Theorems quantifiziert werden. Wenn im einfachsten Fall über alle Untersuchungen a und b konstant sind und vor dem Forschungsprogramm HO und Hl für gleichermaßen wahrscheinlich gehalten wurden, gilt nach t Untersuchungen, daß $p_t(H_0;\text{Daten}) = Q/(Q+1)$, wobei Q = as A(l-a)t-s/(l-b)s Ã bt-s und s die Anzahl signifikanter Resultate ist. Ist $Q > 1$ und damit $p_t(H_0;\text{Daten}) > .5$, bekräftigt die Evidenz eher $H_0$ als H,; bei $Q < 1$ ist es umgekehrt. Was hieraus für die inhaltliche Hypothese folgt, hängt davon ab, ob $H_0$ oder $H_1$ aus ihr abgeleitet wurde. Das Verfahren funktioniert allerdings nur unter der Annahme, daß die Apriori-Wahrscheinlichkeit für alle Parameter in der Indifferenzzone ($H'_1$) Null ist. Ein Wert $Q > 1$ muß deshalb nicht für $H_0$ sprechen, sondern kann auch für $H'_1$ sprechen, die von einer geringeren Effektstärke als $H_1$ ausgeht. Dies kann ohne weitere Datenerhebung untersucht werden, indem für die neu festgelegte Effektstärke das ß in jeder Untersuchung berechnet und Q neu gebildet wird. Ist Q nunmehr kleiner als 1, wird $H'_1$ bekräftigt, was darauf hinweist, daß vor dem Forschungsprogramm die Effektstärke zu groß angesetzt wurde.

## 1.5 Die Bedeutung der Situationsvalidität

Der Geltungsanspruch für PHn wird nie auf bestimmte Versuchsleiter, Versuchsräume usw. relativiert. Hat man Vermutungen darüber, daß derartige Gültigkeitseinschränkungen bestehen oder daß eine Hypothese nur für Personen mit bestimmten Merkmalen gilt, muß dies über den Test sog. disordinaler Wechselwirkungen geprüft werden. Eine Wechselwirkung zweier Wn A und B heißt disordinal, wenn a) die Rangordnung der Erwartungswerte über den B-Faktor sich zwischen den Stufen des A-Faktors unterscheidet, b) sich die Rangordnung der Erwartungswerte über den A-Faktor zwischen den Stufen des B-Faktors unterscheidet oder c) beides der Fall ist (vgl. z.B. Bredenkamp, 1980)[3]. Der Nachweis, daß eine Wechselwirkung disordinal ist, ist nicht einfach zu führen. Verfahren zur Bestimmung der Disordinalität sind Bestandteile einer deduktivistischen Methodologie (vgl. dazu Bredenkamp, 1982; Hager & Westermann, 1983c; Hager, 1987, 1992).

3 Bredenkamp (1980) differenziert darüber hinaus zwischen semidisordinalen Wechselwirkungen (a, b) und vollständig disordinalen Wechselwirkungen (c).

## 2. Eine formale Analyse der statistischen Prüfung psychologischer Hypothesen

Die in Abschnitt 1.1 geschilderten Probleme einer „direkten" empirischen Beurteilung von PHn anhand von Beobachtungen am einzelnen Individuum sind u. E. ein wesentlicher Grund dafür, daß die zuerst in den Agrarwissenschaften populär gewordene Idee statistischer Hypothesentests in der Psychologie begeistert aufgegriffen und angewandt wurde. Kann jedoch auf diese Weise die in der Einleitung dieses Kapitels geforderte „strenge und zugleich faire" Prüfung von PHn erreicht werden? Um diese Frage beantworten zu können, muß zunächst geklärt werden, wie eine „möglichst strenge und zugleich faire" ***statistische*** Prüfung von PHn zu explizieren ist. Im Anschluß daran ist zu untersuchen, welche methodologischen Konsequenzen eine geeignete Explikation nach sich zieht.

### 2.1 Was ist eine strenge und faire statistische Prüfung psychologischer Hypothesen?

Westermann (1987, S. 37f.) und Hager (1987, 1992) haben im Anschluß an Arbeiten von Hager und Westermann (1983a, 1983b) sowie Westermann und Hager (1986) ein auf Gadenne (1976) zurückgehendes Strengekriterium dadurch konkretisiert, daß sie zunächst zwei Fehlerwahrscheinlichkeiten $e_U$, und $f_U$ eingeführt haben. Dabei ist $e_U$ als die Wahrscheinlichkeit eines ***fälschlichen Bewährungsurteils*** und $f_U$ als die Wahrscheinlichkeit eines ***fälschlichen Nichtbewährungsurteils bzgl. einer zu prüfenden PH*** in einer konkreten empirischen Untersuchung U definiert:

$$e_U := p(\text{PH gilt in U als bewährt; PH trifft in U nicht zu}) \quad (2.1)$$

$$f_U := p(\text{PH gilt in U als nichtbewährt; PH trifft in U zu}). \quad (2.2)$$

Das Zeichen ";" ist hier als „unter der Annahme der Gültigkeit von" zu lesen; bedingte Wahrscheinlichkeiten werden demgegenüber durch das Zeichen „|" ausgedrückt. Man kann nun den ***Grad der Strenge der Prüfung von PH in*** U gleichsetzen mit der Wahrscheinlichkeit l-e, für korrekte Nichtbewährungsurteile und analog die ***Fairneß der Prüfung*** von ***PH in*** U mit der Wahrscheinlichkeit $1\text{-}f_U$für korrekte Bewährungsurteile. Es ist leicht zu erkennen, daß eine isolierte Forderung nach „maximaler Strenge" methodologisch nicht sinnvoll ist. Ihr könnte ja ganz einfach dadurch entsprochen werden, daß man unabhängig von der Befundlage ausschließlich Nichtbewährungsurteile bzgl. PH abgibt. Dies aber hieße, daß keinerlei Information über den Wahrheitswert von PH vermittelt wird. Westermann (1987) und Hager (1987) ersetzen deshalb die Forderung nach ***Strenge*** durch die Forderung nach ***Validität.*** Validität

ist für sie der ***Durchschnitt von Strenge und Fairneß.*** Diese Definition hat den Nachteil, daß mangelnde Strenge durch erhöhte Fairneß und umgekehrt mangelnde Fairneß durch erhöhte Strenge kompensierbar ist. Will man sicherstellen, daß eine Untersuchung nur dann valide genannt wird, wenn sie sowohl streng als auch fair ist, empfiehlt sich ein komparativer Validitätsbegriff der folgenden Art (vgl. auch Hager & Westermann, 1983b, S.69/70; eine graphische Veranschaulichung liefert Abbildung 1):

> ***Definition*** 1. Seien U und V zwei empirische Untersuchungen zur Überprüfung von PH; dann erfolgt in U genau dann eine validere Prüfung von PH als in V (kurz: U ist ***valider als*** V), wenn $e_U < e_V$ oder $f_U < f_V$, aber weder $e_U > e_V$ noch $f_U > f_V$.

In Worten ausgedrückt heißt dies, daß mindestens eine der beiden Fehlerwahrscheinlichkeiten in U kleiner und keine größer als in V ist.

> ***Definition*** 2. U und V können ferner ***gleich valide*** (zur Überprüfung von PH) genannt werden, wenn $e_U = e_V$ sowie $f_U = f_V$ gilt.

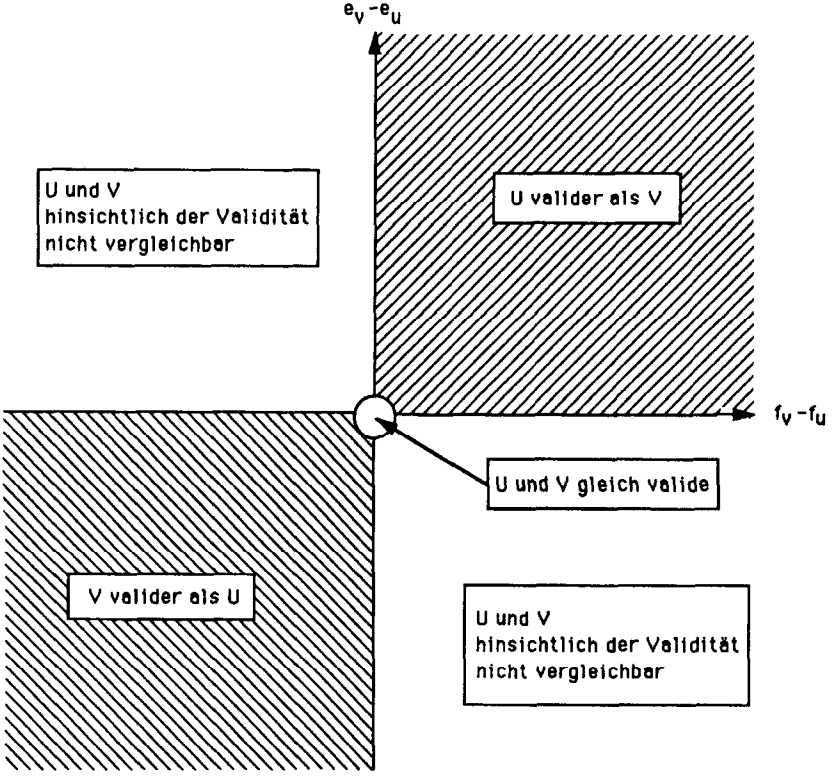


Abb. 1: Graphische Veranschaulichung verschiedener komparativer Validitätsbegriffe

Der Nachteil dieser Definitionsvorschläge, daß zwei Untersuchungen hinsichtlich ***der Validität nicht*** vergleichbar sein können (wenn U nicht valider als V, V nicht valider als U sowie U und V nicht gleich valide sind), fällt gegenüber dem o.g. Vorteil u.E. nicht ins Gewicht.

Wir werden im folgenden die Forderung nach möglichst strengen und zugleich fairen Hypothesenprüfungen mit der Forderung nach Validität im soeben definierten Sinne gleichsetzen (vgl. auch Hager & Westermann, 1983a, 1983b; Erdfelder, 1993a, 1993b). Im Kontext ***statistischer*** Prüfungen von PHn ist nun allerdings die empirische Untersuchung U noch genauer zu spezifizieren. Entscheidend ist hier vor allem das Konzept der ***statistischen Hypothese,*** die in noch näher zu spezifizierender Weise aus der PH gewonnen werden muß. Hiermit ist nicht nur eine einzelne Parameterhypothese (z. B. Mittelwertshypothese) gemeint, sondern eine Parameterhypothese oder eine logische Verknüpfung verschiedener Parameterhypothesen in einem bestimmten ***empirisch interpretierten*** statistischen Modell. Wenn eine bestimmte statistische Hypothese $SH_U$ spezifiziert wird, ist also immer auch eine konkrete Untersuchungssituation einschließlich der unabhängigen und abhängigen (Zufalls-)Variablen, des Versuchsplans und des Versuchsablaufs damit verbunden. Darüber hinaus ist mit der $SH_U$ ein ***Modell*** festgelegt (oder eine statistische Oberhypothese sensu Stegmüller, 1973), auf dessen Hintergrund die Daten statistisch analysiert werden. Hierbei wird es sich häufig um einen Spezialfall des Allgemeinen Linearen Modells oder des Log-Linearen Modells handeln; prinzipiell sind aber auch andere statistische Rahmenmodelle als Oberhypothesen denkbar.

Zusätzlich zur $SH_U$ ist eine ***Entscheidungsstrategie*** $ES_U$ festzulegen, die angibt, welche Klasse denkbarer empirischer Daten in U zu einer Annahme und welche zu einer Ablehnung der $SH_U$ führt[4]. Das Paar ($SH_U$, $ES_U$) bildet dann die empirisch-statistische Untersuchung, mittels der die PH beurteilt werden soll. Die Beurteilung geschieht in der Weise, daß eine Annahme der $SH_U$ zu einem Bewährungsurteil und eine Ablehnung der $SH_U$ zu einem Nichtbewährungsurteil bzgl. PH führt. Die statistische Entscheidung über $SH_U$ wird also gewissermaßen auf die PH übertragen.

Die Forderung nach validen Untersuchungen kann nun wie folgt konkretisiert werden:

> ***Regel $R_1$:*** Zu einer gegebenen PH wähle $SH_U$ und $ES_U$ derart, daß ($SH_U$, $ES_U$) möglichst valide ist.

4 Die Entscheidungsstrategie entspricht dem, was man gemeinhin unter einem „statistischen Test" versteht. Sie umfaßt die Festlegung einer Teststatistik T, einer Entscheidungsregel d(T = t) und einer Menge von Entscheidungsalternativen, die hier als A = {$H_0$, $H_1$} vorgegeben ist (vgl. dazu Ostmann & Wutke, Kapitel 16 dieses Bandes, Abschnitt 2).

## 2.2 Determinanten von Strenge und Fairneß

Die Umsetzung der methodologischen Regel $R_1$ verlangt eine Wahl von $SH_U$ und $ES_U$ in der Weise, daß die Wahrscheinlichkeiten fälschlicher Bewährungs- und fälschlicher Nichtbewährungsurteile bzgl. PH möglichst klein sind. Doch wie lassen sich diese Fehlbeurteilungswahrscheinlichkeiten reduzieren oder u. U. sogar kontrollieren, so daß sie bestimmte (kleine) Werte nicht überschreiten? Erdfelder (1993a) hat gezeigt, daß $e_U$ und $f_U$ von vier Parametern abhängen, die wir in dieser Arbeit wie folgt einführen wollen:

$$\varepsilon := p(\text{Annahme von } SH_U \mid SH_U \text{ trifft nicht zu}) \tag{2.3}$$

$$\phi := p(\text{Ablehnung von } SH_U \mid SH_U \text{ trifft zu}) \tag{2.4}$$

$$g := p(SH_U \text{ trifft zu; PH trifft in U nicht zu}) \tag{2.5}$$

$$h := p(SH_U \text{ trifft nicht zU; PH trifft in U zu}) \tag{2.6}$$

Die beiden zuerst aufgeführten Fehlerwahrscheinlichkeiten, die wir dem Vorschlag Hagers (1987, 1992; siehe auch Hager & Westermann, 1983a, 1983b) folgend mit $\varepsilon$ und $\phi$ notieren, sind ***statistische Irrtumswahrscheinlichkeiten,*** **d. h. Wahrscheinlichkeiten für fälschliche Annahmen und Ablehnungen von** $SH_U$ bei Anwendung der festgelegten Entscheidungsstrategie $ES_U$. Ist z.B. $SH_U$ eine statistische Alternativhypothese ($H_1$), so gilt $\varepsilon = \alpha$ und $\phi = ß$, wobei $\alpha$ und ß die bekannten statistischen Irrtumswahrscheinlichkeiten erster bzw. zweiter Art sind. Umgekehrt gilt $\varepsilon = ß$ und $\phi = \alpha$, wenn $S_U$, eine $H_0$ ist. **Ist** $SH_U$ allgemein eine bestimmte Konjunktion, Disjunktion oder eine andere Verknüpfung verschiedener Null- und/oder Alternativhypothesen, so sind $\varepsilon$ und $\phi$ mehr oder minder komplizierte Funktionen der Irrtumswahrscheinlichkeiten der durchgeführten Einzeltests. Die Kontrolle von $\varepsilon$ und $\phi$ bereitet - wie sich zeigen läßt - vielleicht praktische, aber jedenfalls keine prinzipiellen Probleme (vgl. bereits Hager & Westermann, 1983a, Westermann & Hager, 1986). Wir werden auf diesen Punkt in Abschnitt 3.3 zurückkommen.

In diesem Abschnitt werden wir uns vor allem den in den Gleichungen (2.5) und (2.6) eingeführten Wahrscheinlichkeiten g und h widmen. Beide Wahrscheinlichkeiten betreffen das Verhältnis von PH und daraus gewonnener $SH_U$. Sie können am einfachsten als Wahrscheinlichkeiten für Wahrheitswertdiskrepanzen zwischen PH und $SH_U$ charakterisiert werden: g ist die Wahrscheinlichkeit einer wahren $SH_U$ im Falle einer ungültigen PH, h umgekehrt die Wahrscheinlichkeit einer falschen $SH_U$ im Falle einer zutreffenden PH (Erdfelder, 1993a).

Wie hängen nun $e_U$ und $f_U$ von den o.g. Parametern ab? Es muß lediglich untersucht werden, auf welchen Wegen man zu Fehlbeurteilungen der zu testenden PH gelangen kann (vgl. im Detail Erdfelder, 1993a). Betrachten wir

zunächst die Fehlerwahrscheinlichkeit $e_U$. Wenn PH in U nicht zutrifft, kommt es zu einer Fehlbeurteilung genau dann, wenn $SH_U$ angenommen wird. Dies ist auf zwei Wegen möglich[5]: (a) $SH_U$ trifft zu und wird korrekt angenommen und (b) $SH_U$ trifft nicht zu, wird aber dennoch fälschlich angenommen. Die Wahrscheinlichkeit für Weg (a) ist gleich dem Produkt der Wahrscheinlichkeiten g und $(1-\phi)$, die für Weg (b) gleich dem Produkt der entsprechenden Wahrscheinlichkeiten (I-g) und $\varepsilon$. Da die Wege disjunkt sind, addieren sich die Wahrscheinlichkeiten beider Wege zu $e_U$ auf:

$$\begin{aligned} e_U &= p(\text{Annahme von } SH_U;\ PH \text{ trifft in U nicht zu}) \\ &= g \cdot (1-\phi) + (1-g) \cdot \varepsilon \\ &= \varepsilon + (1-\varepsilon-\phi) \cdot g. \end{aligned} \tag{2.7}$$

Wir sehen, daß $e_U$ für gegebene statistische Irrtumswahrscheinlichkeiten $\varepsilon$ und $\phi$ eine lineare Funktion von g mit Ordinatenabschnitt $\varepsilon$ und Steigung $1-\varepsilon-\phi$ ist. Da diese Steigung niemals negativ werden kann (vgl. Erdfelder, 1993b), ist $\varepsilon$ zugleich der kleinste Wert, den $e_U$ annehmen kann. Mit Wachsendern g wächst auch $e_U$. Falls g und $\phi$ gegeben sind und g kleiner als 1 ist, können wir zusätzlich sagen, daß $e_U$ mit $\varepsilon$ wächst.

Eine analoge Analyse läßt sich für $f_U$ durchführen. Auch hier können zwei Wege zu einer falschen Beurteilung der PH führen, die in diesem Fall aus der Ablehnung von $SH_U$ resultiert. Weg (a) führt über eine gültige $SH_U$ zu einer fälschlichen Verwerfung von $SH_U$ und hat die Wahrscheinlichkeit (1-h) . $\phi$; Weg (b) mit Wahrscheinlichkeit h . $(1-\varepsilon)$ führt über eine ungültige $SH_U$ zu einer korrekten Zurückweisung. Analog zu (2.7) ergibt sich dann:

$$\begin{aligned} f_U &= p(\text{Verwerfung von } SH_U;\ PH \text{ trifft in U zu}) \\ &= (1-h) \cdot \phi + h \cdot (1-E) \\ &= \phi + (1-\phi-\varepsilon) \cdot h. \end{aligned} \tag{2.8}$$

Für gegebene statistische Irrtumswahrscheinlichkeiten ist (2.8) eine linear steigende Funktion von h mit dem kleinstmöglichen Wert $\phi$. Falls $h < 1$ ist, wächst $f_U$ unter sonst gleichen Bedingungen mit $\phi$ und fällt mit $\varepsilon$.

5 Um die Darstellung zu vereinfachen, wird hier angenommen, daß $H_0$ und $H_1$ den Parameterraum ausschöpfen. Die folgenden Formeln (2.7) und (2.8) sind zu erweitern, wenn diese Annahme fallengelassen wird. Die methodologischen Folgerungen in Abschnitt 2.3 werden hierdurch nicht wesentlich tangiert.

## 2.3 Methodologische Folgerungen aus der Forderung nach Strenge und Fairneß

Hager und Westermann (erstmals 1983a, 1983b) haben behauptet, daß $e_U$ und $f_U$ durch die statistischen Irrtumswahrscheinlichkeiten $\varepsilon$ und $\phi$ beeinflußt werden können. Diese Behauptung ist richtig, wie die Gleichungen (2.7) und (2.8) zeigen. Zugleich wird aber deutlich, daß die Art der Abhängigkeit zwischen statistischen und PH-Fehlbeurteilungswahrscheinlichkeiten keineswegs eine einfache ist. Sehen wir einmal von den Grenzfällen g = 1 und h = 1 ab, trifft es zwar zu, daß - ceteris paribus - eine Herabsetzung von $\varepsilon$ eine Herabsetzung von $e_U$ und eine Reduzierung von $\phi$ eine Reduzierung von $f_U$ bewirkt (vgl. z.B. Hager, 1987, 1992 und Hussy & Möller, Kapitel 11 dieses Bandes). Daraus kann aber nicht ohne weiteres geschlossen werden, daß die Reduzierung der statistischen Irrtumswahrscheinlichkeiten eine Erhöhung der Validität zur Folge hat, denn eine Herabsetzung von $\varepsilon$ führt leider i.a. auch zu einer Erhöhung von $f_U$ ebenso wie eine Reduzierung von $\phi$ eine Erhöhung von $e_U$ zur Folge hat. Wir können vorerst festhalten:

> ***Beobachtung 1.*** Seien U und V zwei Untersuchungen mit g > 0 und h > 0, die sich nur darin unterscheiden, daß genau eine der beiden statistischen Irrtumswahrscheinlichkeiten in U kleiner gewählt wurde. Dann sind U und V hinsichtlich der Validität nicht vergleichbar. Insbesondere ist ausgeschlossen, daß U valider als V ist.

Angenommen, wir streben eine Erhöhung der Validität durch Reduzierung beider Irrtumswahrscheinlichkeiten $\varepsilon$ und $\phi$ an. Unter welchen Bedingungen ist dies möglich? Wenn wir $\varepsilon$ um den Betrag $k_\varepsilon$ ($0 < k_\varepsilon < E$) verkleinern, fuhrt dies einerseits zu einer Reduzierung von $e_U$, andererseits aber auch zu einem Anwachsen von $f_U$. Dieser unerwünschte Anstieg muß durch Reduzierung der zweiten statistischen Irrtumswahrscheinlichkeit $\phi$ kompensiert werden, Man kann zeigen (Erdfelder, 1993b), daß die Verkleinerung von $\phi$ um den Betrag $k, = h / (1-h) \cdot k_\varepsilon$ die gerade notwendige Kompensation leistet, d.h. ein Anwachsen von $f_U$ verhindert. Natürlich führt diese $\phi$-Änderung wiederum zu einem unerwünschten Anstieg von $e_U$, der nur dann kleiner als die durch $\varepsilon$-Änderung erzielte Verbesserung ist, falls g + h < 1 (Erdfelder, 1993b). Das gleiche Resultat ergibt sich, wenn man eine Validitätsverbesserung zunächst durch Reduzierung von $f_U$ über $\phi$ versucht. Wir können demnach konstatieren:

> ***Beobachtung*** 2. Seien U und V zwei Untersuchungen, die sich lediglich darin unterscheiden, daß $\varepsilon$ und $\phi$ in U kleiner als in V gewählt wurden. Dann kann U nur dann valider als V sein, falls g + h < 1.

Die Forderung nach Reduktion der statistischen Irrtumswahrscheinlichkeiten ist also lediglich bei Gültigkeit einer Randbedingung aus der Forderung nach Validität ableitbar. Ehe wir kleinere $\varepsilon$- und $\phi$-Werte fordern, muß sichergestellt sein, daß die Wahrscheinlichkeiten g und h für Wahrheitswertdiskrepanzen zwischen PH und $SH_U$ zusammen nicht größer als 1 sind.

Ein naheliegender Gedanke wäre es, mit folgender Forderung zu beginnen:

> ***Regel $R_2$:*** Zu einer gegebenen PH wähle $SH_U$ so, daß die Wahrscheinlichkeiten g und h den Wert 0 annehmen.

Diese Forderung folgt aus Regel R,. Die oben erwähnten Eigenschaften der Gleichungen (2.7) und (2.8) garantieren, daß von zwei Untersuchungen, die sich lediglich hinsichtlich der Wahrscheinlichkeiten g oder h unterscheiden, grundsätzlich diejenige valider sein wird, die Regel $R_2$ genügt. Aber wie kann gezeigt werden, daß eine Untersuchung Regel $R_2$ genügt? Dies kann nur über einen Äquivalenzbeweis geschehen, der Wahrheitswertdiskrepanzen zwischen PH und $SH_U$ prinzipiell ausschließt. Praktisch bedeutet die Umsetzung der Regel $R_2$ daher einerseits eine stochastische Formulierung der PH und andererseits eine Spezifikation von $SH_U$ derart, daß $SH_U$ zugleich notwendig und hinreichend für die Gültigkeit von PH in U ist. Der Äquivalenzbeweis ist im Kalkül der Stochastik zu erbringen.

Diese Forderungen erscheinen auf den ersten Blick ungewöhnlich streng, so daß es angezeigt ist, sich mit ihrer Realisierbarkeit zu beschäftigen. Zunächst muß festgehalten werden, daß es zur stochastischen Beweisführung keine Alternative gibt, wenn man sich einmal auf die methodologische Regel $R_1$ in Verbindung mit den Definitionen 1 und 2 festgelegt hat. Die Definition von Strenge, Fairneß und Validität in termini von Wahrscheinlichkeiten legt die formale Sprache fest, in der die Argumentation zu fuhren ist. Ist die formale Sprache festgelegt, so muß die PH in eben diese Sprache „übersetzt“[6] werden, sofern sie nicht von vornherein schon in dieser Sprache formuliert wurde. Andernfalls sind Behauptungen über Implikations- oder Äquivalenzbeziehungen zwischen PH und $SH_U$ weder beweisbar noch widerlegbar; wie Regel $R_2$ zeigt, können wir aber hierauf nicht verzichten.

Glücklicherweise ist mit der Festlegung auf die Stochastik ein äußerst flexibler Rahmen gegeben. In dieser Sprache lassen sich deterministische und nichtdeterministische Aussagen über qualitative (diskrete) und quantitative ZVn formulieren, die beobachtbar (manifest) oder nicht beobachtbar (latent) sein können (einführend z.B. Steyer, 1988, Kapitel 15 dieses Bandes). Tatsächlich sind

6 Es ist klar, daß ein solcher „übersetzungsvorgang“ im Regelfall auch eine Explikation der zumeist umgangssprachlich formulierten PH beinhalten muß. Dies ist auf die zumeist gegebene Vagheit und Mehrdeutigkeit der ursprünglichen Formulierung in einer nicht-formalen Sprache zurückzuführen und insofern nicht zu vermeiden.

PHn, die in diesem Rahmen nicht angemessen formulierbar sind, nur schwer vorstellbar. So überrascht es nicht, daß PHn, zu denen empirische Überprüfungen vorgeschlagen wurden, die Regel $R_2$ genügen, in der psychologischen Literatur einen immer breiteren Raum einnehmen. Ein prominentes und besonders typisches Beispiel bilden verschiedene probabilistische Modelle additiv-verbundener Messung (vgl. Falmagne, ***1979),*** die u.a. zur Überprüfung von PHn über binaurale Lautheitssummation in der Psychoakustik verwendet wurden (z.B. Falmagne, Iverson & Markovici, 1979).

Allerdings ist es nicht leicht, eine $SH_U$ zu finden, die Regel $R_2$ genügt. Vor allem der Nachweis der Implikation $SH_U$ ti PH ist normalerweise nicht trivial, so daß man u.U. behelfsweise an schwächere Formen der Überprüfung denken muß, als sie Regel $R_2$ verlangt. Auch andere Gründe können ausschlaggebend dafür sein, die Forderung nach Äquivalenz von PH und $SH_U$ fallenzulassen: Wie wir in Abschnitt 1.1 gesehen haben, ist oftmals eine Überprüfung von PHn auf Individuumsebene prinzipiell ausgeschlossen, so daß zwangsläufig auf die Aggregatebene übergegangen werden muß. Aggregataussagen können aber bestenfalls notwendige, niemals aber zugleich auch hinreichende Bedingungen für die Gültigkeit (allgemeinpsychologischer) PHn sein, die für die Individuumsebene Gültigkeit beanspruchen (vgl. Bredenkamp, ***1972, 1980).*** Will man den in der Psychologie sehr häufig anzutreffenden und oftmals gar nicht vermeidbaren übergang auf die Aggregatebene nicht von vornherein als sinnlos abtun, wird man sich der folgenden Abschwächung von Regel $R_2$ zuwenden müssen:

> ***Regel $R_{2'}$:*** Zu einer gegebenen PH wähle $SH_U$ so, daß die Wahrscheinlichkeit h den Wert 0 annimmt.

Praktisch wird mit $R_{2'}$ die Forderung nach Äquivalenz von PH und $SH_U$ auf die Forderung nach einer ***Implikationsbeziehung*** PH $\Rightarrow$ $SH_U$ reduziert. Diese Regel dürfte im Falle einer präzise formulierten, empirisch gehaltvollen PH wesentlich leichter als $R_2$ zu erfüllen sein.

Welche Konsequenzen hat die abgeschwächte Regel $R_{2'}$? Wenn die Wahrscheinlichkeit g keiner Normierung unterworfen wird, so heißt das nach Gleichung (2.7) zunächst, daß die Wahrscheinlichkeit eines fälschlichen Bewährungsurteils für die PH nicht kontrolliert werden kann. ***Nur Nichtbewährungsurteile können demnach bei kontrollierter Fehlerwahrscheinlichkeit $f_U = \phi$ erfolgen (vgl. Gleichung 2.8), nicht aber Bewährungsurteile.*** Dennoch ist Regel $R_{2'}$ streng genug, um eine Erhöhung der Validität über die Reduktion von statistischen Irrtumswahrscheinlichkeiten zu ermöglichen. Die Validität kann problemlos durch Verkleinerung von $\varepsilon$ erhöht werden, da diese Maßnahme $e_U$ senkt (falls $g < 1$) und $f_U$ nicht tangiert. Prinzipiell kann auch an eine Validitätserhöhung über Reduzierung von $\phi$ um den Betrag $k_\phi$ in Verbindung mit einem mindestens um den Betrag g /(l-g) . $k_\phi$ verringerten $\varepsilon$

gedacht werden. Diese Maßnahme würde $f_U$ senken und sicherstellen, daß $e_U$ zumindest nicht erhöht wird. Da g jedoch im hier interessierenden Fall unbekannt ist, nutzt die angegebene E-Adjustierungsformel wenig. Es empfiehlt sich deshalb, $\phi$ nicht zu klein zu wählen, da man ansonsten Gefahr läuft, zu viele fälschliche Bewährungsurteile zu erhalten.

Fassen wir zusammen. Aus dem Validitätsmaximierungspostulat $R_1$, von dem wir ausgegangen sind, folgt, daß wann ***immer möglich*** Regel $R_2$ angewendet werden sollte. In Verbindung mit $R_2$ folgt aus $R_1$ auch die methodologische Regel:

> ***Regel $R_3$: Zu*** einer PH wähle eine Untersuchung U mit einer Entscheidungsstrategie $ES_U$ so, daß die statistischen Irrtumswahrscheinlichkeiten $\varepsilon$ und $\phi$ kontrollierbar und möglichst klein sind.

Die Kombination von $R_2$ und $R_3$ stellt sicher, daß PH-Bewährungsurteile bei kontrollierter (kleiner) Fehlerwahrscheinlichkeit $e_U = \varepsilon$ und Nichtbewährungsurteile bei kontrollierter (kleiner) Fehlerwahrscheinlichkeit $f_U = \phi$ erfolgen.

Läßt sich dieses ideale Vorgehen nicht realisieren, so empfiehlt sich die Realisierung der abgeschwächten Regel $R_{2'}$ in Verbindung mit der methodologischen Regel:

> ***Regel $R_{3'}$: Zu*** einer PH wähle eine Untersuchung U mit einer Entscheidungsstrategie $ES_U$ so, daß $\phi$ und $\varepsilon$ kontrollierbar sind. Ferner sollte $\varepsilon$ möglichst klein gewählt werden.

Die Kombination von $R_{2'}$ und $R_{3'}$ stellt die Kontrolle der Wahrscheinlichkeit $f_U = \phi$ für fälschliche Nichtbewährungsurteile sicher. Die Wahrscheinlichkeit $e_U$ für fälschliche Bewährungsurteile kann mangels Normierung von g zwar nicht kontrolliert, aber u.U. (in unbekanntem Ausmaß) reduziert werden, wenn man die folgende Forderung ergänzt:

> ***Regel $R_{4'}$: Zu*** einer PH wähle $SH_U$ so, daß g möglichst klein (und U somit möglichst streng) ist.

Aus Regel $R_{4'}$ kann abgeleitet werden, welche von mehreren aus einer PH deduzierbaren SHn zur Überprüfung herangezogen werden sollte. Es gilt (Erdfelder, 1993b):

> ***Beobachtung 3.*** Seien ($SH_U$, $ES_U$) und ($SH_V$, $ES_V$) zwei Untersuchungen zur Prüfung einer PH, die beide der Regel $R_{2'}$ genügen. Es gelte ferner $SH_U \Rightarrow SH_U$, nicht aber $SH_U \Rightarrow SH_U$. Dann ist g in U kleiner als in V und - sofern $\varepsilon$ und $\phi$ sich zwischen beiden Untersuchungen nicht unterscheiden- U somit auch valider als V.

Man sollte also immer die „bestimmtere“ (empirisch gehaltvollere) SH als Überprüfungsinstanz heranziehen, sofern Wahlmöglichkeiten bestehen (vgl. auch Bredenkamp, 1984). Diese Maßnahme erhöht die Strenge einer Untersuchung, ohne die Fairneß zu senken.

Die Kombination der Regeln $R_{2'}$, $R_{3'}$ und $R_{4'}$ ist die zweitbeste Umsetzung der Regel $R_{1'}$ die denkbar ist. Andere Umsetzungen von $R_{1'}$ die nicht auf der Forderung basieren, daß g oder h Null sind, sind zwar formal konzipierbar, dürften aber mangels zwingender Begründungen für bestimmte Wahrscheinlichkeitswerte von g und h, die dann größer Null und kleiner Eins sein müßten, nicht realisierbar sein. Wie wir anhand von Beobachtung 2 sehen, ist aber nicht einmal die Forderung nach kleinen statistischen Irrtumswahrscheinlichkeiten $\varepsilon$ und $\phi$ gerechtfertigt, wenn die Möglichkeit besteht, daß g und h zusammen größer als 1 sind.

## 2.4 Randomisierte oder nichtrandomisierte Untersuchungen?

Die im letzten Abschnitt besprochenen Folgerungen aus der Forderung nach möglichst validen Untersuchungen beinhalten Antworten auf konkrete Probleme der psychologischen Versuchsplanung, die den Regeln nicht auf den ersten Blick anzusehen sind. Ein Beispiel hierfür ist die Frage, ob von der Kontrolltechnik der Randomisierung Gebrauch zu machen ist (vgl. Strack & Rehm, Kapitel 12 dieses Bandes). Als Begründung für die Notwendigkeit der Randomisierung wird häufig genannt, daß allein randomisierte Untersuchungen eine kausale Interpretation eines eventuellen Effekts der UV(n) auf die AV(n) ermöglichen. Diese Begründung ist natürlich genauso problematisch wie der Kausalitätsbegriff, auf den sie Bezug nimmt. Da der Kausalitätsbegriff wissenschaftstheoretisch nach wie vor als nicht geklärt betrachtet werden kann und insbesondere unklar ist, in welchem Sinne experimentellen Befunden kausale Interpretierbarkeit zugesprochen werden kann, hat Steyer (1992 und Kapitel 15 dieses Bandes) einige Kausalitätsbegriffe vorgeschlagen, die mit dem (randomisierten) Experiment sinnvoll in Verbindung zu bringen sind. Einer der Steyerschen Kausalitätsbegriffe - der Begriff der schwachen kausal-regressiven Abhängigkeit - ist erfüllt, wenn die interessierende UV und potentielle Störvariablen stochastisch unabhängig sind. Bekanntlich stellt aber das Randomisierungsprinzip genau diese Unabhängigkeit von UV(n) und Störvariablen sicher. Insofern kann man mit Steyer (1992) feststellen, daß Randomisierung immer dann angestrebt werden sollte, wenn eine kausale Interpretierbarkeit der Versuchsergebnisse erwünscht ist.

Man kann nun weiterfragen, ***warum*** eine kausale Interpretierbarkeit der Versuchsergebnisse erwünscht sein sollte. Auf dem in den vorstehenden Abschnitten dargelegten methodologischen Hintergrund läßt sich diese Frage wie folgt

beantworten: ***Randomisierung ist im Regelfall*** deshalb ***anzustreben, weil*** *für* ***viele PHn*** nur ***auf diese Weise sichergestellt werden kann, daß die Implikation PH $\Rightarrow$ $SH_U$ und damit h = 0 erfüllt ist. Hieraus folgt: Relativ zu einer nicht-randomisierten Untersuchung erhöht Randomisierung die Fairneß ($1-e_U$) und damit die Validität einer*** Untersuchung.

Zur Begründung wollen wir den einfachen, aber dennoch typischen Fall einer PH betrachten, die einen Effekt einer dichotomen UV X auf eine AV Y behauptet. Als Beispiel, das auch von Hager (1987) zur Illustration herangezogen wird, sei die PH „Konkrete Wörter (X = 1) werden besser behalten als abstrakte Wörter (X = 0)" herangezogen. Die AV Y entspricht in diesem Fall dem Gedächtnismaß (z. B. einem Reproduktions- oder Rekognitionsmaß), auf das sich die PH bezieht. Bei Licht besehen enthalten derartige PHn oft noch zwei Zusatzannahmen, die zumeist nicht explizit genannt, aber durchaus mitgedacht werden: (a) Neben dem X-Effekt gibt es möglicherweise einen additiven Effekt unbekannter Fehlervariablen mit Erwartungswert 0, so daß nicht Y, sondern der bedingte Erwartungswert von Y unter X linear von X abhängt; (b) die lineare Beziehung zwischen bedingtem Erwartungswert und X gilt ***ceteris*** paribus, d.h. bei Konstanthaltung anderer Störeinflüsse auf Y. Bei den Störgrößen kann es sich z.B. um Persönlichkeitseigenschaften der Vpn oder auch um situative Kontextbedingungen handeln, die neben der interessierenden ***W*** Einfluß auf die interessierende AV der Untersuchung nehmen.

Bezeichnen wir die (u.U. mehrdimensionale) unbekannte Störgröße mit W und ihre möglichen Realisationen mit w, so lautet eine angemessene Formulierung der skizzierten PH:

Für alle Werte ***W = w:***

$$E(Y \mid X,\ W = w) = a_w + b_w \cdot X, \text{ wobei } b_w > 0. \tag{2.9}$$

Für nicht konstantgehaltene Werte der Störvariablen gilt somit

$$E(Y \mid X,\ W) = f(W) + g(W) \cdot X, \tag{2.10}$$

wobei g(W) positiv reellwertig sein muß. Letztlich geprüft - z.B. via t-Test oder Mann-Whitney-Test - wird aber die $SH_U$

$$E(Y \mid X) = a + b \cdot X, \text{ wobei } b > 0. \tag{2.11}$$

Ist dieses Vorgehen gerechtfertigt? Natürlich kann die PH gemäß Gleichung (2.9) nicht aus der $SH_U$ gemäß Gleichung (2.11) deduziert werden und deshalb können PH und $SH_U$ auch nicht äquivalent sein. Der bestmögliche Weg der Überprüfung der PH (mit g = h = 0) ist somit nicht gegeben. Ist aber wenigstens die Implikation PH $\Rightarrow$ $SH_U$ erfüllt, so daß h = 0 gilt? Ohne weitere Zusatzannahmen folgt Gleichung (2.11) nicht aus Gleichung (2.9). Wenn die

Störvariable W nicht konstantgehalten wird und X und ***W*** abhängig sind, kann bekanntlich trotz Gültigkeit von ***(2.9)*** ein negativer Effekt $b < 0$ in Gleichung (2.11) auftreten, der fälschlich die Ungültigkeit der PH nahelegt (vgl. Steyer, 1992, Kapitel 15 dieses Bandes). Dies muß man ausschließen, wenn man von der PH auf die $SH_U$ übergehen will. Ist ***W*** bekannt und vollständig kontrollierbar, so kann dies durch die experimentellen Kontrolltechniken der ***Elimination, Konstanthaltung*** und der ***systematischen Variation*** der Störeinflüsse oder auch über ***Parallelisierung*** erfolgen (vgl. Strack & Rehm, Kapitel 12 dieses Bandes). Ist ***W*** nicht (vollständig) bekannt oder nicht (vollständig) kontrollierbar, so bleibt als ***einzig möglicher Weg*** die ***Randomisierung,*** welche die stochastische Unabhängigkeit (stU) von X und ***W*** garantiert. Man kann dann aus Gleichung (2.10) herleiten:

$$
\begin{aligned}
E(Y|X) &= E((E(Y|X, W) + F)|X) \qquad (\text{da } F := Y - E(Y|X, W)) \\
&= E(E(Y|X, W)|X) + E(F|X) \\
&= E(E(Y|X, W)|X) + 0 \qquad (2.12) \\
&= E((f(W) + g(W) \cdot X)|X) \quad (\text{nach Gleichung 2.10}) \\
&= E(f(W)|X) + E(g(W)|X) \cdot X \\
&= a + b \cdot X,\ b > 0 \qquad (\text{aufgrund stU von } X \text{ und } W).
\end{aligned}
$$

Ähnliche Argumentationen lassen sich auch für andere PHn führen, die ceteris-paribus-Klauseln beinhalten. Aus Platzgründen wollen wir dies jedoch nicht weiter ausführen.

Das Fazit dieses Abschnitts lautet, daß eine Verwendung der Randomisierung bei Überprüfungen von PHn geboten ist, die ceteris-paribus-Klauseln bezüglich unbekannter oder nicht vollständig kontrollierbarer Störeinflüsse beinhalten. In diesem Fall ermöglicht die Randomisierung die Deduktion prüfbarer statistischer Hypothesen, welche die Störgrößen nicht mehr beinhalten. Dies ***ist der eigentliche Zweck der Randomisierung.***

## 2.5 Sind statistische Aggregathypothesen zulässig?

Sehr viele PHn - z.B. die im letzten Abschnitt diskutierte Hypothese über die Abhängigkeit der Gedächtnisleistung von der Konkretheit des Wortmaterials - sind allgemeinpsychologischer Natur und behaupten demzufolge etwas über (beliebige) ***einzelne*** Individuen eines offenen Individuenbereichs (vgl. Abschnitt 1). Überprüft werden diese Hypothesen aber sehr häufig anhand von SHn, die sich auf ***über Individuen aggregierte Daten*** beziehen. Die Frage, ob diese oftmals gar nicht vermeidbare Vorgehensweise methodologisch vertretbar ist, läßt sich in ähnlicher Weise wie die Frage nach dem Zweck der Randomisierung beantworten. Wie schon erwähnt, kann beim übergang ***von***

Individuen auf Aggregate bestenfalls eine notwendige, nicht aber auch eine hinreichende Bedingung für die Gültigkeit der vorgeordneten PH resultieren. Da wir andere Wege der statistischen Prüfung einer PH oben ausgeschlossen haben, bleibt nur eine Möglichkeit, die Überprüfung allgemeinpsychologischer Hypothesen über statistische Aggregathypothesen zu rechtfertigen: ***Die statistische Aggregathypothese muß aus der allgemeinpsychologischen Hypothese logisch folgen*** (vgl. bereits Estes, 1956).

Im Regelfall ist die Frage, ob eine Aggregathypothese aus einer allgemeinpsychologischen Hypothese folgt, nicht so leicht zu beantworten, wie es auf den ersten Blick vielleicht den Anschein hat. Im letzten Abschnitt haben wir beispielsweise gesehen, daß aus der Hypothese „Für alle Individuen gilt, daß konkrete Wörter ***ceteris paribus*** besser als abstrakte Wörter behalten werden" keineswegs ohne weiteres folgt „Im Mittel (über Individuen) werden konkrete Wörter besser als abstrakte Wörter behalten". Die Implikation gilt nur, wenn randomisiert wurde. Ebenso wie in diesem Beispiel muß jede PH daraufhin analysiert werden, welche Aggregathypothesen unter welchen Zusatzannahmen daraus ableitbar sind und welche nicht.

Bei PHn, die quantitative (durch eine mathematische Funktion beschreibbare) Gesetzmäßigkeiten zwischen einer ***W*** X und einer AV Y behaupten, ist besondere Vorsicht geboten. Es gilt im allgemeinen nicht, daß die über Individuen gemittelten Daten demselben Funktionstyp wie die individuellen Funktionen folgen. Z. B. ist das Potenzgesetz für interindividuell gemittelte Daten nicht erfüllt, wenn es für einzelne Individuen mit interindividuell variablem Exponenten gilt. Andere Beispiele dafür, daß ein bestimmter Funktionstyp für individuumsbezogene Daten nicht auch für die interindividuell gemittelten Daten gelten muß, präsentieren u. a. Bakan (1954) und Estes (1956). Die gleichen Autoren haben auch hinreichende Bedingungen dafür angegeben, daß ein bestimmter Funktionstyp invariant gegenüber arithmetischer Mittelwertsbildung ist. Bei der Beantwortung der Frage, ob eine konkrete Aggregathypothese aus einer allgemeinpsychologischen Gesetzeshypothese ableitbar ist, können diese Kriterien wertvolle Dienste leisten.

## 2.6 Skalenniveau und Statistik

Seit Ende der vierziger Jahre wird in der einschlägigen psychologischen Methodenliteratur - insbesondere in der Zeitschrift ***Psychological Bulletin*** - eine Debatte darüber ausgetragen, ob das Skalenniveau psychologischer Variablen die Wahl der (Prüf-) Statistik mitdeterminieren sollte oder nicht. Von bedingungsloser Zustimmung bis hin zu bedingungsloser Ablehnung wurden nahezu alle denkbaren Standpunkte zu dieser Frage vertreten. Eine Zusammenfassung der frühen Debatte findet man bei Bredenkamp (1972); neuere Ar-

beiten sind in Luce, Krantz, Suppes und Tversky (1990) zitiert und knapp zusammengefaßt.

Unsere Auffassung, die weitgehend derjenigen von Luce et al. (1990, S.299) entspricht, kann wie folgt zusammengefaßt werden: ***Bei der Wahl der Prüf-Statistik zu einer vorgegebenen $SH_U$ spielt das Skalenniveau keine Rolle, bei der Wahl einer $SH_U$ zu einer vorgegebenen PH aber eine ganz entscheidende.*** Diese Position ergibt sich zwingend aus den Konsequenzen des Validitätsmaximierungspostulats, die wir am Ende von Abschnitt 2.3 zusammenfassend dargestellt haben. Nehmen wir zunächst an, eine aus der PH ableitbare $SH_U$ sei bereits gegeben. Dann kommt es lediglich darauf an, ein Prüfverfahren festzulegen, das eine Kontrolle von $\varepsilon$ und $\phi$ ermöglicht und dabei möglichst effizient ist, d.h. möglichst kleine Irrtumswahrscheinlichkeiten bei möglichst kleinem Stichprobenumfang sicherstellt. Hieraus folgt in keiner Weise, daß die Prüfstatistik bzw. der Wahrheitswert von Aussagen über numerische Werte der Prüfstatistik invariant gegenüber zulässigen Transformationen von UVn und AVn sein muß. Also kann zu einer vorgegebenen $SH_U$ eine Prüfstatistik völlig unabhängig vom Skalenniveau der Variablen ausgewählt werden.

Ganz anders sieht es bei der Wahl einer $SH_U$ zu einer vorgegebenen PH aus. Wenn wir mit Regel $R_2$ oder $R_{2'}$ fordern, daß die $SH_U$ aus der PH logisch folgt, ***so*** setzt dies die ***Bedeutsamkeit*** der $SH_U$ in dem Sinne voraus, daß der ***Wahrheitswert der $SH_U$ invariant gegenüber zulässigen Transformationen der involvierten ZVn*** ist. Andernfalls wäre die Antwort auf die Frage, ob PH $\Rightarrow$ $SH_U$ gilt, von der willkürlichen Auswahl einer numerischen Zuordnung abhängig.

Auch bei der Untersuchung von Invarianzeigenschaften statistischer Hypothesen ist Sorgfalt geboten. Ein gängiges Vorurteil ist beispielsweise, daß monotone Transformationen von ZVn die Rangordnung von Erwartungswerten invariant lassen, so daß z.B. die gerichtete Mittelwertshypothese $E(Y|X = 1) > E(Y|X = 0)$ auch bei Rangskalenniveau von Y bedeutsam ist. Dies stimmt nicht. Nehmen wir beispielsweise an, daß Y eine Reaktionszeitvariable ist, die in jeder von zwei Gruppen X = 1 bzw. X = 0 lognormalverteilt ist mit $E(Y|X = 1) > E(Y|X = 0)$. Wenn Y Rangskalenniveau aufweist, dürfen wir z.B. die Transformation Y* := ln(Y) anwenden. Unter den genannten Bedingungen muß Y* innerhalb der Gruppen normalverteilt sein. Man kann zeigen, daß die gruppenspezifischen Erwartungswerte von Y und Y* dann in folgender Beziehung zueinander stehen (vgl. Johnson & Kotz, 1970, S. 115, Gleichung 6.1):

$$E(Y|X) = \exp(E(Y^*|X) + 0.5 \cdot \mathrm{Var}(Y^*|X)) \qquad (2.13)$$

Es hängt also von der Varianz von Y* innerhalb der Gruppen ab, ob auch $E(Y^*|X = 1) > E(Y^*|X = 0)$ gilt oder nicht. Wenn man sich im Rahmen eines

statistischen Modells bewegt, das die Varianzen nicht näher normiert, ist demzufolge die gerichtete Mittelwertshypothese bei Rangskalennniveau nicht bedeutsam. Man kann nun darauf verweisen, daß man das ALM zugrunde legt, welches $\mathrm{Var}(Y^* \mid X = 1) = \mathrm{Var}(Y^* \mid X = 0)$ vorschreibt. In diesem Fall wäre zwar die gerichtete Mittelwertshypothese bei Rangskalenniveau bedeutsam. Allerdings wäre dann das zugrundeliegende Modell selbst nicht bedeutsam, denn man verläßt es, wenn man eine beliebige nichtlineare (aber monotone) Transformation von Y vornimmt. Wir können uns also nicht damit begnügen, Invarianzeigenschaften für die Parameterhypothese zu fordern; die Gültigkeit des statistischen Modells selbst muß invariant gegenüber zulässigen Transformationen der ZVn sein. Beides ist gemeint, wenn wir fordern, daß nicht Statistiken, sondern SHn bedeutsam ***(meaningful)*** sein müssen.

## *3. Empfehlungen zur statistischen Entscheidungsstrategie*

Die in den Regeln $R_3$ und $R_{3'}$ geforderte Kontrolle der statistischen Fehlerwahrscheinlichkeiten ist möglich, wenn vor der Durchführung einer Untersuchung neben der Wahrscheinlichkeit $\alpha$ für die fälschliche Zurückweisung der $H_0$ eine Effektgröße spezifiziert wird, die mit der Wahrscheinlichkeit $1-ß$ zu einer korrekten Zurückweisung von $H_0$ führen soll. Bei Gültigkeit bestimmter statistischer Annahmen läßt sich dann der Stichprobenumfang bestimmen, der benötigt wird, um $\alpha$ und ß nicht größer als die Werte werden zu lassen, die festgelegt wurden. Die Bestimmung des Stichprobenumfangs für gegebene Effektstärke sowie gewünschtes $\alpha$ und ß kann z.B. mit Hilfe der Tabellen Cohens (1977 oder 1988) oder Hagers (1987) vorgenommen werden. Inzwischen liegen aber auch komfortable und leicht bedienbare Programme für Personalcomputer vor, die die Planung von Stichprobenumfängen sowie die Durchführung anderer Teststärkeanalysen erheblich erleichtern (z. B. Faul, Erdfelder & Buchner, 1993).

### 3.1 Zur Festlegung von $\alpha$ und ß

Wie in Abschnitt 2.3 ausgeführt, lassen sich Strenge und Fairneß einer Untersuchung nur dann vollständig über $\alpha$ und ß kontrollieren, wenn g und h Null sind. Wird dagegen die $SH_U$ durch die PH nur impliziert, ist also $h = 0$ und $g > 0$, so ist die Kontrolle von Strenge und Fairneß nicht mehr so einfach. Wir schlagen vor, in diesem Fall von der Fairneß auszugehen, die $1-\phi$ beträgt (vgl. Gleichung 2.8, Abschnitt 2.2). Ist festgelegt worden, wie groß die Fairneß mindestens sein soll, so läßt sich sagen, daß die Strenge der Prüfung um so

größer wird, je kleiner $\varepsilon$ gewählt wird. Für verschiedene Werte von $\varepsilon$ und g läßt sich nach der Formel

$$1-e_U = (1-\varepsilon) \cdot (1-g) + \phi \cdot g \qquad \textbf{(3.1)}$$

die Strenge bestimmen. Diese beträgt z.B. .82 für g = .1, $\phi$ = .1 und $\varepsilon$ = .1. Bei g = .1, $\phi$ = .1 und $\varepsilon$ = .2 beträgt die Strenge .73. Wie klein $\varepsilon$ gewählt werden kann, hängt natürlich auch von den zur Verfügung stehenden Ressourcen ab. Wenn z.B. eine PH die $SH_U$ $H_1$: $\mu_1 > \mu_2$ in einem Zwei-Gruppen-Experiment impliziert, sind nach Formel (3.2)

$$n = \frac{(z_{(1-\alpha)} + z_{(1-\beta)})^2}{2f^2} \qquad \textbf{(3.2)}$$

N = 2 . n = 2 . 33 Vpn notwendig, um einen Effekt der Größe $f^2$ = .1 (vgl. Cohen, 1977) bei einseitigem Test zum Niveau $\varepsilon = \alpha$ = .l mit der Wahrscheinlichkeit $1-\phi = 1-\beta$ = .9 entdecken zu können. In dieser Formel sind $z_{1-\alpha}$ bzw. $z_{1-\beta}$ die z-Werte zum (l-a).l00- bzw. (l-ß)l00-Perzentil der Standardnormalverteilung; sie gilt, wenn die AV in den Gruppen normal und mit gleichen Varianzen verteilt ist (vgl. Bredenkamp, 1969). Wird $\alpha$ auf .2 erhöht, reduziert sich der Stichprobenumfang auf 2 . 23 Vpn. Dies ändert nichts an der Fairneß, reduziert aber die Strenge der Prüfung.

## 3.2 Zur Festlegung der Effektstärke

Erfahrungsgemäß bereitet die Festlegung der Effektstärke die meisten Problemc. Möglicherweise ist das scheinbare Fehlen rationaler Kriterien der Effektgrößenfixierung dafür entscheidend, daß Signifikanztests so häufig ohne Kontrolle von ß durchgeführt werden. Allerdings löst dieses Vorgehen das Problem keineswegs. Nach Formel (3.2) hat man sich im Falle einer unreflektierten Wahl des Stichprobenumfangs für jedes $\alpha$ und ß implizit auf eine unbekannte kritische Effektstärke festgelegt (vgl. Bredenkamp, 1969). Durch die Planung des Stichprobenumfangs nach vorgegebenen Werten für $\alpha$, ß und die Effektgröße macht man die Prämissen der Entscheidungsstrategie dagegen explizit und durchschaubar.

Nach welchen Kriterien können Effektgrößen festgelegt werden? Ohne Anspruch auf Vollständigkeit seien hier drei Möglichkeiten aufgeführt. Die erste Möglichkeit besteht darin, aus vorangegangenen Untersuchungen die Effektstärke zu schätzen (vgl. dazu Bredenkamp, 1980). Weiterhin kann man Überlegungen zur Reliabilität der AV einbeziehen. Wenn man annimmt, daß die experimentelle Fehlervarianz sich im Sinne der klassischen Testtheorie additiv aus wahrer Varianz und Meßfehlervarianz zusammensetzt, und festlegt, daß die durch die experimentellen Effekte bedingte Varianz nicht kleiner als die

Meßfehlervarianz sein sollte, impliziert dies die Spezifikation $f^2 = 1\text{-}r_{vv}$, wobei $r_{yy}$ die Reliabilität der AV ist (vgl. Bredenkamp, 1972). Bei reliablen Variablen lohnt demnach die Aufdeckung kleiner Effekte, die relativ viele Vpn erforderlich macht'. Schließlich bleibt die dritte Möglichkeit, sich an den von Cohen (1977) geschaffenen Konventionen zu orientieren, der „kleine", „mittlere" und „große" Effekte ($f^2 = .01$, $f^2 = .0625$ und $f^2 = .16$) unterscheidet. Wenn $SH_U$ eine $H_0$ ist, mag man bereits an der Aufdeckung „kleiner" Abweichungen von $H_0$ interessiert sein. Ist $SH_U$ dagegen eine $H_0$, wird man vielleicht „mittlere" oder „große" Effekte fordern, ehe man die zugrundeliegende PH als bewährt betrachtet.

Allerdings können Cohens (1977) Konventionen auch Probleme aufwerfen. Diese sollen am Beispiel einer monotonen Trendhypothese erläutert werden. Grundlegend ist hier die Unterscheidung zwischen der Effektgröße zu Lasten des Kontrastes zweier Mittelwerte $f_c^2$ und der gesamten Effektstärke $f_t^2$ über alle Bedingungen hinweg. Werden zwei Mittelwerte $\mu_j$ und $\mu_{j'}$ verglichen, gilt

$$f_c^2 = (\mu_j - \mu_{j'})^2 \,/\, (2m\sigma^2), \tag{3.3}$$

wobei $m$ die Gesamtzahl der Versuchsgruppen und $\sigma^2$ die Varianz innerhalb der Bedingungen ist (vgl. Bredenkamp, 1984). Formel (3.3) ist ein Spezialfall von

$$f_c^2 = \frac{(\sum_{j=1}^{m} u_j \mu_j)^2 / (m \sum_{j=1}^{m} u_j^2)}{\sigma^2}, \tag{3.4}$$

wobei $u_j$ das Gewicht ist, das dem j-ten Mittelwert im Kontrast zugewiesen wird. Der in Gleichung (3.4) über dem großen Bruchstrich stehende Ausdruck ist die durch die lineare Regression von $\mu_j$ auf $u_j$ $(j = 1, \ldots, m)$ aufgeklärte Varianz, so daß $f_c^2$ nichts anderes als das Verhältnis dieser Varianz zur Binnenvarianz $\sigma^2$ ist. Bei orthogonalen Kontrasten gilt (vgl. Bredenkamp, 1984):

$$\sum_{c=1}^{m-1} f_c^2 = f_t^2 \,. \tag{3.5}$$

Für nonorthogonale Kontraste jeweils zweier Mittelwerte machen wir uns folgende Beziehung zunutze:

7 Auch wenn die Reliabilität nicht bekannt ist, kann diese Formel hilfreich sein. Dann ist für $r_{yy}$ eine erwartete Reliabilität einzusetzen und $f^2$ entsprechend darauf abzustimmen.

$$\frac{\sum_{j<j'}(\mu_j - \mu_{j'})^2}{m(m-1)} = \frac{\sum_{j=1}^{m}(\mu_j - \mu)^2}{m-1} = \frac{m}{m-1}\sigma^2 f_t^2 . \tag{3.6}$$

Ersetzen von $(\mu_j - \mu_{j'})^2$ durch $2m\sigma^2 f_c^2$ (vgl. Gleichung 3.3) liefert:

$$\frac{2}{m}\sum_{c=1}^{\frac{m(m-1)}{2}} f_c^2 = f_t^2 . \tag{3.7}$$

Die Summe auf der linken Seite der Gleichung läuft hierbei über alle m(m-1) / 2 möglichen paarweisen Mittelwertsvergleiche eines m-Gruppen-Designs. Bei $m$ = 6 besagt z. B. die Hypothese eines streng monoton steigenden Trends $H_1$: $\mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5 < \mu_6$. Angenommen, es wird für jeden der fünf Tests, die sich auf die $H_0$ der Gleichheit zweier benachbarter Mittelwerte beziehen, $f_{c1}^2$ = .01 festgelegt. Bis auf die Konstante $2m\sigma^2$ beträgt dann der Unterschied zwischen benachbarten Mittelwerten mindestens $f_{c1}$ = .l, zwischen Mittelwerten, deren Indexdifferenz 2 beträgt, $f_{c2}$ = .2 usw.; diese $f_c$-Werte sind zu quadrieren und aufzuaddieren (s. Gleichung 3.7). So entstehen fünf Werte $f_{c1}^2$, = .01, vier Werte $f_{c2}^2$ = .04, drei Werte $f_{c3}^2$ = .09, zwei Werte $f_{c4}^2$ = .16 und ein Wert $f_{c5}^2$ = .25. Ihre Summe beträgt 1.05. Multiplikation mit dem Faktor *2/m* = l/3 ergibt $f_t^2$ = .35. Das heißt: Soll für jeden Test ein laut Cohen (1977) ,,kleiner" Effekt entdeckt werden, so ist impliziert, daß der globale Effekt $f_t^2$ = .35 beträgt und damit größer ist als ein ,,großer" Effekt. Bei $f_{c1}^2$ = .16 ergibt sich $f_t^2$ = 5.6. Um einen ,,großen" Effekt $f_{c1}^2$ = .16 bei einseitigem Test, $\alpha$ = .05 und ß = .05 entdecken zu können, sind aber bereits unter jeder Bedingung n = 12, insgesamt also 72 Vpn erforderlich. Dieser Wert ergibt sich aufgrund von Formel (3.8) (Bredenkamp, 1984), die eine Verallgemeinerung von (3.2) ist[8]:

$$n = \frac{(z_{(1-\alpha)} + z_{(1-\beta)})^2}{m f_{c1}^2} . \tag{3.8}$$

Für die Festlegung der Effektgröße resultiert nach den vorangegangenen Ausführungen folgendes: Da die $H_1$ einer Varianzanalyse wegen ihrer Unbestimmtheit wohl kaum jemals die durch die PH implizierte oder mit ihr äquivalente $SH_U$ ist, sollte sich die Effektgrößenbestimmung auf die interessierenden Kontraste beziehen. Hierbei ist immer zu berücksichtigen, was aus diesen Festlegungen für $f_t^2$ resultiert. Da $f_t^2 = \rho^2 / (1-\rho^2)$ bzw. $\rho^2 = f_t^2 / (1+f_t^2)$ -

8 Formel (3.8) führt approximativ zu denselben Ergebnissen wie eine Stichprobenumfangsplanung nach Cohen (1977, Kap. 9), wenn man berücksichtigt, daß Cohen von *zweiseitigen* Tests ausgeht. Die Differenz der berechneten n-Werte ist maximal 1.

wobei $\rho^2$ der in der zugrundeliegenden Population durch die experimentellen Bedingungen aufgeklärte Anteil an der Gesamtvarianz ist -, bedeutet z.B. ein $f_t^2 = 5.6$, daß 85 % der Gesamtvarianz durch die Bedingungsvariation aufgeklärt wird. Das ist ein unrealistisch großer Wert, auf den man sich in der Regel kaum festlegen würde. Reduziert man $f_c^2$, so resultiert ein realistisches $f_t^2$, aber die erforderlichen Stichprobenumfänge werden so groß, daß das Experiment kaum zu realisieren ist. Realisierbare Werte ergeben sich, wenn man sich entschließt, die Anzahl der Bedingungen zu reduzieren und $\alpha$ oder ß zu erhöhen.

Die obigen Erörterungen gelten in analoger Form auch für andere spezielle Mittelwertskontraste. Wenn dagegen die $H_0$ einer Varianzanalyse die durch die PH implizierte $SH_U$ ist, muß bei der Effektgrößenbestimmung von $f_t^2$ ausgegangen werden.

## 3.3 Mehrfache Signifikanztests

Manchmal besteht die abgeleitete $SH_U$ aus der Konjunktion verschiedener SHn. Bei der eben besprochenen $SH_U$ eines monotonen Trends und m = 3 gilt z.B. $SH_U \Leftrightarrow ((H_1^{(1)} : \mu_1 < \mu_2) \wedge (H_1^{(2)}: \mu_2 < \mu_3))$.

Tabelle 1: Statistische Fehlerwahrscheinlichkeiten bei der Entscheidung über eine Konjunktion zweier Alternativhypothesen

| | Zutreffend ist | | | |
|---|---|---|---|---|
| | $SH_U$ | | $\sim SH_U$ | |
| Entscheidung für | $H_1^{(1)} \wedge H_1^{(2)}$ | $H_0^{(1)} \wedge H_0^{(2)}$ | $H_0^{(1)} \wedge H_1^{(2)}$ | $H_1^{(1)} \wedge H_0^{(2)}$ |
| $H_1^{(1)} \wedge H_1^{(2)}$ | $1-\phi = (1-\beta)^2$ | $\varepsilon_1 = \alpha^2$ | $\varepsilon_2 = \alpha(1-\beta)$ | $\varepsilon_3 = \alpha(1-\beta)$ |
| $\sim(H_1^{(1)} \wedge H_1^{(2)})$ | $\phi = 1-(1-\beta)^2$ | $1-\varepsilon_1 = 1-\alpha^2$ | $1-\varepsilon_2 = 1-\alpha(1-\beta)$ | $1-\varepsilon_3 = 1-\alpha(1-\beta)$ |
| $H_1^{(1)} \vee H_1^{(2)}$ | $1-\phi = 1-\beta^2$ | $\varepsilon_1 = 1-(1-\alpha)^2$ | $\varepsilon_2 = 1-\beta(1-\alpha)$ | $\varepsilon_3 = 1-\beta(1-\alpha)$ |
| $\sim(H_1^{(1)} \vee H_1^{(2)})$ | $\phi = \beta^2$ | $1-\varepsilon_1 = (1-\alpha)^2$ | $1-\varepsilon_2 = \beta(1-\alpha)$ | $1-\varepsilon_3 = \beta(1-\alpha)$ |

Tabelle 1 verdeutlicht die Wahrscheinlichkeiten richtiger und falscher Entscheidungen für $SH_U$ unter der Annahme, daß bei den beiden erforderlichen Signifikanztests $\alpha_1$ = CZ, = $\alpha$ und $\beta_1 = \beta_2 = \beta$ ist. Wenn die PH die $SH_U$ eines streng monotonen Trends impliziert, gilt weiterhin, daß die Wahrscheinlichkeit h für das Nichtzutreffen dieser Hypothese bei Gültigkeit der PH Null ist. Die Fairneß ist dann unter Zugrundelegung der Entscheidungsstrategie, nur dann für $SH_U$ zu entscheiden, wenn $H_1^{(1)}$ und $H_1^{(2)}$) angenommen werden, genau $1-\phi = (1-\beta)^2$. Wenn die Fairneß der Untersuchung nicht kleiner als ein

bestimmter Wert werden soll, ist in diesem Fall für die einzelnen Tests ß zu adjustieren, während eine Adjustierung des a-Fehlers keine Vorteile bringt, da die Wahrscheinlichkeit für die fälschliche Annahme von $H_1^{(1)}$ *und* $H_1^{(2)}$) ohnehin kleiner als $\alpha$ ist (vgl. Tabelle 1, obere Hälfte). Eine Adjustierung nach der Formel $\beta\cdot$ = ß / ***(Anzahl Tests)*** stellt sicher, daß die Fairneß nicht unter den gewünschten Wert $1-\beta$ fällt.

Westermann und Hager haben verschiedentlich (z.B. 1986) auf die Notwendigkeit der Adjustierung von ß hingewiesen, wenn eine konjunktive Verknüpfung von Alternativhypothesen aus der PH abgeleitet wird. Bei einer konjunktiven Verknüpfung von Nullhypothesen ist dagegen die Adjustierung von a erforderlich. Zum Fall mehrerer $H_0$- und $H_l$-Hypothesen vgl. Westermann und Hager (1986).

Zurück zum Beispiel. Wie groß ist die Strenge der Prüfung? $SH_U$ kann auf dreierlei Weise nicht zutreffen (vgl. Tabelle l), und die Wahrscheinlichkeiten hierfür seien $g_1$, $g_2$ und $g_3$ genannt. In diesem Fall gilt:

$$e_U = g_1\varepsilon_1 + g_2\varepsilon_2 + g_3\varepsilon_3 + (1 - g_1 - g_2 - g_3)(1 - \phi). \qquad \textit{(3.9)}$$

Ohne Kenntnis der g-Wahrscheinlichkeiten läßt sich über die Strenge der Entscheidungsstrategie nichts sagen.

Die bisher betrachtete Entscheidungsstrategie (obere Hälfte der Tabelle 1) sei mit Westermann und Hager (1986) streng genannt, während die Strategie, auf Richtigkeit der Hypothese $H_1^{(1)} \wedge H_1^{(2)}$) zu entscheiden, wenn wenigstens einer der Tests $H_1$ bekräftigt, schwach genannt sei. Die Wahrscheinlichkeit für richtige und falsche Entscheidungen beim Verfolgen dieser Strategie gibt die untere Hälfte von Tabelle 1 wieder. Wie man erkennt, ist die Fairneß der „strengen" Strategie kleiner als die der „schwachen", während die Strenge größer ist, da jeder einzelne Summand in Gleichung (3.9) für die strenge Strategie kleiner ist. Durch eine Adjustierung des ß-Fehlers kann man diesen Nachteil der „schwachen" Strategie nicht ausgleichen, während der Fairneß-Nachteil der „strengen" Strategie durch eine Reduktion des ß-Fehlers kompensierbar ist.

## 3.4 Nonparametrische Verfahren

Die bislang besprochenen Stichprobenumfangsplanungen sind an Annahmen über die Verteilung der Teststatistik gebunden. Annahmeärmere Hypothesen-Prüfungen lassen sich in experimentellen Untersuchungen mit Hilfe von Randomisierungstests durchfuhren. Da die asymptotische relative Effizienz (A.R.E.) des Randomisierungstests relativ zum F-Test bzw. t-Test im Normalverteilungsfall gleich Eins ist, kann man die Tabellen Cohens (1977, 1988)

oder das Programm von Faul et al. (1993) benutzen, um den Stichprobenumfang für einen Randomisierungstest zu bestimmen (Willmes, 1987), sofern ein solcher für die Testsituation existiert. Seien $N_p$ und $N_v$ die Stichprobenumfänge eines auf der Annahme der Normalverteilung basierenden parametrischen bzw. eines verteilungsfreien Tests. Bei gleichem ß wird mit Wachsendern $N_p$ die entdeckbare Effektgröße immer kleiner werden. $N_v$ ist der Stichprobenumfang des verteilungsfreien Tests, der diese Effektgröße mit gleicher Teststärke entdeckt. Die A.R.E. ist der Wert, dem der Quotient $N_p$ / $N_v$ mit Wachsendern $N_p$ zustrebt. Wenn A.R.E. = 1, kann man erwarten, daß auch im finiten Fall bei nicht zu kleinem Stichprobenumfang die Teststärkeunterschiede Vernachlässigbar sind. Sind die Verteilungsvoraussetzungen des parametrischen Tests nicht erfüllt, dürfte der Randomisierungstest das teststärkere Verfahren sein.

Bei großem N ist allerdings die Anzahl möglicher Permutationen so groß, daß die Durchführung eines exakten Randomisierungstests unmöglich wird. Durchführbar sind aber Tests, die auf einer Zufallsstichprobe aus allen möglichen Permutationen beruhen. Für derartige Monte-Carlo-Lösungen kann entsprechend den Empfehlungen von Willmes (1987) der Stichprobenumfang durch das N im Normalverteilungsfall, dividiert durch die sog. Dwass-Effizienz, bestimmt werden. Wenn $T'(a)$ und $T_s(a)$ die Teststärken des exakten Tests bzw. der Monte-Carlo-Lösung mit dem Simulationsumfang S sind, gilt (vgl. Willmes 1987, S.400):

$$T(\alpha) - T_S(\alpha) \leq T(\alpha) \cdot (1 - e^D_{S,\alpha}), \tag{3.10}$$

wobei $e^D_{S.\alpha}$ die Dwass-Effizienz zum Simulationsumfang S ist. Willmes (1987, S. 400f.) präsentiert Tabellen und Approximationsformeln, welche die Bestimmung des erforderlichen Simulationsumfangs S für vorgegebene Dwass-Effizienzen und $\alpha$-Niveaus ermöglichen. Umgekehrt ist natürlich auch die Bestimmung der Dwass-Effizienz für gegebenes $\alpha$ und S möglich. Bei $\alpha$ = .05 und S = 1219 ist $e^D$ z.B. .95 (Willmes, 1987, S.401, Tab. 8.2).

Viele der auf Rängen beruhenden nonparametrischen Verfahren wie der Kruskal-Wallis-H-Test, der U-Test nach Mann-Whitney usw. sind **de** facto Randomisierungstests. Das erforderliche N läßt sich approximativ bestimmen, wenn der für den Normalverteilungsfall bestimmte Stichprobenumfang durch die A.R.E. des nonparametrischen Tests dividiert wird, die für verschiedene Verfahren z. B. bei Lienert (1973) angegeben ist (Bredenkamp, 1980; Willmes, 1987). Ob derartige Verfahren eingesetzt werden sollten, hängt von der aus der PH abgeleiteten $SH_U$ ab. Die auf Rängen beruhenden Randomisierungstests prüfen die $H_0$ gleicher Rang-Erwartungswerte für verschiedene experimentelle Bedingungen. Nur wenn sich die $SH_U$ auf Erwartungswerte von Rängen bezieht, sind diese Tests indiziert. Dies wird z.B. dann häufig der Fall

sein, wenn für die AV Ordinalskalenniveau unterstellt werden muß. Mittelwertshypothesen sind in diesem Fall nicht bedeutsam (vgl. Abschnitt 2.6), wohl aber Hypothesen über Erwartungswerte von Rängen.

## *4. Stochastische Modelle mit latenten Variablen als Bestandteile einer deduktivistischen Methodologie*

Wenn PHn theoretische Größen beinhalten, deren Ausprägungen der direkten Beobachtung nicht zugänglich sind (wie z. B. „Intelligenz", „Gedächtnisspur", „Empfindungsintensität", „Erfolgsmotiv" usw.), stellt sich die grundlegende Frage, wie eine logische Beziehung zwischen derartigen PHn und beobachtbaren Tatbeständen hergestellt werden kann, die dann eventuell in Form von SHn formulierbar sind (vgl. Abschnitt 1.4). Macht man sich klar, daß PHn, die - wie die Invarianzhypothese des verbalen Lernens (vgl. Abschnitt 1.1) - ausschließlich beobachtbare Größen enthalten, eher die Ausnahme als die Regel sind, wird deutlich, welcher Stellenwert dem Theorie-Empirie-Überbrükkungsproblem in der psychologischen Forschung zukommt.

### 4.1 Probleme der Operationalisierung

Die in der psychologischen Forschung übliche Methode des Umgangs mit dem Theorie-Empirie-Überbrückungsproblem ist die sog. Operationalisierung theoretischer Größen. „Operationalisierung" darf dabei nicht im Sinne von „operationaler Definition" verstanden werden. Operationale Definitionen ziehen bekanntlich Implikationen auch für nicht beobachtete Untersuchungseinheiten nach sich und sind schon aus diesem Grunde abzulehnen (vgl. Herrmann, 1973). Eine „Operationalisierung" entspricht eher dem, was in der analytischen Wissenschaftstheorie als „bilateraler Reduktionssatz" bezeichnet wird. Es handelt sich hierbei um eine bedingte Definition, die eine Äquivalenz von theoretischer und empirischer Größe unter einer ***empirischen (situativen) Randbedingung*** behauptet (vgl. Westmeyer, 1972; Herrmann, 1973). Während allerdings in bilateralen Reduktionssätzen von perfekten, deterministischen Beziehungen zwischen theoretischen und empirischen Größen ausgegangen wird, beruhen Operationalisierungen eher auf der vagen Idee einer nicht-perfekten, „ungefahren" Korrespondenz zwischen theoretischer und empirischer Größe.

Eine theoretische Größe operationalisieren bedeutet in praktischer Hinsicht, ihr (mindestens) eine empirische Indikatorvariable zuzuordnen, die mehr oder minder fehlerbelastet ist, d.h. in nicht perfekt eineindeutiger Beziehung zur theoretischen Größe steht. Wenn diese Zuordnung erfolgt ist, werden alle

theoretischen Größen der PH durch ihre empirischen Indikatoren (z.B. HAWIE-IQ, Reproduktions- oder Rekognitionsleistung, Größenschätzung, Kategorialurteil, TAT-Score usw.) ersetzt. Aus der so resultierenden modifizierten Hypothese, die wir in Anlehnung an Hager (1987) als ***empirische Hypothese*** (EH) bezeichnen wollen, wird dann eine $SH_U$ abgeleitet und statistisch geprüft (vgl. Hussy & Möller, Kapitel 11 dieses Bandes). Die statistische Entscheidung wird anschließend der Beurteilung der PH zugrunde gelegt.

Ist dieses Vorgehen gerechtfertigt? Dies ist selbst dann nicht notwendigerweise der Fall, wenn eine Implikation EH $\Rightarrow$ $SH_U$ oder gar eine Äquivalenz EH $\Leftrightarrow$ $SH_U$ nachweislich vorliegt. Zu fordern ist nämlich aufgrund von Regel $R_2$ bzw. $R_{2'}$ eine Äquivalenz- oder Implikationsbeziehung nicht zwischen EH und $SH_U$, sondern zwischen PH und $SH_U$. Über die Beziehung zwischen PH und $SH_U$ ist jedoch zunächst nichts bekannt, da EH und PH in keiner logisch determinierten Beziehung zueinander stehen.

Wie kann gezeigt werden, daß die aus einer EH gewonnene $SH_U$ nicht aus der eigentlich zu prüfenden PH folgt? Hierzu muß man jeweils im Einzelfall ein stochastisches Modell formulieren, das (a) eine geeignete Darstellung der zentralen theoretischen Aussagen der PH beinhaltet (sog. ,,Strukturmodell“) und (b) Beziehungen zwischen den als latente ZVn aufzufassenden theoretischen Größen und den empirischen ZVn der $SH_U$ herstellt (sog. ,,Meßmodell“). Kann man nun zeigen, daß zu einem gegebenen Strukturmodell (der „eigentlichen“ PH also) mindestens ein plausibles Meßmodell denkbar ist, derart, daß die Negation von $SH_U$ aus der Konjunktion von Struktur- und Meßmodell ableitbar ist, dann ist damit der Nachweis des Nichtbestehens einer Implikationsbeziehung PH $\Rightarrow$ $SH_U$ erbracht. Beispiele findet man z.B. bei Batchelder und Riefer (1986) sowie Erdfelder (1991).

## 4.2 Die stochastische Formulierung psychologischer Hypothesen mit theoretischen Größen

In der Psychologie gibt es eine lange Tradition psychophysikalischer Skalierungsmodelle, deren Bedeutung für die Lösung des Theorie-Empirie-Überbrückungsproblems außerhalb der Psychophysik lange Zeit übersehen wurde. Der Grundgedanke dieser Modelle besteht darin, theoretische psychologische Größen - wie z.B. die durch einen Reiz ausgelöste Empfindungsintensität - durch latente ZVn oder Parameter der Verteilung latenter ZVn (kurz: latente Parameter) zu repräsentieren. ,,Latent“ soll in diesem Zusammenhang einfach besagen, daß die Ausprägungen der entsprechenden ZVn nicht beobachtbar sind oder zumindest nicht beobachtet wurden, so daß die Schätzung der latenten Parameter nicht anhand einer Stichprobe aus der Verteilung der latenten

ZVn erfolgen kann. Lösbar wird das Problem der Schätzung latenter Parameter dadurch, daß über ein geeignetes Meßmodell eine Beziehung zwischen den Verteilungen latenter und beobachteter ZVn hergestellt wird, derart, daß Schätzer für die latenten Parameter aus Schätzern für die Parameter der Verteilung beobachtbarer ZVn ableitbar sind.

Wann ist ein Meßmodell „geeignet"? Unseres Erachtens sind einerseits formale, andererseits aber auch sog. inhaltliche Kriterien ausschlaggebend dafür, ob ein bestimmtes stochastisches Modell als Werkzeug zur Überprüfung einer bestimmten PH eingesetzt werden sollte. Die formalen Kriterien lassen sich am besten anhand der durch das Meßmodell (bzw. die entsprechenden Modellgleichungen) definierten Abbildung f: $\Lambda \rightarrow 0$ erläutern, wobei $\Lambda$ die Menge möglicher latenter Parameterwerte und 0 die Menge der möglichen Parameterwerte der beobachtbaren Verteilung ist. Durch die Abbildung f wird jedem $\lambda \in \Lambda$ (jedem u.U. mehrdimensionalen latenten Parameter) genau ein $o \in 0$ (ein u. U. mehrdimensionaler Parameter der beobachtbaren Verteilung) zugeordnet.

Es lassen sich nun drei Kriterien formulieren: (1) Die Abbildung f sollte möglichst nicht surjektiv sein, d.h. die Bildmenge f(A) sollte eine echte Teilmenge von 0 sein. Konkret bedeutet dies, daß Parameterkonstellationen der beobachtbaren Verteilung existieren (nämlich alle $o \in O \backslash f(\Lambda)$), die bei Gültigkeit der Modellgleichungen nicht vorkommen dürften. Ist diese Situation gegeben, so ist es prinzipiell möglich, einen Modellgeltungstest zu konstruieren. Da dieser Test nur eine von mehreren Möglichkeiten der empirischen Adäquatheitsprüfung darstellt (vgl. Punkt 3 weiter unten), wollen wir die Nichtsurjektivität von f nicht zwingend fordern. Prinzipiell sollen also auch Modelle zugelassen werden, die keine beobachtbaren Verteilungen ausschließen und für die sich demzufolge kein Modellanpassungstest konstruieren läßt[9]. Wenn allerdings die Abbildung f nicht surjektiv ist, so wollen wir fordern, daß der Modellanpassungstest positiv ausfällt. Auf statistische Probleme, die sich im Zusammenhang mit der Modellgeltungsprüfung stellen, gehen wir in Abschnitt 4.4 ein.

(2) Die Abbildung f sollte injektiv sein, d. h. unterschiedlichen latenten Parametern sollten auch unterschiedliche Parameter der beobachtbaren Verteilung zugeordnet sein. Ist diese Bedingung erfüllt, so heißt das Modell auch „global identifizierbar". Die globale Identifizierbarkeit stellt sicher, daß eine Funktion g: $f(\Lambda) \rightarrow \Lambda$, nämlich $g \Leftrightarrow f^{-1}$, definiert werden kann, die es erlaubt, die interessierenden latenten Parameter aus den Parametern der beobachtbaren Verteilung rückzurechnen. Dies ist eine Voraussetzung für die Schätzbarkeit der

9 Die Two-High Threshold-Theorie (vgl. Snodgrass & Corwin, 1988) ist ein Beispiel für ein derartiges saturiertes Modell, sofern sie auf lediglich eine Versuchsbedingung (eine Trefferund eine falsche Alarmrate) angewendet wird.

latenten Parameter. Dennoch ist die globale Identifizierbarkeit keine generell unverzichtbare Forderung. Impliziert nämlich die zu prüfende PH keine bestimmte $SH_U$ über die latenten Parameter, sondern lediglich die Gültigkeit des entsprechenden Modells ***unabhängig von konkreten Parameterwerten, so*** ist die Forderung nach Injektivität von f Verzichtbar. Grundsätzlich müssen nur die Parameter identifizierbar sein, für die sich aus der zu prüfenden PH konkrete Hypothesen ableiten lassen. Probleme, die sich bei der Modellgeltungsprüfung nichtidentifizierbarer Modelle stellen, werden ebenfalls in Abschnitt 4.4 angesprochen.

(3) Ist ein Modell global identifizierbar oder sind zumindest einige latente Parameter des Modells identifizierbar, so muß das Modell bezüglich der identifizierbaren Parameter ***konstruktvalide*** sein. Der Begriff der Konstruktvalidität wird hier in Anlehnung an Cronbach und Meehl (1956) gebraucht. Gemeint ist damit, daß experimentelle UVn, die gemäß der zu prüfenden PH bestimmte theoretische Größen beeinflussen, einen hypothesenkonformen Effekt auf genau die latenten Parameter ausüben müssen, welche die entsprechenden theoretischen Größen im stochastischen Modell repräsentieren.

Alle drei genannten Kriterien lassen sich sehr schön am Beispiel der Signalentdeckungstheorie (Green & Swets, 1974) erläutern. Zunächst muß die Abbildung f untersucht werden, die den latenten Parametern d' (Sensitivität) und ß (Antworttendenz) Parameter der beobachtbaren Verteilung - nämlich Wahrscheinlichkeiten eines Treffers und eines falschen Alarms - zuordnet. Eine nähere Analyse zeigt, daß diese Abbildung surjektiv und injektiv ist, wenn nur eine Versuchsbedingung realisiert wurde, d.h. nur eine Treffer- und nur eine falsche Alarmrate vorliegt. Das Modell ist unter diesen Umständen somit nicht testbar, wohl aber global identifizierbar, d.h. es lassen sich Schätzer für ***d'*** und ß ableiten. Eine Möglichkeit, die Konstruktvalidität des Modells zu überprüfen, besteht darin, Treffer- und falsche Alarmraten unter ***k*** verschiedenen Versuchsbedingungen zu erheben, die sich z.B. hinsichtlich der Auszahlungen für falsche und korrekte Ja-Antworten unterscheiden. Es läßt sich nun die PH prüfen, daß die Auszahlungen die Antworttendenzen $\beta_i$, nicht aber die Sensitivitäten $d'_j$ (j = 1, . . . , ***k)*** beeinflussen. Hierzu ist lediglich ein simultanes signalentdeckungstheoretisches Modell für alle k Versuchsbedingungen mit der Parameterrestriktion $d'_1 = d'_2 = \ldots = d'_k$ zu formulieren. Dieses Modell entspricht der zu testenden $SH_U$. Es umfaßt nicht mehr 2, sondern ***2k*** Modellgleichungen (für ***k*** Paare von Treffer- und falschen Alarmraten). Während 0 somit nun 2k-dimensional ist, ist $\Lambda$ lediglich k+l-dimensional, da neben ***k*** Antworttendenzparametern $\beta_i$ lediglich ein Sensitivitätsparameter ***d'*** vorgesehen ist. Eine Analyse dieser Abbildung zeigt, daß sie nicht surjektiv und injektiv ist, so daß (a) ein Modellgeltungstest konstruierbar und (b) das Modell global identifizierbar ist. Der Modellgeltungstest hätte in diesem Fall die $H_0$ zu überprüfen, daß alle ***k*** Paare von Treffer- und falschen Alarmwahr-

scheinlichkeiten auf einer durch die Modellgleichungen definierten ROC-Kurve liegen. Geht der Test insignifikant aus und beeinflußt die UV „Auszahlung" die ß-Parameter in der erwarteten Richtung, so liegt eine Bestätigung der zu prüfenden PH vor. Das Modell kann als ***für diese empirische Anwendungssituation*** konstruktvalide gelten. Weitere Möglichkeiten der Konstruktvalidierung bestehen darin, andere UVn zu variieren, für die PHn einen Effekt auf die Sensitivität ***d'*** oder die Antworttendenz ß behaupten.

## 4.3 Eine Auswahl wichtiger stochastischer Rahmenmodelle

Die Möglichkeit, das Theorie-Empirie-Überbrückungsproblem mittels stochastischer Skalierungsmodelle anzugehen, wird in der Psychophysik schon seit längerer Zeit genutzt. Erst in jüngerer Zeit zeichnet sich jedoch ab, daß das Prinzip der stochastischen Modellbildung auch in anderen Bereichen der Psychologie nutzbar gemacht werden kann. Dies gilt vor allem für die Bereiche, in denen die psychophysikalischen Skalierungsmodelle ohne methodische Modifikationen zur Beantwortung neuer psychologischer Fragestellungen einsetzbar sind, z.B. bei der Analyse von Rekognitionsdaten in der Gedächtnispsychologie (vgl. Snodgrass & Corwin, 1988; Bredenkamp & Erdfelder, 1993). Die Verfügbarkeit schneller Computer und damit die praktische Realisierbarkeit aufwendiger numerischer Verfahren der Parameterschätzung und Modellprüfung läßt darüber hinaus inzwischen die Anwendung weitaus komplizierterer stochastischer Modelle mit latenten Variablen zu. Die entsprechenden stochastischen Rahmenmodelle sind durchweg nicht neu, schieden jedoch in der Vergangenheit häufig einfach deshalb als Werkzeuge der Hypothesenprüfung aus, weil die zugehörige statistische Analyse nicht praktikabel war.

Ohne Anspruch auf Vollständigkeit sollen einige wichtige Rahmenmodelle angesprochen werden, die als Instrumente der Überprüfung von PHn in ***Zu***kunft noch an Bedeutung gewinnen könnten. Zunächst wären hier ***Strukturgleichungsmodelle*** zu nennen. Sie gehören zu den derzeit populärsten Modellen mit latenten Variablen, nicht zuletzt wohl aufgrund der allgemeinen Verfügbarkeit benutzerfreundlicher Computerprogramme (z. B. Jöreskog & Sörbom, 1988). PHn, die sich als lineare Regressionshypothesen formulieren lassen, d.h. einfache und multiple Regressions- sowie Pfadmodelle, können im Rahmen von Strukturgleichungsmodellen angemessen formuliert und geprüft werden; dies gilt insbesondere dann, wenn die involvierten ZVn „latent" sind und lediglich meßfehlerbelastete Indikatoren vorliegen. Strukturgleichungsmodelle decken damit den Bereich komplett ab, der sich mit den Schlagworten „Faktorenanalyse" und „Pfadanalyse mit latenten und manifesten Variablen" grob charakterisieren läßt. Die Modellgleichungen linearer Strukturgleichungsmodelle drücken die Kovarianzen beobachteter ZVn als

Funktion latenter Parameter (latenter Pfadkoeffizienten und Faktorladungen sowie Varianzen und ggf. auch Kovarianzen latenter ZVn) aus. Zugrunde gelegt wird dabei immer ein lineares Meßmodell, d.h. es wird angenommen, daß beobachtete und latente ZVn in linearer Beziehung zueinander stehen. Eine Darstellung der Grundlagen linearer Strukturgleichsmodelle unter besonderer Berücksichtigung nichtrekursiver Modelle und der Identifizierbarkeitsproblematik gibt Andres (1990). Die Anwendungen linearer Strukturgleichsmodelle dürften in erster Linie im Bereich der Differentiellen Psychologie und im Bereich der Entwicklungspsychologie liegen. Für die experimentelle Psychologie dürften primär simultane Strukturgleichungsmodelle über mehrere Gruppen von Bedeutung sein. Mit diesem Spezialfall, der z.B. via LISREL (Jöreskog & Sörbom, 1988) analysierbar ist, lassen sich Varianzanalysen für latente AVn durchführen, soweit mehrere Indikatorvariablen dieser AV verfügbar sind (intraexperimentelle Replikation der AV) und ein lineares Meßmodell unterstellt werden kann (LISREL-MANOVA).

Eine weitere wichtige Modellklasse, die in der Psychologie erst in jüngster Zeit Beachtung findet, bilden ***sog. finite Mischverteilungen*** (Titterington, Smith & Makov, 1985; Erdfelder, 1990a; Rost & Langeheine, 1991). Finite Mischverteilungen stellen einen geeigneten formalen Rahmen zur Prüfung von PHn zur Verfügung, die die Untersuchungseinheiten (z.B. Vpn) in mehrere disjunkte und exhaustive „latente Klassen" unterteilen, für die jeweils unterschiedliche Verteilungen beobachteter ZVn gelten. „Latent" bedeutet hier wiederum, daß eine exakte Indikatorvariable für die Klassenzugehörigkeit einer Untersuchungseinheit nicht vorliegt. Es können daher nicht die Verteilungen der beobachteten ZVn innerhalb der latenten Klassen, sondern nur ihre „Mischung" über alle Klassen hinweg beobachtet werden.

Ein wichtiger Spezialfall finiter Mischverteilungsmodelle ist das auf P. Lazarsfeld zurückgehende Modell mehrerer latenter Klassen ***(latent class model, vgl.*** Lazarsfeld & Henry, 1968). Es ergibt sich aus dem allgemeinen finiten Mischverteilungsmodell durch die Annahme, daß die beobachteten ZVn zum einen qualitativ und zum anderen lokal (d.h. innerhalb der latenten Klassen) stochastisch unabhängig sind. Obwohl dieses Modell bislang vorwiegend in der Soziologie und in der Differentiellen und Diagnostischen Psychologie eingesetzt wurde, ist es für alle Bereiche der Psychologie von außerordentlicher Bedeutung. So läßt sich etwa zeigen, daß multinomiale probabilistische Ereignisbäume - von Riefer und Batchelder (1988) als Instrumente zur Prüfung kognitionspsychologischer Hypothesen und zur Messung kognitiver Prozesse propagiert - nichts anderes sind als spezielle Modelle mehrerer latenter Klassen (Erdfelder, 1990b).

Von großer Bedeutung sind ferner ***Markoff-Modelle*** (Wickens, 1982), deren Einsatz bislang auf stochastische Lerntheorien (vgl. Tack, 1976) weitgehend

beschränkt blieb, die sich aber zur Formulierung von PHn, welche sich auf zeitlich erstreckte Prozesse beziehen, auch in anderen Bereichen der Psychologie prinzipiell eignen. ***Reaktionszeitmodelle*** (Townsend & Ashby, 1983; Luce, 1986) bilden eine in formaler Hinsicht heterogene Modellklasse, welche eine Vielzahl kognitionspsychologischer PHn zu formulieren und zu prüfen gestattet.

## 4.4 Probleme der Modellgeltungsprüfung

In Abschnitt 4.2 wurde festgestellt, daß die Modellgleichungen stochastischer Modelle möglichst eine Abbildung f definieren sollten, die nicht surjektiv ist. In diesem Fall ist eine Teilklasse denkbarer beobachtbarer Verteilungen mit der Gültigkeit des Modells unvereinbar. Prinzipiell ist es somit möglich, einen Modellgeltungstest zu konstruieren, der prüft, ob der (u.U. mehrdimensionale) Parameter $o_e$, welcher die empirische Verteilung charakterisiert, im „erlaubten Bereich", d.h. in der Bildmenge $f(\Lambda)$ des Raumes latenter Parameter, oder außerhalb des erlaubten Bereichs liegt. Ein solcher Modellgeltungstest kann natürlich nur ein statistischer Hypothesentest sein, da die Modellgeltungshypothese sich nicht auf Statistiken (Kennwerte der beobachteten Stichprobe), sondern auf Parameter (Kennwerte der zugrundeliegenden Verteilung beobachtbarer ZVn) bezieht. Das Problem besteht somit zunächst darin, eine Teststatistik und eine Stichprobenverteilung dieser Teststatistik unter der Nullhypothese

$$H_0\colon o_e \in f(\Lambda) \tag{4.1}$$

abzuleiten. Diese $H_0$ entspricht der $SH_U$, die aus der vorgeordneten PH abgeleitet wurde. Eine Entscheidung für $H_0$ ist demnach als Bewährungsurteil bzgl. PH, eine Entscheidung gegen $H_0$ als Nichtbewährungsurteil bzgl. PH zu interpretieren.

Wilks (1938) hat vorgeschlagen, die transformierte Likelihood-Quotienten-Statistik

$$L^2 := -2 \cdot \ln(L(M_0) \,/\, L(M_1)) \tag{4.2}$$

zum Vergleich der durch verschiedene stochastische Modelle erzielten Datenanpassungen heranzuziehen. In Gleichung (4.2) bezeichnet $L(M_0)$ die maximale Likelihood[10] für eine gegebene Datenstichprobe unter einem stochastischen Modell $M_0$ und $L(M_1)$ analog die maximale Likelihood für den gleichen Datensatz unter einem „allgemeineren" stochastischen Modell $M_1$. Die Be-

10 Gemeint ist die Likelihood, die resultiert, wenn die freien Parameter des Modells durch ihre Maximum-Likelihood (ML) - Schätzer ersetzt werden. Die Existenz von ML-Schätzern wird vorausgesetzt.

deutung von „allgemeiner" wird klarer, wenn man sich die Modellgleichungen der beiden Modelle anschaut, welche zwei Abbildungen $f_o$: $\Lambda_0 \rightarrow 0$ und $f_1$: $\Lambda_1 \rightarrow 0$ in die Menge möglicher empirischer Verteilungen definieren. Gilt für die beiden Bildmengen die Teilmengenrelation $f_0(A_0) \subset f_1(A_1)$, ist also jeder $M_o$-konforme Parameter $o \in 0$ zugleich $M_1$-konform, so heißt $M_1$ „allgemeiner" als $M_0$. Geht es - wie im hier zu diskutierenden Fall der Modellgeltungshypothese gemäß Gleichung (4.1) - in erster Linie nicht um den Vergleich zweier stochastischer Modelle, sondern um die statistische Evaluation eines bestimmten Modells $M_0$, wählt man zweckmäßigerweise $M_1$ als saturiertes Modell, so daß $f_1(A_1) = 0$.

Wilks (1938) hat gezeigt, daß die Statistik $L^2$ unter bestimmten Regularitätsbedingungen bei Gültigkeit von $H_0$ asymptotisch zentral $\chi^2$-verteilt ist, wobei sich die Freiheitsgrade df aus der Anzahl *unabhängiger* Parameterrestriktionen ergeben, die notwendig sind, um das Modell $M_1$ in den Spezialfall $M_0$ zu überführen. Im Regelfall entspricht df gerade der Differenz zwischen der Anzahl freier Parameter in $M_1$ und in $M_0$. Darüber hinaus läßt sich nachweisen, daß die Wilksschen Bedingungen ebenfalls hinreichend sind, um für $N \rightarrow \infty$ eine nonzentrale $X^2_{(df)}$-Verteilung von $L^2$ bei Ungültigkeit des Modells $M_0$ (d.h. $o \in O \backslash f_0(\Lambda)$) zu garantieren. Dies eröffnet die Möglichkeit der Kontrolle von ß, z.B. mit Hilfe der $\chi^2$-Tabellen von Cohen (1977 oder 1988) oder mit Hilfe des Programms von Faul et al. (1993).

Die simultane Kontrolle von $\alpha$ und ß scheint also bei Verwendung der Wilksschen Statistik auf den ersten Blick keine Probleme zu bereiten. Leider täuscht dieser Eindruck, da gerade bei Modellgeltungstests für die im letzten Abschnitt erwähnten Modelle die Regularitätsbedingungen, die zur Herleitung der asymptotischen Verteilung von $L^2$ benötigt werden, oftmals nicht erfüllt sind. Grundsätzlich gibt es die Möglichkeit, die genannten Probleme in jedem Einzelfall durch Adjustierungen der $L^2$-Statistik anzugehen, derart, daß nach einer „korrigierten" $L^2$-Statistik gesucht wird, die unter $H_0$ annähernd einer Standard-$\chi^2$-Verteilung folgt. Dieses Vorgehen war in einigen Fällen erfolgreich (vgl. Titterington et al., 1985), verlangt allerdings als Rechtfertigung immer eine Simulationsstudie, die die Adäquatheit der verwendeten Korrekturformel belegt. Wenn ein solcher Aufwand schon getrieben werden muß, gibt es u. E. bessere, generell anwendbare Alternativen, wie z.B. die von Aitkin, Anderson und Hinde (1981) zur Prüfung von Latent-Class-Modellen vorgeschlagene Vorgehensweise: Sie approximierten die Stichprobenverteilung von $L^2$ über eine Monte-Carlo-Studie, in der wiederholt Stichproben des (empirisch realisierten) Umfangs N nach einem Modell generiert wurden, das aus der Ersetzung aller freien Modellparameter durch ihre ML-Schätzer resultiert.

Aus der Sicht einer deduktivistischen Methodologie ergibt sich ein weiteres Problem, das nicht ausgeklammert werden soll: das Problem der Rechtferti-

gung der Verteilungsannahmen, die den Inferenzverfahren (und auch den ML-Schätzern) zugrunde liegen. Beziehen sich die aus PHn abgeleiteten SHn auf Mittelwerte verschiedener experimenteller Bedingungen, so gibt es - wie in Abschnitt 1.3 ausgeführt - prinzipiell die Möglichkeit, die üblichen parametrischen Tests als approximative Randomisierungstests ohne ***jede*** Verteilungs-***annahme*** zu rechtfertigen. Gibt es eine ähnliche Möglichkeit auch für Modellgeltungstests, die auf $L^2$ basieren? Eine endgültige Antwort auf diese Frage kann derzeit noch nicht gegeben werden. Wir vermuten jedoch, daß die von Efron (1979) vorgeschlagene Bootstrap-Methode hier u. U. weiterhilft. Die Bootstrap-Methode ist im Kern eine Konkretisierung der Idee, daß die empirisch vorliegende Stichprobe die bestmögliche Schätzung für die zugrundeliegende Population ist. Ist über die Population nichts bekannt, was über die vorliegende Stichprobe hinausgeht, so liegt es nahe, die Stichprobenverteilung einer Statistik in der Weise zu approximieren, daß man aus der empirisch vorliegenden Stichprobe des Umfangs N weitere Stichproben des Umfangs N ***mit Zurücklegen*** zieht (sog. Bootstrap-Stichproben). Für jede aus der empirischen Stichprobe gezogene Bootstrap-Stichprobe wird die zur Diskussion stehende Statistik berechnet, so daß eine Bootstrap-Stichprobenverteilung dieser Statistik resultiert.

Efron hat diese Methode vor allen Dingen zur verteilungsfreien Bestimmung von Konfidenzintervallen und Standardschätzfehlern herangezogen. Hierzu liegt inzwischen auch ein beachtliches Arsenal an mathematisch-statistischer Literatur vor, das die positiven Eigenschaften der Bootstrap-Methode analytisch zu begründen erlaubt (vgl. das Literaturverzeichnis von Sievers, 1990). Simulationsstudien von Sievers (1990) deuten darüber hinaus an, daß die Leistungsfähigkeit der Bootstrap-Methode möglicherweise weit über das hinausreicht, was bislang in der mathematischen Statistik untersucht wurde. Seine Ergebnisse zeigen in beeindruckender Weise, daß einiges. dafür spricht, auch statistische Inferenzverfahren im Lichte der Bootstrap-Methode zu betrachten: Die zentralen Verteilungen einiger üblicher F-Statistiken (ANOVA und Hotellings $T^2$) werden bei Gültigkeit der parametrischen Verteilungsannahmen nahezu perfekt durch die entsprechenden Bootstrap-Verteilungen approximiert. Sind die Standardannahmen dagegen verletzt, so halten die parametrischen Tests das a-Niveau z.T. nicht ein, während das Bootstrap-Verfahren unter allen Bedingungen die nominellen Fehlerrisiken bewahrt. Damit wird der Versuch nahegelegt, die Bootstrap-Methode auch zur Absicherung von Modellgeltungstests heranzuziehen. Ob dieser Versuch in eine generelle Empfehlung einmünden kann, muß von weiteren Resultaten zur Bootstrap-Methode abhängig gemacht werden.

## 4.5 Einwände gegen stochastische Modellbildung zwecks Überprüfung psychologischer Hypothesen

Gegen die Verwendung formaler Modelle in der Psychologie sind verschiedentlich Einwände erhoben worden, auf die abschließend kurz eingegangen werden soll. Eine recht ausführliche Liste dieser Argumente hat Deppe (1977, Kapitel 8) zusammengestellt und zugleich Gegenargumente geliefert, die aufzeigen, daß die Kritikpunkte eigentlich nicht formale Modelle per se, sondern lediglich bestimmte Formen ihrer Verwendung betreffen. Sehr viele Kritikpunkte werden hinfällig, wenn deutlich gemacht wird, daß Modellbildung nicht auf eine kaum zu leistende isomorphe Abbildung des Gegenstandsbereiches abzielt. Die Art der Verwendung, von der in dieser Arbeit die Rede ist, hat z.B. wesentlich bescheidenere Ziele. Stochastische Modelle interessieren lediglich insoweit, als sie mathematische Formulierungen (notwendiger Bedingungen) von PHn für bestimmte Untersuchungssituationen darstellen. Sie sind somit lediglich ***Werkzeuge der Hypothesenprüfung*** und sonst nichts.

Sind sie geeignete Werkzeuge in dem Sinne, daß sie eine optimale Überprüfung von PHn erlauben? Die Antwort auf diese Frage hängt davon ab, was man genau unter einer PH versteht. Betrachten wir den Prototyp der Frustrations-Aggressions-Hypothese, die für eine bestimmte Untersuchungssituation formuliert wird: „Für alle Vpn x der Untersuchung gilt: Wenn x frustriert wird, reagiert x aggressiv." Diese Hypothese läßt sich als latente 2·2-Kontingenztafel mit einer leeren Zelle im Rahmen eines finiten Mischverteilungsmodells formulieren: Die latente Klasse der frustrierten und nicht aggressiven Personen muß hypothesengemäß leer sein, alle anderen latenten Klassen (frustriert und aggressiv, nicht frustriert und aggressiv, nicht frustriert und nicht aggressiv) können dagegen beliebig frequentiert sein. Damit hat man allerdings noch kein testbares Modell. Das ***Strukturmodell*** bedarf der Ergänzung durch ein ***Meßmodell,*** welches die Beziehung zu beobachtbaren ZVn herstellt. Erst die Konjunktion von Struktur- und Meßmodell kann über einen Modellgeltungstest geprüft werden.

Versteht man unter der PH eine Konjunktion von Struktur- und Meßmodell, so kann offenbar ohne Einschränkung davon gesprochen werden, daß stochastische Modelle eine geeignete Methode der Überprüfung von PHn sind. Versteht man dagegen unter einer PH ausschließlich das, was im Rahmen des stochastischen Modells „Strukturmodell" genannt wird, so kann der Modellgeltungstest nicht ohne weiteres als Test der PH interpretiert werden. Denkbar ist ja, daß das stochastische Modell aufgrund eines ungeeigneten Meßmodells verworfen werden muß, obwohl der „psychologische Kern" - das Strukturmodell - durchaus zutrifft. Dieses Problem ist als Duhem-Quine-Problem bekannt (z.B. Gadenne, 1984, Kapitel 9 dieses Bandes) und keineswegs spe-

zifisch für Hypothesenpriifungen mittels stochastischer Modelle. Grundsätzlich gelingt es fast nie, eine direkte Implikationsbeziehung zwischen einer primär interessierenden wissenschaftlichen Hypothese und bestimmten Datenklassen herzustellen. Fast immer ist eine ***Konjunktion*** von wissenschaftlicher Hypothese und Hilfsannahmen erforderlich, um empirisch prüfbare Folgerungen ableiten zu können. Treten diese Folgerungen empirisch nicht ein, ist es logisch immer vertretbar, das Scheitern auf die Hilfsannahmen und nicht auf die wissenschaftliche Kernhypothese zu attribuieren.

Als Ausweg aus dem Duhem-Quine-Problem bleibt nur die nüchterne Einsicht, daß isolierten psychologischen Kernhypothesen ohne Meßmodell für die involvierten theoretischen Größen nicht sinnvoll empirische Prüfbarkeit attestiert werden kann. Die o. g. Frustrations-Aggressions-Hypothese ist also keine prüfbare PH, solange man sich nicht auf ein bestimmtes Meßmodell für ,,Frustration“ und ,,Aggression“ festgelegt hat. Ohne ein solches Meßmodell ist die Frustrations-Aggressions-Hypothese lediglich eine Heuristik, die in Verbindung mit Intuitionen über adäquate Operationalisierungen prüfbare PHn unregen kann. Bezüglich einer Heuristik fragt man aber sinnvollerweise nicht nach Wahrheit und Falschheit, sondern nach Fruchtbarkeit oder Unfruchtbarkeit.

## Literatur


Aitkin, M., Anderson, D. & Hinde, J. (1981). Statistical modeling of data on teaching styles (with discussion). ***Journal of the Royal Statistical Society, A,*** 144, 419-461.

Andres, J. (1990). ***Grundlagen linearer Strukturgleichungsmodelle.*** Frankfurt: Lang.

Baddeley, A. ***(1990). Human memory.*** London and Hove: Erlbaum.

Bakan, D. (1954). A generalization of Sidman’s results on group and individual functions, and a criterion. ***Psychological Bulletin,*** 51, 63-64.

Batchelder, W. H. & Riefer, D.M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. ***British Journal of Mathematical and Statistical Psychology,*** 39, 129-149.

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: Akademische Verlagsgesellschaft.

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

Bredenkamp, J. (1982). Verfahren zur Ermittlung des Typs der statistischen Wechselwirkung. ***Psychologische Beiträge,*** 24, 56-75 und 309.

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. & Erdfelder, E. (1993). Methoden der Gedächtnispsychologie. In D. Albert & K.-H. Stapf (Hrsg.), ***Gedächtnis*** (= Enzyklopädie der Psychologie, Themenbereich C, Serie II, Band 4). Göttingen: Hogrefe (im Druck).

Bugelski, B. R. (1962). Presentation time, total time, and mediation in paired-associate learning. ***Journal of Experimental Psychology, 63,*** 409-412.

Campbell, D.T. & Stanley, J. C (1963). Experimental and quasi-experimental designs for research on teaching. In ***N.*** L. Gage (Ed.), ***Handbook of research on teaching.*** Chicago: Randy McNally.

Cohen, J. ***(1977). Statistical power analysis for the behavioral sciences*** (revised edition). New York: Academic Press.

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

Cronbach, L.J. & Meehl, P. E. (1956). Construct validity in psychological tests. In H. Feigl & M. Scriven (Eds.), ***Minnesota studies in the philosophy of science. Volume Z: The foundations of science and the concepts of psychology and psychoanalysis.*** Minneapolis: University of Minnesota Press.

Deppe, W. (1977). ***Formale Modelle in der Psychologie.*** Stuttgart: Kohlhammer.

Edgington, E. ***S.*** (1969). ***Statistical inference: the distribution-free approach. New*** York: Mc Graw-Hill.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. ***Annals of Statistics,*** 7, 1-26.

Erdfelder, E. (1990a). Deterministic developmental hypotheses, probabilistic rules of manifestation, and the analysis of finite mixture distributions. In A. von Eye (Ed.), ***Statistical methods in longitudinal research. Volume II: Time series and categorical longitudinal data*** (pp. 471-509). Boston: Academic Press.

Erdfelder, E. (1990b). Probabilistische Ereignisbäume als restringierte univariate Latent-Class-Modelle. In D. Frey (Hrsg.), ***Bericht über den 37. Kongreß der*** Deutschen Gesellschaft für Psychologie in Kiel 1990. Band 1: Kurzfassungen (S. 590-591). Göttingen: Hogrefe.

Erdfelder, E. (1991). Prädiktionsanalyse und die Analyse latenter Klassen: Welches Verfahren für welche Fragestellung? In A. von Eye (Hrsg.), ***Prädiktionsanalyse. Vorhersagen mit kategorialen Variablen*** (S. 157-191). Weinheim: Psychologie Verlags Union.

Erdfelder, E. (1993a). ***Entscheidungen über statistische und wissenschaftliche Hypothesen:*** Welche ***Fehlerwahrscheinlichkeiten sind zu reduzieren?*** (Manuskript, zur Veröffentlichung eingereicht). Bonn: psychologisches Institut der Universität.

Erdfelder, E. (1993b). ***Entscheidungen über statistische und wissenschaftliche Hypothesen: Konsequenzen des Validitätspostulats.*** (Manuskript, zur Veröffentlichung eingereicht). Bonn: Psychologisches Institut der Universität.

Estes, W. K. (1956). The problem of inference from curves based on groups data. ***Psychological Bulletin, 53,*** 134-140.

Falmagne, J. C. (1979). On a class of probabilistic conjoint measurement models: Some diagnostic properties. ***Journal of Mathematical Psychology, 19, 73-88.***

Falmagne, J. C., Iverson, G. J. & Marcovici, S. (1979). Binaura1 loudness summation: Probabilistic theory and data. ***Psychological Review,*** *86, 25-43.*

Faul, F., Erdfelder, E. & Buchner, A. (1993). ***GPOWER: A general power analysis program.*** (Manuskript, zur Veröffentlichung eingereicht). Bonn: Psychologisches Institut der Universität.

Festinger, L. & Carlsmith, J.M. (1959). Cognitive consequences of forced compliance. ***Journal*** *of* ***Abnormal and Social Psychology,*** *58, S.203-210.*

Gabriel, K. R. & Hsu, C.-F (1983). Evaluation of the power of revandomization tests, with application to weather modification experiments. ***Journal*** *of the* ***American Statistical Association,*** *78, 766-775.*

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

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

Green, D.M. & Swets, J.A. (1974). ***Signal detection theory and psychophysics. New*** York: Krieger.

Groeben, N. & Westmeyer, H. ($1981^2$). ***Kriterien psychologischer Forschung.*** München: Juventa.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen der Psychologie. In ***G.*** Lüer (Hrsg.), ***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. & Westermann, R. (1983a). Entscheidung über statistische und wissenschaftliche Hypothesen: Probleme bei mehrfachen Signifikanttests zur Prüfung ***einer*** wissenschaftlichen Hypothese. ***Zeitschrift*** *für* ***Sozialpsychologie,*** *14,* 106-117.

Hager, W. & Westermann, R. (I983b). Zur Wahl und Prüfung statistischer Hypothesen in psychologischen Untersuchungen. ***Zeitschrift*** *für* ***experimentelle und angewandte Psychologie,*** *30, 67-94.*

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

Herrmann, ***T.*** (1973). ***Persönlichkeitsmerkmale. Bestimmung und Verwendung in der psychologischen Wissenschaft.*** Stuttgart: Kohlhammer.

Jöreskog, K.G. & Sörbom, D. (1988). ***LZSREL*** *7.* ***A*** guide ***to the program and its applications.*** Chicago: SPSS Inc.

Johnson, ***N.*** L. & Kotz, ***S.*** (1970). ***Distributions in statistics. Continuous*** univariute ***distributions*** *- 1.* New York: Wiley.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (eds.), ***Criticism and the growth*** *of* ***knowledge*** (pp. 91-195). Cambridge: University Press.

Lazarsfeld, P. F. & Henry, ***N.*** W. (1968). ***Latent*** *structure* ***analysis.*** Boston: Houghton Mifflin.

Lienert, ***G.*** A. ***(1973). Verteilungsfreie Methoden der Biostatistik, Band*** 1. Meisenheim am Glan: Hain.

Luce, R. D. (1986). ***Response times. Their role in inferring elementary mental organization.*** New York: Oxford University Press.

Luce, R.D., Krantz, D.H., Suppes, P. & Tversky, A.P. (1990). ***Foundations of*** measurement. Volume ***III: Representation, axiomatization, and invariance.*** San Diego: Academic Press.

Popper, K. R. (1982[7]). ***Logik der Forschung.*** Tübingen: Mohr.

Riefer, D. M. & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. ***Psychological Review, 91,*** 318-339.

Rost, J. & Langeheine, J. (1991). Mischverteilungsmodelle: die Methodologie der kommenden Jahre. In D. Frey (Hrsg.), ***Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990, Band*** 2 (S. 622-626). Göttingen: Hogrefe.

Sievers, W. (1990). Bootstrap-Konfidenzintervalle und Bootstrap-Akzeptanz-Bereiche hypothesenprüfender Verfahren. ***Zeitschrift für experimentelle und angewandte Psychologie, 37,*** 85-123.

Snodgrass, J. G. & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. ***Journal of Experimental Psychology: General, 117, 34-50.***

Stegmüller, W. ***(1973). Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band 4: Personelle*** und ***statistische Wahrscheinlichkeit. 2. Halbband: Statistisches Schließen.*** Berlin: Springer.

Steyer, R. (1988). Conditional expectations: An introduction to the concept and its applications in empirical sciences. ***Methodika, 2, 53-78.***

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

Tack, W. H. (1976). ***Stochastische Lernmodelle.*** Stuttgart: Kohlhammer.

Titterington, D.M., Smith, A. F. M. & Makov, U.E. (1985). ***Statistical analysis of finite mixture distributions.*** New York: Wiley.

Townsend, J. T. & Ashby, F. ***G.*** (1983). ***Stochastic modeling of elementary psychological processes.*** Cambridge: Cambridge University Press.

Wald, A. (1948). ***Sequential analysis. New*** York: Wiley.

Westermann, R. (1987). Wissenschaftstheoretische Grundlagen der experimentellen Psychologie. In ***G.*** Lüer (Hrsg.), ***Allgemeine Experimentelle Psychologie (S. 5-42).*** Stuttgart: Fischer.

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

Westmeyer, H. (1972). ***Logik der Diagnostik.*** Stuttgart: Kohlhammer.

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

Wickens, T.D. (1982). ***Models for behavior. Stochastic processes in psychology.*** San Francisco: Freeman.

Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. ***Annals of Mathematical Statistics, 9, 60-62.***

Willmes, K. (1987). ***Beiträge zu Theorie** und **Anwendung von Permutationstests in der uni- und multivariaten Datenanalyse.*** (Unveröffentlichte Dissertation). Trier: Fachbereich 1 - Psychologie der Universität.

## *Autorenhinweis*

Die Verfasser danken den Mitautoren dieses Bandes sowie den Herren Dr. J. Andres und Prof. Dr. A. Iseler für kritische Kommentare zur Erstfassung des vorliegenden Kapitels. Lisa Irmen hat das Manuskript korrekturgelesen und Formulierungsänderungen vorgeschlagen, die wir größtenteils übernommen haben. Auch ihr sei herzlich gedankt.