

Schewe, Gerhard

**Working Paper — Digitized Version**

## Nicht-Befund-Analyse: Ein Konzept zur Re-Analyse empirischer Forschungsarbeiten

Manuskripte aus den Instituten für Betriebswirtschaftslehre der Universität Kiel, No. 369

**Provided in Cooperation with:**

Christian-Albrechts-University of Kiel, Institute of Business Administration

*Suggested Citation:* Schewe, Gerhard (1995) : Nicht-Befund-Analyse: Ein Konzept zur Re-Analyse empirischer Forschungsarbeiten, Manuskripte aus den Instituten für Betriebswirtschaftslehre der Universität Kiel, No. 369, Universität Kiel, Institut für Betriebswirtschaftslehre, Kiel

This Version is available at:

<https://hdl.handle.net/10419/181058>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

Nr. 369

Gerhard Schewe

**Nicht-Befund-Analyse:  
Ein Konzept zur Re-Analyse empirischer  
Forschungsarbeiten**

Kiel, März 1995

## Zusammenfassung

In der empirischen Forschung läßt sich die Beobachtung machen, daß bei der Vielzahl der veröffentlichten empirischen Ergebnissen diejenigen überwiegen, die eine Forschungshypothese mit einer a priori vereinbarten Irrtumswahrscheinlichkeit beibehalten. Nicht signifikante Ergebnisse werden nur unzureichend interpretiert. Die in diesem Beitrag entwickelte Nicht-Befund-Analyse soll dazu beitragen, auch empirische Ergebnisse, die ohne Befund bleiben, einer inhaltlichen Würdigung zugänglich zu machen. Dabei gilt es zum einen Artefaktgefahren abzuschätzen, die insbesondere für einen Nicht-Befund von Bedeutung sind und zum anderen Möglichkeiten aufzuzeigen, wie ein Nicht-Befund auch von einem externen Analytiker einer Forschungsarbeit zu verarbeiten und anschließend zu interpretieren ist.

## 1. Problemstellung und Begriffsabgrenzung

Die empirische Forschung ist heutzutage fester Bestandteil der Betriebswirtschaftslehre. Kein Teilgebiet des Faches ist ohne sie denkbar. Konsequenterweise ist zwischenzeitlich eine Flut empirischer betriebswirtschaftlicher Forschungsergebnisse zu verzeichnen. Interessant ist in diesem Zusammenhang eine Zahl, die Hauschildt/Grün<sup>1)</sup> nennen. Sie analysierten für den Zeitraum 1972 bis 1992 800 englischsprachige Zeitschriften dahingehend, ob in ihnen empirische Ergebnisse vorgelegt oder diskutiert wurden. Sie konnten allein für diese Zeitspanne ca. 20000 Artikel identifizieren; eine sicherlich beeindruckende Zahl und Vielfalt. Insofern scheint es angezeigt, neben der Durchführung weiterer empirischer Studien sich auch einmal der Frage zuzuwenden, wie sich die Vielzahl der bisher gesammelten empirischen Ergebnisse verdichten und ordnen lassen, um so die Übersicht über ein Fach zu bekommen, das scheinbar in der Fülle seiner Forschungsergebnisse zu versinken droht.

Entsprechend ihres Resultats bei der Prüfung bestimmter hypothetischer Zusammenhänge, Abhängigkeiten und Unterschiede lassen sich *empirische Ergebnisse* anhand zweier Untergruppen klassifizieren: Zum einen diejenigen, die die Beibehaltung der Forschungshypothese erlauben - im folgenden als *empirischer Befund* bezeichnet - und zum anderen diejenigen, die eine Ablehnung der Hypothese erfordern - im folgenden als *empirischer Nicht-Befund* bezeichnet. *Ein Nicht-Befund liegt damit immer dann vor, wenn zwischen Variablen eine hypothetisch unterstellte Beziehung als nicht signifikant klassifiziert wird.* Als Grenze der Signifikanzprüfung wollen wir hier den allgemein üblichen Wert von 5%  $\alpha$

---

1) Vgl. Hauschildt/Grün (1993) S. X.

Fehlerwahrscheinlichkeit vereinbaren.<sup>1)</sup> Wie wir noch zeigen werden, wird den beiden Gruppen in der Literatur in unterschiedlich starkem Maße Aufmerksamkeit bei der theoretischen Würdigung geschenkt. Während Befunde in vielfältiger Weise interpretiert werden, sind Nicht-Befunde oftmals dadurch gekennzeichnet, daß sie - wenn überhaupt beachtet - nur oberflächlich und knapp gewürdigt werden.

An diesem Problem setzt der vorliegende Beitrag an. Es wird ein Konzept erarbeitet, welches es insbesondere dem externen Analytiker einer Forschungsarbeit ermöglichen soll, empirische Ergebnisse, die ohne Befund bleiben, über Studien hinweg zu analysieren und zu interpretieren. Ziel ist dabei das Erkennen und Abschätzen möglicher Forschungsartefakte sowie eine Kompensation von zu erwartenden Interpretationsdefiziten speziell bei Nicht-Befunden.

## 2. Beweggründe für eine Re-Analyse von Nicht-Befunden

Die Nicht-Befund-Analyse wollen wir im folgenden aus der Sicht des externen Analytikers einer veröffentlichten Forschungsarbeit vornehmen. Diesem fällt als erstes die Tatsache auf, daß ein bestimmtes empirisches Ergebnis in einer wissenschaftlichen Zeitschrift veröffentlicht wurde. Sodann fällt sein Blick auf die Qualität der Datenauswertung und -interpretation. Schließlich erfolgt noch eine kritische Überprüfung der Art und Weise der Datengewinnung. Ein derartiger Prozeß der Evaluation empirischer Ergebnisse wird sich meist auf diejenigen richten, die eine Aufrechterhaltung einer Forschungshypothese bei einer a priori fixierten

---

1) Streng genommen müßte auch die  $\beta$ -Fehlerwahrscheinlichkeit bei der Nicht-Befund-Klassifikation beachtet werden. Sie gibt an, mit welcher Wahrscheinlichkeit eine Nullhypothese - also der Nicht-Befund - irrtümlicherweise beibehalten wird. Um den  $\beta$ -Fehler jedoch zu berechnen, benötigt man gesicherte Annahmen über eine kritische Mindesteffektstärke zwischen den untersuchten Merkmalen, ab der von einer Beziehung zwischen ihnen gesprochen werden kann. Liegt diese vor, kann ein kritischer Stichprobenumfang berechnet werden, ab dessen überschreiten eine vereinbarte Grenze der  $\beta$ -Fehlerwahrscheinlichkeit unterschritten wird. Bei zu geringen Stichprobenumfängen wäre damit eine Nicht-Befund-Interpretation nicht sinnvoll. Leider ist es vielfach unmöglich, eine Mindesteffektstärke zwischen bestimmten Merkmalen a priori festzulegen. In einem derartigen Fall ist dann auch die Berechnung des  $\beta$ -Fehlers nicht möglich.

Grenze der Irrtumswahrscheinlichkeit nahelegen. Nicht signifikante Ergebnisse werden nicht näher betrachtet.

Die nachfolgenden Ausführungen sollen nun zeigen, daß es durchaus Gründe gibt, die es notwendig erscheinen lassen, sich auch mit Nicht-Befunden näher zu befassen.

## 2.1. Analyse von Nicht-Befunden als Konsequenz des Publikations- und Forschungsverhaltens

Der "Zwang zur Befundpräsentation" bei empirischen Forschungsarbeiten ist in der sozialwissenschaftlichen Literatur vereinzelt bereits Gegenstand umfangreicher Forschungsarbeiten geworden. Im folgenden soll kurz auf die - aus der Sicht der Analyse von Nicht-Befunden - wichtigsten Ergebnisse eingegangen werden:

(i) Atkinson, Furlong und Wampold<sup>1)</sup> befragten 50 Mitglieder der Herausgeberbeiräte der Zeitschriften "Journal of Counseling Psychology" und "Journal of Consulting and Clinical Psychology". Sie wurden gebeten, Forschungs-Manuskripte zu bewerten, die sich nur hinsichtlich des Grades des Vorhandenseins von signifikanten Befunden unterschieden. Die theoretischen Überlegungen, die formale Präsentation usw. waren bei allen Manuskripten gleich. Diejenigen Personen, die die Manuskripte mit nicht signifikanten Befunden analysierten, schlugen in nur 39% der Fälle diese zur Publikation vor. Hingegen wurde bei Manuskripten mit signifikanten Befunden in 82% der Fälle eine Empfehlung zur Publikation ausgesprochen.

(ii) Von Woodman und Wayne<sup>2)</sup> wurden 50 empirische Untersuchungen, die sich mit Fragen der Organisationsentwicklung befaßten, begutachtet. Diese Artikel fanden ihre Veröffentlichung in den Jahren 1978 bis 1983 in den Zeitschriften "Journal of Applied Behavioral Science" und "Group

---

1) Vgl. Atkinson/Furlong/Wampold (1982) S. 189 ff.

2) Vgl. Woodman/Wayne (1985) S. 894 ff. sowie eigene Berechnungen. Wir haben für die bei Woodman/Wayne (1985) in Tab. 2 als Item (8) genannte Variable "significance level" die entsprechenden relativen Häufigkeiten berechnet.

& Organization Studies". Es zeigte sich, daß bei 12 von 50 Studien auch nicht signifikante Ergebnisse berichtet wurden, bei ca. 75% der Untersuchungen aber ausschließlich signifikante Befunde.

(iii) In ihrer Befragung von 301 Gutachtern bei 19 amerikanischen sozialwissenschaftlichen Zeitschriften kommen Kerr, Tolliver und Petree<sup>1)</sup> zu dem Ergebnis, daß ein Manuskript höchstwahrscheinlich dann zur Publikation angenommen wird, wenn es eine neue Theorie beinhaltet, starke signifikante empirische Ergebnisse liefert und von einer in diesem Forschungsgebiet bereits etablierten Forschungspersönlichkeit verfaßt wurde.

(iv) Campbell<sup>2)</sup> analysiert das Begutachtungsverhalten des "Journal of Applied Psychology" (JAP). Die Angemessenheit der Untersuchungsmethode wie die Relevanz der Forschungsfrage sind hier die entscheidenden Kriterien. Studien mit nicht signifikanten Befunden müssen diese schon in besonderer Weise und anspruchsvoll erklären, damit sie veröffentlicht werden. Dieser hohe theoretische Anspruch mag auch erklären, warum nur 10% der empirischen Ergebnisse der analysierten Studien nicht signifikant waren.

(v) Dieser erstaunliche Befund ist allerdings nicht neu. Bereits 1959 kommt Sterling in seiner Analyse empirischer Forschungsergebnisse zu dem Schluß:

"There is some evidence that in fields where statistical tests of significance are commonly used, research which yields nonsignificant results is not published. Such research being unknown to other investigators may be replaced independently until eventually by change a significant result occurs - an 'error of the first kind' - and is published."<sup>3)</sup>

Belegt wird diese These mit einer Analyse von 294 Signifikanztests, die 1955 in psychologischen Zeitschriften veröffentlicht wurden. Hiervon kam es in 286 Fällen zu einer Zurückweisung der Nullhypothese. In nur 8 Fällen wurde über nicht signifikante Ergebnisse berichtet.

---

1) Vgl. Kerr/Tolliver/Petree (1977) S. 135 ff.

2) Vgl. Campbell (1982) S. 699 f.

3) Sterling (1959) S. 30.

Der mehr oder minder manifeste Zwang zur Publikation und damit zur Präsentation signifikanter Befunde in Zusammenhang mit dem Vorhandensein immer anwendungsfreundlicherer statistischer Computer-Software, mag bei einigen empirischen Forschern dazu führen, daß sie der Versuchung erliegen, die Daten "solange zu quälen, bis sie gestehen", d.h. das Ergebnis letztendlich signifikant wird.<sup>1)</sup> Natürlich wird eine solche Vorgehensweise von der Literatur scharf kritisiert.<sup>2)</sup>

Das Wissen um mögliche Probleme bei der Datenauswertung läßt sich anhand einer einzelnen Studie meist nur schwer belegen. Es kann allenfalls dazu dienen, empirische Forschungsergebnisse mit der notwendigen Vorsicht zu interpretieren. Und genau diese Vorsicht bei der Analyse signifikanter Befunde führt dazu, sich auch den nicht signifikanten Befunden zuzuwenden. Je vorsichtiger signifikante Befunde zu beurteilen sind, desto relevanter wird der Nicht-Befund, d.h. die theoretische Würdigung des nicht signifikanten Ergebnisses. Der mangelnde qualitative Gehalt eines Befundes läßt auf eine stärkere Gültigkeit des Nicht-Befundes schließen, der lediglich entsprechend "auswertungstechnisch" verändert wurde, damit er den Charakter eines Befundes aufweist.

## 2.2. Analyse von Nicht-Befunden als Konsequenz der Kritik an der Durchführung von Signifikanzprüfungen

Ein Problem des Signifikanztests ergibt sich aus der inhaltlichen Würdigung der Tatsache der Ablehnung einer Nullhypothese zugunsten der alternativen Forschungshypothese. Ein Signifikanztest prüft lediglich die Gültigkeit der Nullhypothese. Der Schluß auf die alternative Forschungshypothese erfolgt indirekt. Es wird nichts über die Wahrscheinlichkeit der Forschungshypothese ausgesagt. Hierzu müßten die Verteilungsform wie auch bestimmte Parameter genau festgelegt sein, was im allgemeinen nur für die Nullhypothese möglich ist. Ihr liegt die Vorstellung eines reinen Zufallsexperiments zugrunde. Als Konsequenz hieraus ergibt sich,

---

1) Vgl. Truesdell (1984) S. 595 ff., wonach die Mittel (der Computer) in erster Linie empirischer Forschung zu bestimmen scheinen und nicht die Ziele (die zu prüfende Theorie).

2) Vgl. hierzu beispielhaft Opp (1972) S. 198 ff.

daß eine Zurückweisung der Nullhypothese nicht bedeutet, daß die in der Forschungshypothese behauptete Beziehung die einzige inhaltlich gültige und sinnvolle Beziehung darstellt. Eine Forschungshypothesen sollte einer theoretischen Fundierung entsprechen, deren Plausibilität bereits vor Durchführung der Signifikanzprüfung gezeigt wurde. Es ist somit Kriz zuzustimmen, wenn er hervorhebt, daß trotz der Wahrscheinlichkeitsprüfung der Nullhypothese, dem Forscher "nach wie vor die alleinige Verantwortung für den Sinn oder Unsinn seiner Arbeitshypothesen"<sup>1)</sup> obliegt.

Ein weiteres Problem des Signifikanztests ergibt sich aus der Abhängigkeit von Signifikanz und Stichprobengröße. Mit zunehmender Stichprobengröße verringert sich für einen bestimmten Stichproben-Kennwert die Irrtumswahrscheinlichkeit.<sup>2)</sup> "Bei großen Stichproben wird alles signifikant."<sup>3)</sup> Sehr geringe - unter Umständen sogar bedeutungslose - Effekte werden bei hinreichend großem Stichprobenumfang immerhin noch signifikant. Man muß sich insofern auf der einen Seite die Frage stellen, ob relativ niedrige Kennwerte, die allerdings signifikant sind, nur deshalb keinen Nicht-Befund darstellen, weil der Stichprobenumfang relativ groß ist? Auf der anderen Seite können aber auch geringe signifikante Effektstärken einer großen Stichprobe durchaus erhebliche praxeologische Relevanz besitzen. Es sei in diesem Zusammenhang auf Gage verwiesen, der dies am Beispiel der relativ niedrigen aber signifikanten Korrelation von 0,14 zwischen den Items "Rauchen" und "Erkrankung an Lungenkrebs" verdeutlicht.<sup>4)</sup>

Es stellt sich damit die grundlegende Frage, wo denn letztendlich eine Grenze zu ziehen ist, ab der ein empirisches Ergebnis als unbedeutend zu bezeichnen ist. Im Rahmen unserer Nicht-Befund-Analyse versuchen wir hierauf eine Antwort zu geben.

---

1) Kriz (1981) S. 138.

2) Vgl. Bortz (1989) S. 157 sowie ausführlich Meehl (1967) S. 109 ff.

3) Witte (1977) S. 299.

4) Vgl. Gage (1979) S. 12.

### 2.3. Analyse von Nicht-Befunden als Konsequenz des Auftretens von Forschungsartefakten

In der Literatur sind vielfältige Ursachen für das Entstehen von Forschungsartefakten bekannt: Zu nennen wären hier beispielhaft:

- Der sog. "Halo-Effekt", d.h. ein durch die Reihenfolge der Fragen verändertes Antwortverhalten der Befragten.
- Die Wahl sozial wünschenswerter Antworten durch die Befragten bzw. das Vorliegen bestimmter Antwortmuster.
- Die unterschiedliche Bereitschaft von Unternehmungspersonen an der Teilnahme einer empirischen Untersuchung. Hier insbesondere der "Return-Bias" bei postalisch verschickten Fragebögen.
- Protokollfehler des Befragenden in Richtung der gewünschten Hypothesenbestätigung.

Aus der Sicht des externen Analytikers fällt es generell schwer, Forschungsartefakte, die sich im Rahmen der Datenerhebung ergeben, zu erkennen. Soweit die Ursache für den Artefakt in der Untersuchungsperson zu sehen ist, ergibt sich meist nur die Möglichkeit, diesen über einen metaanalytischen Vergleich mehrerer Studien aufzudecken. D.h. es ist zu prüfen, ob ein artefaktverdächtiges Ergebnis sich auch in anderen Studien findet, bzw. Teilpopulationen auszugrenzen, für die ein empirisches Ergebnis kein Artefakt mehr darstellt.

Liegt die Artefaktursache eher in der Person des empirisch Forschenden, so gewinnt das empirische Ergebnis des Nicht-Befunds erhebliche Bedeutung, da es höchstwahrscheinlich in geringerem Ausmaß unter einem entsprechenden Artefaktverdacht steht als ein Befund. Das Streben nach Bestätigung der eigenen Forschungshypothesen wird eher zu Verzerrungen der Antwortwahrnehmung in Richtung eines Befunds führen, als in Richtung eines Nicht-Befunds. Wird dann ein Nicht-Befund auch noch publiziert, so müßte sich der externe Analytiker relativ sicher sein, daß ein Artefakt aufgrund der Person des Forschenden nicht vorliegt. Ein Befund, insbesondere wenn er sich durch eine relativ geringe Effektstärke auszeichnet, gibt da schon eher Anlaß zu einer Artefaktvermutung.

### 3. Vorgehensweise der Nicht-Befund-Analyse

#### 3.1. Grundlegender Aufbau der Analyse

Die vorangegangenen Ausführungen haben gezeigt, daß die gängige Praxis empirischer Forschung eindeutig signifikante Ergebnisse präferiert, obwohl dies stellenweise sehr problematisch ist und unter Umständen zu gravierenden Fehlinterpretationen führt. Es kann damit die Auffassung vertreten werden, daß auch die Analyse von Nicht-Befunden theoretische und praxeologische Relevanz besitzt, da hier in der Regel nicht in dem Maße Einfluß auf den Ausgang der statistischen Signifikanzprüfung genommen wird, wie bei der "Befundgenerierung". Mißt man dem Nicht-Befund einen ebenso hohen Stellenwert zu, wie dem signifikanten Befund, so erhebt sich die Frage nach der Vorgehensweise bei der Analyse von Nicht-Befunden. Ist hier in ähnlicher Weise zu verfahren wie bei der Befundinterpretation? D.h. sollte sich der externe Analytiker bereits bei Identifikation eines Nicht-Befundes auf den Weg machen, hierfür eine theoretische Erklärung zu suchen? Wir meinen, daß ein derartiger Ablauf bei der Nicht-Befund-Analyse nicht angebracht sei, vielmehr sollten vorher eine Vielzahl von Prüfungen durchgeführt werden, die es verhindern, daß mit der Nicht-Befund-Interpretation die Würdigung eines Artefaktes einhergeht. Die Nicht-Befund-Analyse ist in unseren Augen damit von der Befundanalyse und Befundinterpretation nicht nur inhaltlich, sondern auch in der Vorgehensweise zu trennen. Wir führen dies auf den allgemein üblichen Ablauf empirischer Forschung zurück. In der Regel wird aus der Literatur oder aus Einzelbeobachtungen eine Theorie zumindestens aber ein theoretischer Bezugsrahmen abgeleitet, dessen Operationalisierung dann die Grundlage für die Bildung und Prüfung von Forschungshypothesen liefert. Ein sich dabei ergebender Befund steht allgemein im Einklang mit den a priori durchgeführten theoretischen Überlegungen. Er fügt sich in das Bild, welches die Literatur vorzeichnet. Man kann auf Erklärungen zurückgreifen, die nicht nur Ergebnis der Überlegungen desjenigen sind, der eine empirische Untersuchung durchgeführt hat. Will man hingegen einen Nicht-Befund interpretieren, ergibt sich eine gänzlich andere Situation. Unterstellt man wieder einen wie oben skizzierten Ablauf empirischer Forschung, so steht der Nicht-Befund im Gegensatz zum theoretischen Bezugsrahmen, der Meinung der Literatur

oder getätigten Einzelbeobachtungen. Vordergründig muß insbesondere der externe Analytiker darauf gefaßt sein, daß eine Nicht-Befund-Interpretation unter Umständen einer Artefaktinterpretation gleichkommt. Folglich sollte unserer Überzeugung nach eine Nicht-Befund-Interpretation erst dann erfolgen, wenn sichergestellt ist, daß es sich beim vorliegenden empirischen Ergebnis höchstwahrscheinlich nicht um ein Forschungsartefakt handelt.

Der von uns im folgenden skizzierte Ablaufalgorithmus zur Prüfung eines empirischen Forschungsergebnisses trägt diesen Überlegungen Rechnung. Zur besseren Übersicht haben wir ihn in einzelne Teile zerlegt. Den Zusammenhang der einzelnen Prüfmodule gibt die nachfolgende Abbildung wieder:

*Abb. 1: Modularer Aufbau der Nicht-Befund-Analyse*  
- hier Abb. 1 einfügen -

## 3.2. Ablauf der Nicht-Befund-Analyse

### 3.2.1. Modul A: "Hypothesentest"

Aufgabe des Moduls "Hypothesentest" ist zum einen die Trennung empirischer Ergebnisse danach, ob sie aufgrund direkter Messung in der jeweiligen Stichprobe zustande kommen oder ob sie das Ergebnis einer Zusammenfassung bestimmter direkt beobachteter Variablen zu einem Merkmalskonstrukt sind. Zum anderen soll der Einfluß möglicher Verzerrungen im Prozeß der Datenerhebung bzw. Merkmalsmessung abgeschätzt werden. Abbildung 2 gibt das Flußdiagramm des Moduls "Hypothesentest" wieder:

*Abb. 2: Modul A: "Hypothesentest"*  
- hier Abb. 2 einfügen -

Grundlegende Annahme unseres Prüfalgorithmus ist das Vorliegen eines Hypothesentests. D.h. dem externen Analytiker liegt eine veröffentlichte

Studie vor, die eine Hypothese anhand eines bestimmten Datenmaterials prüft. Der externe Analytiker hat dabei a priori noch keine Wertung hinsichtlich der Qualität und Relevanz der abgeleiteten Hypothese sowie der Angemessenheit der untersuchten Stichprobe vorgenommen. Entsprechende Abschätzungen erfolgen - soweit dies dem externen Analytiker überhaupt möglich ist - erst im Laufe des Prüfalgorithmus. Ausgeschlossen ist damit der Fall einer axiomatischen Prüfung. Möglich ist hingegen, daß die Prüfung nicht mittels eines statistischen Testverfahrens erfolgt, sondern allein aufgrund von Plausibilitätsüberlegungen vorgenommen wird.

Eine zweite zentrale Entscheidung ist im Rahmen dieses Moduls zu treffen: Handelt es sich bei dem zu analysierenden empirischen Ergebnis um eines, welches sich auf Einzelmerkmale bezieht oder aber auf ein oder mehrere Merkmalskonstrukte? (Entscheidung A<sub>2</sub>) Je nach Art der Bildung der Hypothesenvariablen sind unterschiedlich detaillierte Prüfungen vorzunehmen.

Besteht der Verdacht, daß bereits bei der Messung der einzelnen Hypothesenvariablen Probleme aufgetreten sind, die höchstwahrscheinlich zu einer derartigen Verzerrung der Messung geführt haben, daß die sich anschließende Hypothesenprüfung u.U. zu einem Forschungsartefakt führt, so sollte generell Abstand davon genommen werden, eine entsprechende Würdigung des empirischen Hypothesenprüfergebnisses vorzunehmen. Derartige Effekte sind meist nur sehr schwer durch einen externen Analytiker abzuschätzen. Er ist mit seinem Urteil in erheblichem Maße von der Qualität der publizierten Daten abhängig. Insofern muß hier in erster Linie an die Forschungsverantwortung des empirisch Forschenden appelliert werden. Dem externen Analytiker ist es in der Regel nur möglich abzuschätzen, inwieweit eventuell ein Bias in der Antwortbereitschaft vorliegt. Die Literatur liefert hierfür bereits Anhaltspunkte. So haben beispielsweise Untersuchungen gezeigt, daß Unternehmungen sehr viel eher bereit sind, über von ihnen erfolgreich abgeschlossene Projekte zu berichten als über erfolglose.<sup>1)</sup>

---

1) Vgl. Myers/Marquis (1969) S. 11, Baker (1979) S. 346 und Gold/Rosegger/Boylan (1980) S. 56.

### 3.2.2. Modul B: "Merkmalskonstrukt"

Zielsetzung dieses Moduls ist in erster Linie das Erkennen möglicher Artefakte, die daraus resultieren, daß einzelne Merkmale zu Merkmalskonstrukten kombiniert wurden. Die Ableitung einzelner Hypothesen bezieht sich vielfach auf nicht direkt beobachtbare sogenannte latente Variablen. Immer dann, wenn mehrere Merkmale zur Messung herangezogen werden, wird meist auch eine Kombination dieser Merkmale zu einem einzigen Meßwert vorgenommen. Dem empirischen Forscher eröffnen sich dabei meist vielfältige Eingriffsmöglichkeiten, je nach individueller Zielsetzung. Unterstellt man ein Bestreben, wonach es wünschenswert erscheint, einen signifikanten Hypothesenbefund zu ermitteln, so mag dies dazu führen, daß verstärkt nach derjenigen Kombinationsmöglichkeit gesucht wird, die dieses Ziel auch erreicht. Eine Prüfung auf mögliche Artefakte erscheint damit insbesondere beim Vorliegen von Merkmalskonstrukten zwingend notwendig. Eine solche Prüfung erstreckt sich sowohl auf Nicht-Befunde als auch auf Befunde.

Liegt ein Merkmalskonstrukt vor, so analysiert der Prüfalgorithmus im wesentlichen Problemaspekte, die möglicherweise verantwortlich sind für die Entstehung eines Forschungsartefaktes:

- *Die Art der Konstruktbildung:* Wird sie subjektiv vom empirisch Forschenden durchgeführt oder liegen ihr "objektivierbare" Kriterien zugrunde? Ist aus der Sicht des externen Analytikers auch eine alternative Kombination der direkt gemessenen Variablen plausibel und durchführbar?
- *Die Stabilität des empirischen Ergebnisses vor dem Hintergrund der Konstruktbildung:* Hier interessieren insbesondere Fragestellungen, die die Ursprungsvariablen des Konstrukts betreffen. Sind bei ihnen gegenläufige Beziehungen feststellbar bzw. lassen sich Ausreißer erkennen, die u.U. die Stabilität des Merkmalskonstruktes und damit das Ergebnis des Hypothesentests beeinflussen. Aber auch die Stabilität der Hypothesenprüfung bei alternativer Merkmalskombination gilt es zu analysieren.

Abbildung 3 zeigt das Voranschreiten des Moduls "Merkmalskonstrukt".

**Abb. 3: Modul B: "Merkmalskonstrukt"**

- hier Abb. 3 einfügen -

Von der Vielzahl der Einzelentscheidungen erfordern hier einige eine detailliertere Erläuterung:

(i) Unter einer subjektiven Kombination der Merkmale wird eine Vorgehensweise verstanden, bei der der empirisch Forschende selbst die Art der Merkmalskombination bestimmt. Insbesondere sämtliche Varianten von Scoring-Modellen lassen sich hierunter fassen. Meist wird dabei versucht, eine subjektiv "gemessene" ordinale Rangreihe kardinal abzubilden - eine sicherlich manchmal sehr willkürliche Vorgehensweise.

Eine scheinbar objektive Merkmalskombination liegt vor, wenn die unterschiedlichen Dimensionen eines nicht direkt beobachtbaren Merkmals mit Hilfe eines multivariaten, statistischen Verfahren zu einer latenten Variablen zusammengefaßt werden. Am häufigsten wird sich hier wohl eine Zusammenfassung der Meßvariablen entsprechend ihrer korrelativen Zusammenhänge finden - die unterschiedlichsten Verfahren der Faktoranalyse. Wenn die Ergebnisse dieser Methoden als nicht-subjektive Merkmalskonstrukte bezeichnet werden, so darf nicht verkannt werden, daß auch bei ihnen der Forscher subjektive Entscheidungen zu fällen hat. Dies wird meist schon zu Beginn der Analyse der Fall sein, wenn entschieden wird, auf welche direkt gemessenen Merkmale ein solches Verfahren angewandt wird oder aber welche Algorithmen der Kombination zugrunde liegen.

(ii) In einer nächsten Prüfstufe, ist die Stabilität der Merkmalskonstruktion und des sich anschließenden Hypothesenprüfergebnisses zu testen. Lassen sich andere als die gewählte Form des Merkmalskonstruktion ebenfalls theoretisch rechtfertigen und praktisch durchführen, so besteht die Möglichkeit, daß sich das zu analysierende Hypothesenprüfergebnis u.U. ändert. Eine solche Prüfung muß unabhängig davon erfolgen, ob die Bildung einer latenten Variablen allein aufgrund subjektiver Entscheidungen und Gewichtungen stattfindet oder aber mit Hilfe eines multivariaten statistischen Verfahrens.

Wir wollen bei einer Prüfung auf "Stabilität des Hypothesentests" nur dann von einer Instabilität des Ergebnisses sprechen, wenn sich die Richtung des beobachteten Effekts bei der Hypothesenprüfung ändert. Eine Änderung im Signifikanzniveau - es wird eine kritische Grenze der Irrtumswahrscheinlichkeit überschritten - soll nur dann als instabiler Befund bezeichnet werden, wenn sich gleichzeitig das Vorzeichen des Effekts ändert.

(iii) Ein weiteres Indiz für das mögliche Auftreten eines Artefaktes ergibt sich beim Vergleich der Ergebnisse des Merkmalskonstruktes mit denen der ursprünglich gemessenen Variablen. Ziel dieses Prüfprozesses soll es sein, eventuell gegenläufige Ergebnisse der Ursprungsvariablen zu erkennen, um so gegebenenfalls Ausreißervariablen zu identifizieren. Als konträres Ergebnis wollen wir auch hier nur solche Ergebnisse bezeichnen, bei denen sich das Ergebnis-Vorzeichen ändert, unabhängig davon, ob sich auch das Ergebnis der Signifikanzprüfung ändert.

### 3.3. Modul C: "Ergebnis eines Hypothesentests"

Das Modul der Abbildung 4 zeigt das zentrale Entscheidungsproblem der Klassifikation von empirischen Ergebnissen in solche mit und ohne Befund:

*Abb. 4: Modul C: "Ergebnis eines Hypothesentests"*

- hier Abb. 4 einfügen -

Aufgabe dieses Moduls ist es insbesondere, die folgenden Probleme abzuschätzen:

- Inwieweit gilt es, dem Einfluß von Drittvariablen Rechnung zu tragen?
- Reicht bei vergleichsweise geringen Effektstärken das Signifikanzniveau als alleiniges Kriterium der Nicht-Befund-Klassifikation aus?
- Liegt eine theoretisch gerechtfertigte ein- oder zweiseitige Signifikanzprüfung vor?
- Führt die Wahl der Stichprobe dazu, daß bei Hypothesenmerkmalen

eine Verringerung oder eine Erhöhung der Varianz wahrscheinlich ist?

### 3.3.1. Entscheidung $C_1$ , $C_7$ , $C_{13}$ : "Partial-Effekt"

Bevor ein Ergebnis - unabhängig davon, ob es sich um einen Befund oder Nicht-Befund handelt - eingehend interpretiert wird, muß geprüft werden, ob hier nicht eventuell ein "Scheineffekt" vorliegt, d.h. eine Korrelation beispielsweise nur deshalb auftritt, weil die beiden Merkmale von einem dritten Merkmal abhängen. Um die tatsächliche Stärke des Zusammenhangs feststellen zu können, muß folglich derjenige Anteil des Effekts herausgerechnet werden, der auf die Drittvariable zurückzuführen ist. Dies erfolgt üblicherweise mit Hilfe der Berechnung der entsprechenden Partialkorrelationen. Liegt eine Korrelationsmatrix über alle untersuchten Variablen vor, so ist die Berechnung solcher Partialkorrelationen vor dem Hintergrund unserer Analyse auch für den externen Analytiker unproblematisch.

### 3.3.2. Entscheidung $C_2$ , $C_6$ , $C_8$ , $C_{12}$ , $C_{14}$ , $C_{18}$ , $C_{12}$ : "Grenze der Signifikanzprüfung und Stärke des Effekts"

Der von uns entwickelte Algorithmus unterscheidet zwischen Nicht-Befunden im engeren und solchen im weiteren Sinne. Ein Nicht-Befund im engeren Sinne ergibt sich dann, wenn ein als kritisch definiertes Signifikanzniveau verletzt wird. Wie oben bereits ausgeführt, wählen wir hier eine Grenze von 5% Irrtumswahrscheinlichkeit ( $\alpha$ -Fehlerwahrscheinlichkeit). Ein Blick in die Literatur zeigt, daß diese Grenze allgemein bei der Signifikanzprüfung als noch zulässige Toleranz akzeptiert wird. Von einem Nicht-Befund im weiteren Sinne wollen wir sprechen, wenn zwar das Signifikanzniveau von 5% nicht überschritten wird, die Effektstärke aber sehr gering ist - bei Korrelationen z.B. unter  $r = |0,3|$  liegt.

Abschließend sei noch darauf verwiesen, daß für den Fall, daß eine Signifikanzprüfung in der zu analysierenden Studie nicht vorgenommen wurde, der externe Analytiker mit Hilfe einschlägiger Tabellenwerke in

der Lage ist, diese selbst vorzunehmen. Im Falle des Ausweisens korrelativer Zusammenhänge muß hierfür nur der Stichprobenumfang sowie die Stärke des Effekts bekannt sein. Die Abschätzung der  $\alpha$ -Fehlerwahrscheinlichkeit ist dann möglich. Ferner ergibt sich auch die Möglichkeit, eine Vielzahl statistischer Effekte in Korrelationskoeffizienten umzurechnen.<sup>1)</sup>

### 3.3.3. Entscheidung $C_3$ , $C_9$ , $C_{15}$ : "Richtung der Signifikanzprüfung"

Ausgangslage dieser Entscheidungen ist ein Hypothesenprüfergebnis, welches auf seine Gültigkeit bei einer 5%igen Irrtumswahrscheinlichkeit getestet wurde. Problematisch ist eine solche Prüfung dann, wenn obwohl eine ungerichtete Hypothese vorliegt, nur eine einseitige Signifikanzabschätzung vorgenommen wurde. Für den externen Analytiker kann es sich unter Umständen als unmöglich erweisen, nachträglich eine zweiseitige Abschätzung der Irrtumswahrscheinlichkeit vorzunehmen. Dies wird häufig dann der Fall sein, wenn die exakten Werte der Irrtumswahrscheinlichkeit nicht publiziert werden oder aber andere Kennzahlen nicht vorhanden sind, so daß mit den einschlägigen Tabellenkennwerten hier keine entsprechende Abschätzung vorgenommen werden kann.

### 3.3.4. Entscheidung $C_4$ , $C_{10}$ , $C_{16}$ , $C_{17}$ , $C_{19}$ : "Varianzerhöhung"

Die Problematik dieser Entscheidung sei am Beispiel des Produkt-Moment-Korrelationskoeffizienten erläutert. Der Korrelationskoeffizient ist abhängig von der Standardabweichung bzw. Varianz der beiden Merkmale. Je höher die Varianz, desto höher der Effekt.<sup>2)</sup> Durch "Manipulation" der Stichprobe ist es jetzt möglich, die Varianz der untersuchten

---

1) Vgl. hierzu ausführlich Kendall/Stuart (1967) S. 557 ff., Hays (1973) S. 683 f., Rosenthal/Rubin (1982), S. 166 ff., Glass/McGaw/Smith (1984) S. 148 ff. und Wolf (1986) S. 35.

2) Wir gehen in unserem Beispiel davon aus, daß für beide untersuchten Merkmale die Normalverteilungsannahme Gültigkeit besitzt.

Merkmale zu erhöhen.<sup>1)</sup> Hierzu wird bei einer oder sogar bei beiden Merkmalen auf diejenigen Untersuchungseinheiten verzichtet, bei denen die Merkmalsausprägung(en) in einem "mittleren" Bereich angesiedelt ist (sind). Berechnet man Korrelationen nur aus den Extremfällen einer Stichprobe, so wird der Effekt damit zwangsläufig stärker, als wenn alle Untersuchungseinheiten in die Berechnung eingehen.

In unserem Prüfalgorithmus wollen wir diesen Fall jedoch nicht gleich als Artefakt eingestufen. Grund hierfür ist der bereits oben angesprochene Zusammenhang zwischen Stichprobenumfang und  $\alpha$ -Fehlerwahrscheinlichkeit. Die Erhöhung der Effektstärke wird dadurch erreicht, daß lediglich Untersuchungseinheiten berücksichtigt werden, bei denen die untersuchten Merkmale möglichst extreme Merkmalsausprägungen annehmen - die Stichprobenvarianz der Merkmale wird erhöht. Damit ist nun jedoch aber zwangsläufig auch eine Reduktion des Stichprobenumfangs verbunden. Dies hat zur Folge, daß sich der  $\alpha$ -Fehler vergrößert. Das Ergebnis einer "manipulierten" Stichprobe ist damit ein vergleichsweise höherer Effekt als bei einer "nicht manipulierten" Stichprobe, aber ebenso auch eine relativ höhere  $\alpha$ -Fehlerwahrscheinlichkeit. Es ist damit nicht a priori auszuschließen, daß in einer "nicht manipulierten" Stichprobe der Hypothesentest zu einem nicht signifikanten Ergebnis führen würde.

### 3.3.5. Entscheidung $C_5$ , $C_{11}$ : "Varianzreduktion"

Wesentlich häufiger als das soeben angeführte Problem wird höchstwahrscheinlich der Fall auftreten, daß eine meist unbewußte "Manipulation" der einbezogenen Untersuchungseinheiten eine Varianzreduktion bei bestimmten Merkmalen zur Folge hat. Hypothesenprüfungen werden damit verstärkt zur Falsifikation der Forschungshypothese führen.

Die Analyse des Nicht-Befundes muß folglich berücksichtigen, daß die Wahl der Stichprobe eine Varianzreduktion bei bestimmten Merkmalen

---

1) Vgl. ausführlich zu den Möglichkeiten der "Stichprobenmanipulation" Stelzl (1982) S. 148 f. Vgl. ferner McCall (1970) S. 127 ff., Wendt (1976) S. 292 ff. und Bortz (1989) S. 267 ff., die ausführlich die Bildung verzerrter Korrelationen bei einer unvollständigen Stichprobenselektion diskutieren.

zur Folge haben kann, und man demzufolge nicht zweifelsfrei von einem Nicht-Befund sprechen kann. Wird die Nicht-Signifikanz des empirischen Ergebnisses als Indiz für die Ablehnung der Forschungshypothese gewertet, so würde damit ein Methodenartefakt zur Interpretation gelangen. Ist es nicht möglich, hier auf eine "nicht-manipulierte" Stichprobe zurückzugreifen, so ist die vorliegende Stichprobe ungeeignet, die entsprechende Forschungshypothese zu testen.

Der externe Analytiker muß äußerst sorgfältig die Beschreibung des empirischen Designs analysieren, um so Hinweise auf Verzerrungen der Merkmalsvarianzen zu erhalten. Zu denken ist dabei z.B.

- an eine "Größenbias", der sich immer dann ergibt, wenn die in der Stichprobe analysierten Unternehmungen einer bestimmten Größenklasse entstammen (z.B. Fortune's 500), aber auch
- an Verzerrungen, die sich aus Betrachtung einzelner Branchen ergeben können, wie z.B. der pharmazeutischen Industrie, deren Produktions- und Marktverhalten wesentlich stärker vom Gesetzgeber vorgeprägt ist, als das in anderen Industriezweigen der Fall ist.
- Aber auch die Beschränkung der empirischen Untersuchung auf bestimmte Zeiträume kann zu Verzerrungen führen, wenn diese durch Ereignisse gekennzeichnet sind, die den Untersuchungsgegenstand in maßgeblicher Weise beeinflussen. Zu denken wäre hier beispielsweise an die rezessiven Folgen des "Ölpreisschocks" Anfang der 70iger Jahre denkt.

Ein solcher Katalog ist leicht erweiterbar. Er soll dem externen Analytiker lediglich Anhaltspunkte liefern, die ihn davor schützen, einen Methodenartefakt als Nicht-Befund zu interpretieren.

### 3.4. Modul D: "Nicht-Befund im weiteren Sinne"

Wie oben gezeigt wurde, gibt es eine Art von empirischen Prüfergebnissen, die nicht eindeutig als Befund oder Nicht-Befund zu klassifizieren sind: Die sogenannten Nicht-Befunde im weiteren Sinne, also Ergebnisse, die sich zwar durch das Einhalten einer kritischen Grenze von Irrtumswahrscheinlichkeit auszeichnen, die auf der anderen Seite aber durch

eine sehr geringe Effektstärke gekennzeichnet sind. Ziel dieses Prüfalgorithmus ist es abzuschätzen, ob der Nicht-Befund im weiteren Sinne eher wie ein Befund oder aber eher wie ein Nicht-Befund - Nicht-Befund im engeren Sinne - zu behandeln ist. Diese Aufgabe ist für den externen Analytiker von erheblicher Bedeutung. Er muß erkennen, ob es sich bei dem veröffentlichten Hypothesenprüfergebnis um ein Ergebnis handelt, das möglicherweise auf "Signifikanz getrimmt" wurde, damit es eine vermeintlich größere Beachtung findet. Abbildung 5 zeigt das entsprechende Prüfmodul:

*Abb. 5: Modul D: "Nicht-Befund im weiteren Sinne"*

- hier Abb. 5 einfügen -

Das Modul "Nicht-Befund im weiteren Sinne" beinhaltet im wesentlichen nur eine Entscheidung. Diese steht in engem Zusammenhang mit der Abschätzung der sogenannten  $\beta$ -Fehlerwahrscheinlichkeit, welche unter Umständen eine Interpretation eines Nicht-Befundes verbietet.

Läßt sich ein  $\beta$ -Fehler berechnen, dann müssen auch Vorstellungen darüber existieren, wie hoch eine kritische Mindesteffektstärke sein, damit von einem Zusammenhang zwischen zwei Merkmalen gesprochen werden kann. Da wir im vorangegangenen Modul eine korrelative Effektstärke von  $|r| = 0,3$  als Klassifikationsgrenze vereinbart hatten, diese jedoch unabhängig ist von einer auf theoretischen Überlegungen zum jeweiligen Merkmalszusammenhang basierenden kritischen Mindesteffektstärke, gilt es jetzt noch den tatsächlich berechneten Effekt mit der eventuell vorhandenen Mindesteffektstärke zu vergleichen. Nur wenn dieser größer ist, wollen wir von einem Befund sprechen, den es dann zu interpretieren gilt. Ist der berechnete Effekt geringer als der Mindesteffekt, sind weitere Prüfungen im Rahmen des Moduls "Nicht-Befund im engeren Sinne" notwendig.

### 3.5. Modul E: "Nicht-Befund im engeren Sinne"

Das folgende Prüfmodul soll zeigen, daß auch bei nicht signifikanten empirischen Ergebnissen durchaus eine Interpretation möglich ist, wenn vorher entsprechende Prüfungen vorgenommen werden. Abbildung 6 gibt das Flußdiagramm des Moduls "Nicht-Befund im engeren Sinne" wieder.

*Abb. 6: Modul E: "Nicht-Befund im engeren Sinne"*

- hier Abb. 6 einfügen -

Auch hier sollen wieder einige Entscheidungen detaillierter erläutert werden.

#### 3.5.1. Entscheidung $E_1$ : "Stichprobengröße"

Im Zentrum dieser Entscheidung steht der Zusammenhang von Stichprobenumfang und  $\alpha$ -Fehlerwahrscheinlichkeit. Hintergrund hierfür ist der Umstand, daß eine Hypothesenprüfung unter Umständen an einer sehr kleinen Stichprobe vorgenommen wurde. In diesem Fall besteht die Gefahr, daß Effekte, die bereits eine erhebliche Stärke aufweisen, die kritische Grenze der  $\alpha$ -Fehlerwahrscheinlichkeit aber überschreiten. Ist es trotzdem zulässig, solche Ergebnisse als Nicht-Befund zu interpretieren, da sie unser eingangs gewähltes Kriterium - die Irrtumswahrscheinlichkeit bei Prüfung der Nullhypothese sei größer als 5% - erfüllen?

Wir meinen, daß zur Beantwortung dieser Frage eine differenzierte Vorgehensweise erforderlich ist. Die  $\alpha$ -Fehlerwahrscheinlichkeit darf bei sehr kleinen Stichproben nicht das alleinige Kriterium der Nicht-Befund-Klassifikation sein, es muß zusätzlich auch die ausgewiesene Effektstärke berücksichtigt werden. Es erhebt sich das Problem, ab welchem kritischen Umfang eine Stichprobe als zu klein anzusehen ist, damit die Nicht-Befund-Klassifikation ausschließlich aufgrund des Signifikanzniveaus des  $\alpha$ -Fehlers erfolgen kann. Hierbei müssen wir uns an der Effektstärke orientieren. Wir wählen die bereits oben zur Abgrenzung des Nicht-Befundes im weiteren Sinne gewählte korrelative Effektstärke von  $|r| = 0,3$ . Es gilt jetzt denjenigen Stichprobenumfang zu bestimmen, für dessen Unterschreitung ein Effekt von größer als  $|r| = 0,3$  trotz des Überschreitens

der  $\alpha$ -Fehlerwahrscheinlichkeit von 5% als Befund zu klassifizieren ist. Dieser Umfang läßt sich mit Hilfe einschlägiger Formeln berechnen:<sup>1)</sup> Bei einer einseitigen Signifikanzprüfung sind danach Stichproben mit weniger als 31 Untersuchungseinheiten und bei zweiseitiger Signifikanzprüfung mit weniger als 42 Untersuchungseinheiten als in unserem Sinne "zu klein" zu bezeichnen. Für Stichprobenumfänge, die unterhalb dieser Grenze liegen, wäre demnach eine Klassifikation als Nicht-Befunde ungeeignet, wenn der Effekt eine Stärke von größer als  $\pm 0,3$  annimmt.

### 3.5.2. Entscheidung $E_2, E_{10}, E_{11}, E_{12}, E_{15}, E_{16}$ : "Art des Zusammenhangs"

Betrachtet man empirische Hypothesenprüfergebnisse, so fällt auf, daß überwiegend lineare Beziehungen analysiert werden. Da wir für unsere Analyse in erster Linie Korrelationen betrachten, besteht hier natürlich eine erhebliche Gefahr der Artefaktbildung. Korrelationen geben lediglich an, ob zwischen zwei Merkmalen ein linearer Zusammenhang besteht. Problematisch wird dies insbesondere für die Analyse von Nicht-Befunden. Liegt ein Hypothesenprüfergebnis vor, welches sich durch eine  $\alpha$ -Fehlerwahrscheinlichkeit von größer als 5% auszeichnet, so ist dies nicht zwangsläufig dahingehend zu interpretieren, daß zwischen den beiden gemessenen Merkmalen keine Beziehung besteht. Es existiert lediglich kein linearer Zusammenhang. Durchaus möglich wäre allerdings ein u-förmiger oder umgekehrt u-förmiger Verlauf, beide ökonomisch beachtlich.

Im Vergleich zu einem Befund besteht bei einem Nicht-Befund wesentlich eher die Gefahr, daß hier ein Forschungsartefakt interpretiert wird, wenn eine nicht-lineare Beziehung vorliegt. Ein Befund würde wenigstens richtig erkennen, daß zwischen zwei Merkmalen eine Beziehung besteht; nur die Art der Beziehung wird nicht oder nur unzureichend erkannt.

Das Erkennen solcher Artefakte ist äußerst schwierig. Der externe Analytiker ist dabei entscheidend von der Publikationsgüte der vorliegenden Studie abhängig. Nur in Ausnahmefällen dürfte es der Fall sein, daß

---

1) Vgl. Glass/Hopkins (1984) S. 302.

Scatterplots oder ähnliche graphische Formen der Visualisierung der Merkmalsausprägungen in der untersuchten Studie veröffentlicht werden. Meist bleibt nur die Möglichkeit nach theoretischen Anhaltspunkten zu suchen, die gegen einen linearen Verlauf sprechen. Die folgenden Fragestellungen könnten hierfür eine Hilfe sein:

- Ist es plausibel, daß bestimmte Sättigungsgrenzen existieren, d.h. die lineare Beziehung nur für ein bestimmtes Intervall der Merkmalsbeziehung Gültigkeit besitzt?
- Besteht die Möglichkeit, daß sich die Richtung des Zusammenhangs umkehrt?
- Ergeben sich Anhaltspunkte dafür, daß die funktionale Beziehung zwischen den Merkmalen in bestimmten Intervallen unter- bzw. überproportional verläuft?

Für den Fall, daß sich Indizien für einen nicht-linearen Verlauf ergeben und eine nicht-lineare Prüfung nicht möglich ist, sollte insbesondere dann, wenn ein Nicht-Befund vorliegt, von dessen Interpretation abgesehen werden. Hier sieht man sich einem Artefakt gegenüber, welches nicht nur den Interpretationsspielraum einschränkt, sondern eine Interpretation auch geradezu verbietet.

### 3.6. Modul F: "Metaanalyse"

In unserem abschließenden Modul verlassen wir nun die Einzelbefund-Interpretation und wenden uns der vergleichenden Interpretation mit anderen empirischen Studien zu. Ein solcher Vergleich muß nicht auf Nicht-Befunde oder Befunde beschränkt sein, in der Regel wird eine integrative Gegenüberstellung von sämtlichen zu einem Hypothesenkomplex vorliegenden Befunden und Nicht-Befunden vorgenommen. Zielsetzung ist dabei die Errechnung, Analyse sowie Interpretation einer Effektstärke, die sich über sämtliche in die Metaanalyse einfließende Untersuchungsergebnisse erstreckt.

Unabhängig von der Berechnung der Effektstärke sind auch bei der Metaanalyse einige Entscheidungen zu treffen, die den Analytiker davor schützen sollen, hier einen Forschungsartefakt zu interpretieren. Die

nachfolgende Abbildung zeigt den entsprechenden Prüfalgorithmus.

*Abb. 7: Modul F: "Metaanalyse"*

- hier Abb. 7 einfügen -

Im weiteren soll auf zwei Entscheidungsbereiche näher eingegangen werden: Das Problem der Invarianz des metaanalytischen Befunds hinsichtlich des sog. "file drawer"-Problems<sup>1)</sup> und die Frage nach der Existenz möglicher moderierender Variablen.<sup>2)</sup> Derartige Fragestellungen bedürfen natürlich nur dann einer Beantwortung, wenn eine Metaanalyse überhaupt möglich ist, d.h. hinreichend viele Studien zu dem zu analysierenden Problembereich vorliegen.

### 3.6.1. Entscheidung F<sub>2</sub>: "File drawer"-Problem

Grundlage dieses Entscheidungsproblems ist die bereits oben ausführlich diskutierte Tatsache, daß sich in empirischen Studien ein deutliches Übergewicht von signifikant bestätigten empirischen Ergebnissen findet. Nicht-Befunde gelangen nur selten zur Publikation, sie wandern in die "Schublade" (file drawer). Für die Metaanalyse hat dieser Umstand gravierende Folgen. Er führt dazu, daß eine Vielzahl von Nicht-Befunden, obwohl sie existent sind, nicht in die Metaanalyse einfließen können, da sie nicht veröffentlicht wurden. Das generierte metaanalytische Ergebnis kann insofern verzerrt sein. Es muß Aufgabe des externen Analytikers sein, diese Verzerrung abzuschätzen. Konkret heißt dies: Wieviele zusätzliche Studien mit Nicht-Befunden müßten in die Metaanalyse einbezogen werden, bevor der metaanalytische Befund zum Nicht-Befund wird, das metaanalytische Ergebnis also eine kritische Irrtumswahrscheinlichkeit übersteigt? Die umgekehrte Abschätzung des metaanalytischen Ergebnisses ist entbehrlich, wenn wir oben genanntes Publikationsverhalten als gegeben unterstellen. Es ist dann nicht damit zu rechnen, daß Befunde in großer Zahl existent sind, die jedoch nicht veröffentlicht werden und deshalb nicht in eine Metaanalyse einfließen.

---

1) Vgl. Rosenthal (1979) S. 638 ff.

2) Vgl. hierzu erstmals Saunders (1956) S. 209 ff.

Um den Einfluß des "Schubladen-Bias" abschätzen zu können, werden in der Literatur unterschiedliche Methoden vorgeschlagen, je nach dem, welche Qualität das Datenmaterial besitzt.<sup>1)</sup>

### 3.6.2. Entscheidung F<sub>3</sub>, F<sub>4</sub>, F<sub>5</sub>, F<sub>6</sub>, F<sub>7</sub>: "Moderatoreinfluß"

Die in der Metaanalyse zusammengefaßten Einzelbefunde und Nicht-Befunde werden sich mit Sicherheit durch eine gewisse Varianz hinsichtlich ihrer Effektstärken auszeichnen. Im Rahmen der Metaanalyse muß nun geprüft werden, ob die Varianz mit Hilfe bestimmter - u.U. sogar mit nicht direkt gemessenen - Variablen zu erklären ist. Lassen sich also sogenannte "moderator variables" identifizieren?<sup>2)</sup>

Betrachtet man neuere Anwendungen der Metaanalysetechnik, so läßt sich feststellen, daß allgemein eine 75% bzw. 60%-Heuristik Anwendung findet.<sup>3)</sup> D.h., wenn 75% bzw. 60% der beobachteten Varianz der Einzelkorrelationen sich durch Stichprobenfehler, Meßfehler oder Rangunterschiede erklären lassen, dann ist eine Moderatorsuche nicht notwendig.

## 4. Fazit der Nicht-Befund-Analyse

Der von uns entwickelte Algorithmus zur Prüfung von empirischen Ergebnissen, die ohne Befund bleiben, soll in erster Linie dem externen Analytiker helfen abzuschätzen, ob der vorliegende Nicht-Befund einer inhaltlichen Interpretation zugänglich ist. Nur dann kann ein Interpretationsdefizit bei dieser Art empirischer Ergebnisse beseitigt werden, wie es sich aus dem allgemein verbreiteten Publikationsverhalten heraus ergibt.

Will man eine Interpretation eines Nicht-Befundes unabhängig von einer sich eventuell anschließenden Metaanalyse vornehmen, so könnte man

---

1) Vgl. hierzu ausführlich Rosenthal (1979) S. 640.

2) Vgl. hierzu die sehr frühe Arbeit von Saunders (1956) S. 209 ff. sowie Hunter/Schmidt/Jackson (1982) S. 47 ff.

3) Vgl. z.B. Petty/McGee/Cavender (1984) S. 714, Gooding/Wagner (1985) S. 467 und Damanpour (1992) S. 381.

sich an der folgenden Vorgehensweise orientieren:

Diese umfaßt nicht nur Erkenntnisse, die sich aus dem Modul "Nicht-Befund im engeren Sinne" ergeben haben, sondern auch solche aus vorgelegerten Prüfungen. Generell sollte eine Nicht-Befund-Interpretation - dies gilt übrigens auch für eine Befundinterpretation - in zwei Richtungen gehen: Zum einen gilt es, dem Nicht-Befund solche Erkenntnisse gegenüberzustellen, die andere wissenschaftliche Untersuchungen (empirischen oder theoretischen Charakters) liefern. Die Interpretation richtet den Blick nach außen. Sie löst sich von der vorliegenden Untersuchung. Zum anderen gilt es, den Gültigkeitsbereich des Nicht-Befundes möglichst scharf abzugrenzen, um hierauf aufbauend eigene theoretische und verallgemeinernde Erklärungen anzustellen. Die Interpretation richtet den Blick nach innen. Sie analysiert die vorliegende Untersuchung.

Will man den Gültigkeitsbereich des empirischen Ergebnisses bestimmen, so könnte man sich an dem folgenden Fragenkatalog orientieren:

- Werden spezielle definitorische Abgrenzungen bei den untersuchten Merkmalen gewählt, die im Gegensatz zum allgemeinen Verständnis in der Literatur eine eher enge oder eher weite Sichtweise zulassen?
- Lassen sich spezielle Merkmale der gewählten Stichprobe inhaltlich isolieren, die die Repräsentativität des empirischen Ergebnisses in gewisser Weise einschränken? Gibt es eigentümliche Charakteristika der untersuchten Analyseobjekte, die hierfür ein Indiz wären?
- Gibt es Einschränkungen der Gültigkeit aufgrund der gewählten Merkmalsoperationalisierung? Dies ist immer dann von besonderer Bedeutung, wenn es sich bei den untersuchten Merkmalen um nicht direkt beobachtbare Phänomene handelt.
- Besitzt der Zeitpunkt der Untersuchung einen Einfluß auf den Gültigkeitsbereich des empirischen Ergebnisses? Fand die Datenerhebung zu einem Zeitpunkt statt, bei welchem spezielle externe Gegebenheiten vorlagen?

Ist der Gültigkeitsbereich der untersuchten Hypothesenprüfung bestimmt, so gilt es, das Ergebnis inhaltlich zu würdigen. Dies hat sicherlich zuerst vor dem Hintergrund der Erkenntnisse der Literatur zu erfol-

gen. Dabei ist jedoch zu beachten, daß der bestimmte Gültigkeitsbereich der vorliegenden Studien sich mit demjenigen anderer Studien oder Ansätze decken muß, die zu einem solchen Vergleich herangezogen werden. Es ist zu fragen, ob die dort gewählten Erklärungen oder Schlußfolgerungen auch auf den zu interpretierenden Nicht-Befund Anwendung finden können. Fallen die Gültigkeitsbereiche von Studien auseinander, so ist zu prüfen, ob nicht auch dann noch inhaltliche Transfers möglich sind. In der Regel wird man jedoch derartige Fälle erst im Zuge einer metaanalytischen Prüfung eingehend behandeln, indem man versucht, die unterschiedlichen Gültigkeitsbereiche mit Hilfe eines Geflechts von Moderatorvariablen abzugrenzen.

Besonders schwierig gestaltet sich eine interpretatorische Würdigung eines Nicht-Befundes dann, wenn keine Analogien bei der Literatur gemacht werden können. Dies umso mehr, wenn eine a priori Ableitung einer Forschungshypothese erfolgte, die es nun aufgrund des Nicht-Befundes abzulehnen gilt. Hier hat eine Erforschung der Ursachen eines derartigen Ausgangs der Hypothesenprüfung zu erfolgen. Diese wird sich im wesentlichen an denjenigen Variablen zu orientieren haben, die bereits bei der Bestimmung des Gültigkeitsbereichs der Untersuchung identifiziert wurden. Letztendlich ist hierbei die Kreativität des Forschers gefragt. Das Aufzeigen möglicher Erklärungsmuster muß sich dabei immer der Frage stellen, welche Konsequenzen hieraus für den weiteren Forschungsprozeß zu ziehen sind. D.h. welche Fragen und Probleme bleiben offen bzw. verschließen sich einer Lösung. In welche Richtung hat damit zukünftige Forschung zu gehen? Läßt sich ein Nicht-Befund nur schwer theoretisch würdigen, so sollte er jedoch zumindest zum Anlaß genommen werden, hier weiterführende Fragen zu stellen. Eventuell ergeben sich daraus bereits Anhaltspunkte für eine umfassende nach Erklärungen suchende Metaanalyse, für deren Durchführung allerdings Grundvoraussetzung ist, daß bereits eine hinreichend große Zahl von empirischen Studien sich dem untersuchten Phänomen bereits zugewandt hat.

## Literatur:

- Atkinson, D.R.; Furlong, M.J. und Wampold, B.E.: Statistical Significance, Reviewer Evaluations, and the Scientific Process: Is there a (Statistically) Significant Relationship?, in: *Journal of Counseling Psychology* 29 (1982) S. 189-194.
- Baker, M.J.: Industrial Buying Behavior and the Adoption of Innovation, in: M.J. Baker (Hrsg.), *Industrial Innovation: Technology, Policy, Diffusion*, London 1979, S. 345-366.
- Berg, I.A.: *Response Set in Personality Assessment*, Chicago 1967.
- Bortz, J.: *Statistik für Sozialwissenschaftler*, 3. Auflage, Berlin u.a. 1989.
- Bungard, W. und Lück, H.E.: *Forschungsartefakte und nicht-reaktive Meßverfahren*, Stuttgart 1974.
- Campbell, J.P.: Editorial: Some Remarks from the Outgoing Editor, in: *Journal of Applied Psychology* 67 (1982) S. 691-700.
- Damanpour, F.: Organizational Size and Innovation, in: *Organization Studies* 13 (1992), S. 375-402.
- Gage, N.L.: *Unterrichten: Kunst oder Wissenschaft?*, München 1979.
- Glass, G.V.; McGaw, B. und Smith, M.C.: *Meta-Analysis in Social Research*, Beverly Hills 1981.
- Glass, G.V. und Hopkins, K.D.: *Statistical Methods in Education and Psychology*, 2. Auflage, Englewood Cliffs (NJ) 1984.
- Gold, B.; Rosegger, G.; Boylan, M.G.: *Evaluating Technological Innovations*, Lexington und Toronto 1980.
- Gooding, R.Z und Wagner, J.A.: A Meta-Analytic Review of the Relationship between Size and Performance, in: *Administrative Science Quarterly* 30 (1985), S. 462-481.
- Goudy, W.J.: Interim Response to a Mail Questionnaire: Impacts on Variable Relationships, in: *The Sociological Quarterly* 19 (1978), S. 253-265.
- Hauschildt, J. und Grün, O.: Auf dem Wege zu einer Realtheorie der Unternehmung, in: J. Hauschildt und O. Grün (Hrsg.), *Ergebnisse empirischer betriebswirtschaftlicher Forschung: Zu einer Realtheorie der Unternehmung*, Stuttgart 1993, S. IX-XII.
- Hays, W.L.: *Statistics for the Social Sciences*, 2. Auflage, New York 1973.
- Henkel, R.E.: *Test of Significance*, Beverly Hills und London 1976.
- Kendall, M.G. und Stuart, A.: *The Advanced Theory of Statistics, Band 2: Inference and Relationships*, 2. Auflage, London 1967.
- Kerr, S.; Tolliver, J. und Petree, D.: Manuscript Characteristics which Influence Acceptance for Management and Social Science Journals, in: *Academy of Management Journal* 20 (1977), S. 132 -141.
- Kriz, J.: *Methodenkritik empirischer Sozialforschung: Eine Problemanalyse Sozialwissenschaftlicher Forschungspraxis*, Stuttgart 1981.
- Meehl, P.E.: Theory Testing in Psychology and Physics: A Methodological Paradox, in: *Philosophy of Science* 34 (1967), S. 103-115.
- McCall, R.B.: *Fundamental Statistics for Psychology*, New York 1970.
- Myers, S. und Marquis, D.G.: *Successful Industrial Innovations. A Study of Factors Underlying Innovation in Selected Firms*, National Science Foundation, Washington 1969.
- Opp, K.D.: *Methodologie der Sozialwissenschaften*, 2. Auflage, Reinbeck bei Hamburg 1972.
- Overall, J.E.: Classical Statistical Hypotheses Testing within the Context of Bayesian Theory, in: *Psychological Bulletin* 71 (1969), S. 285-292.
- Petty, M.M.; McGee, G.W. und Cavender, J.W.: A Meta-Analysis of the Relationship between Individual Job Satisfaction and Individual Performance, in: *Academy of Management Review* 9 (1984), S. 712-721.
- Rosenthal, R.: The "File Drawer Problem" and Tolerance for Null Results, in: *Psy-*

- chological Bulletin 86 (1979), S. 638-641.
- Rosenthal, R. und Rubin, D.B.: A Simple, General Purpose Display of Magnitude of Experimental Effect, in: Journal of Educational Psychology 74 (1982), S. 166-169.
- Saunders, D.R.: Moderator Variables in Prediction, in: Educational Psychological Measurement 16 (1956), S. 209-222.
- Stelzl, I.: Fehler und Fallen der Statistik, Bern 1982.
- Sterling, T.D.: Publications Decisions and their Possible Effects on Inferences Drawn from Tests of Significance - or Vice Versa, in: Journal of the American Statistical Association 54 (1959), S. 30-34.
- Truesdell, C.: An Idiot's Fugitive Essays on Science: Methods, Criticism, Training, Circumstances, New York u.a. 1984.
- Wendt, H.W.: Superior Correlation, Revisited: A New Look at the Quantitative Outcomes of Sampling Heterogenous Groups and/or the Wrong Time, in: Archiv für Psychologie 128 (1976), S. 292-315.
- Witte, E.H.: Zur Logik und Anwendung der Inferenzstatistik, in: Psychologische Beiträge 19 (1977), S. 290-303
- Wolf, F.M.: Meta-Analysis: Quantitative Methods for Research Synthesis, Sage University Paper Series on Quantitative Applications in the Social Science, Series No. 07-059, Beverly Hills 1986:
- Woodman, R.W. und Wayne, L.E.: An Investigation of Positive-Findings Bias in Evaluation of Organization Development Interventions, in: Academy of Management Journal 28 (1985), S. 889-913.

Abb. 1: Modularer Aufbau der Nicht-Befund-Analyse

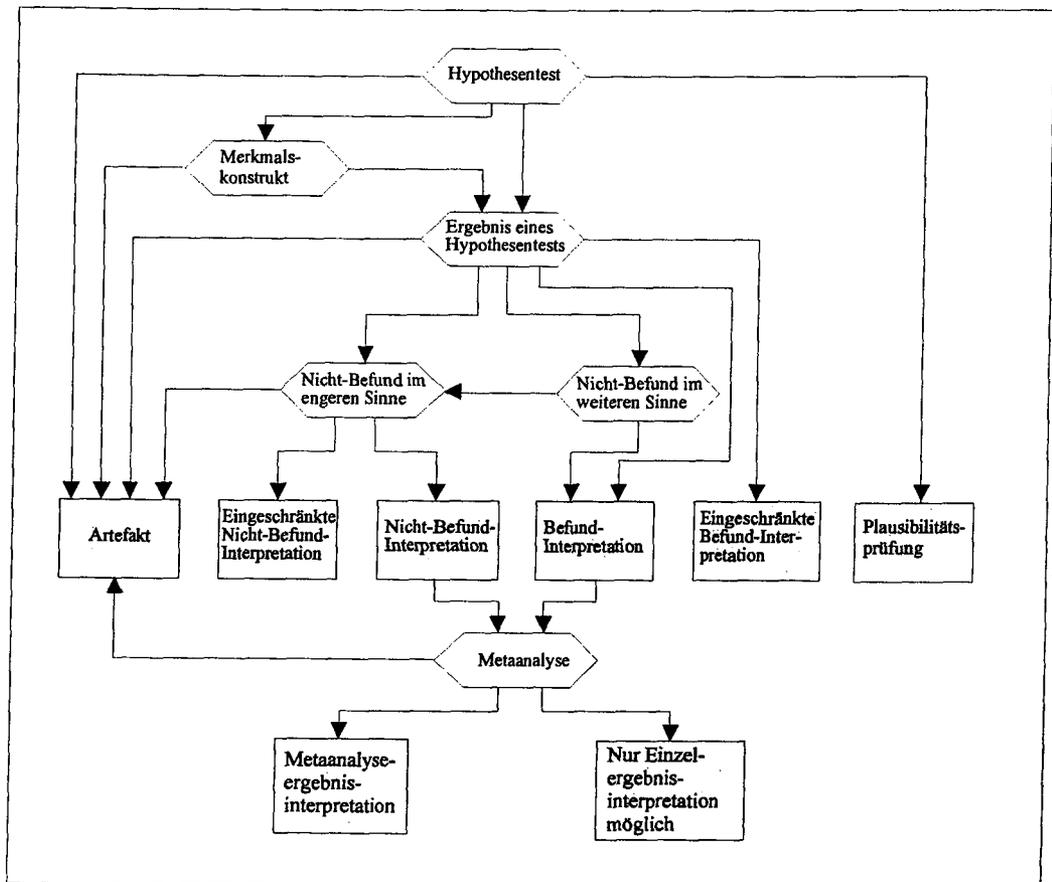


Abb. 2: Modul A: "Hypothesentest"

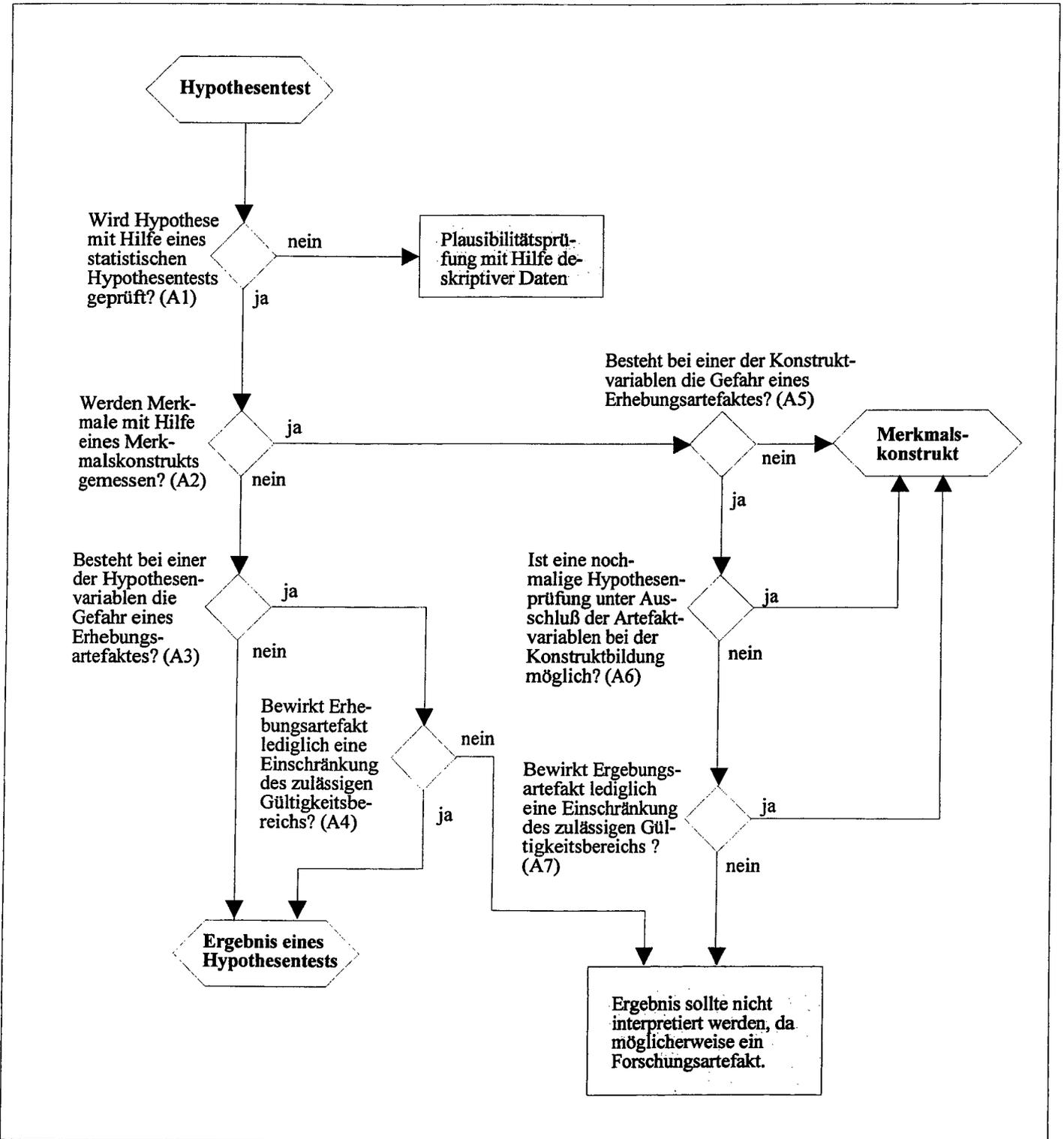


Abb. 3: Modul B: "Merkmalskonstrukt"

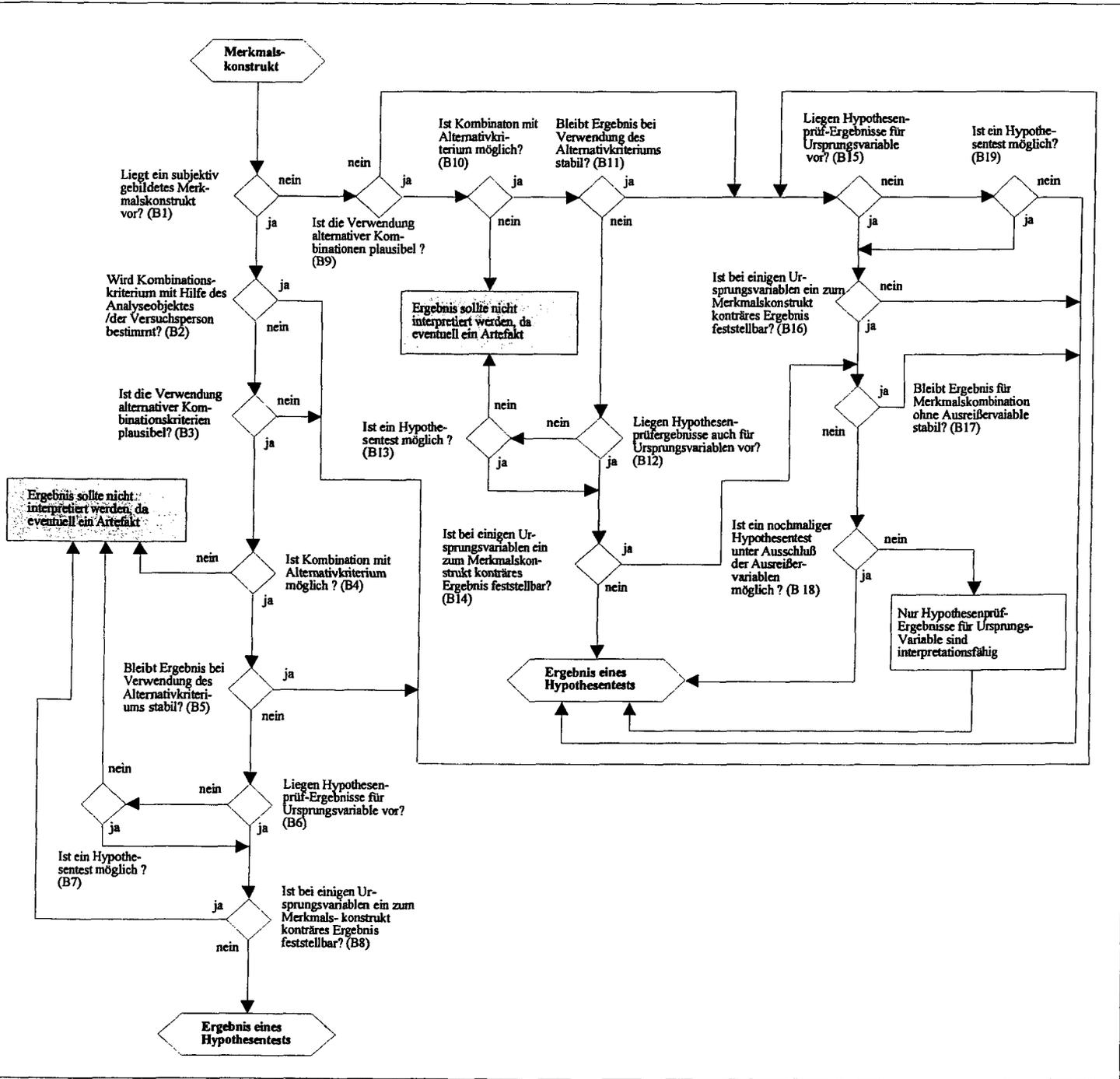


Abb. 4: Modul C: "Ergebnis eines Hypothesentests"

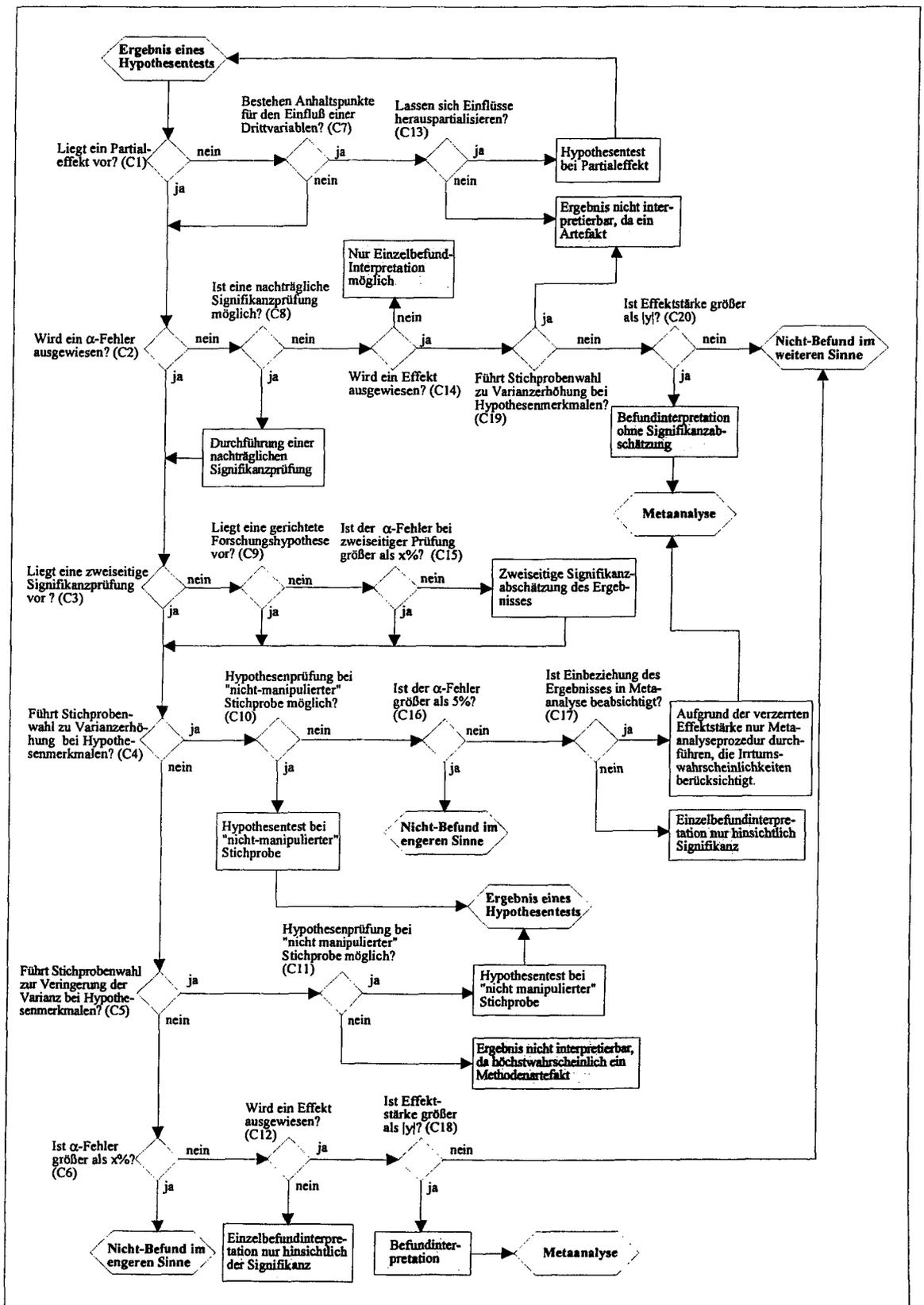


Abb. 5: Modul D: "Nicht-Befund im weiteren Sinne"

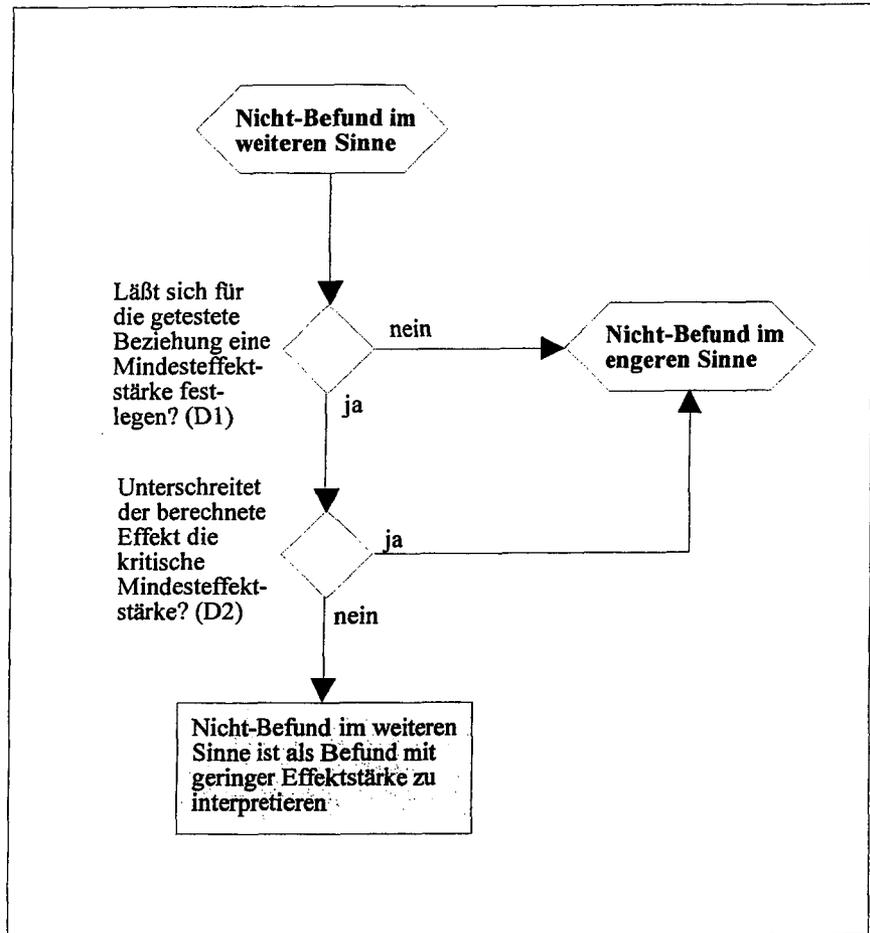


Abb. 6: Modul E: "Nicht-Befund im engeren Sinne"

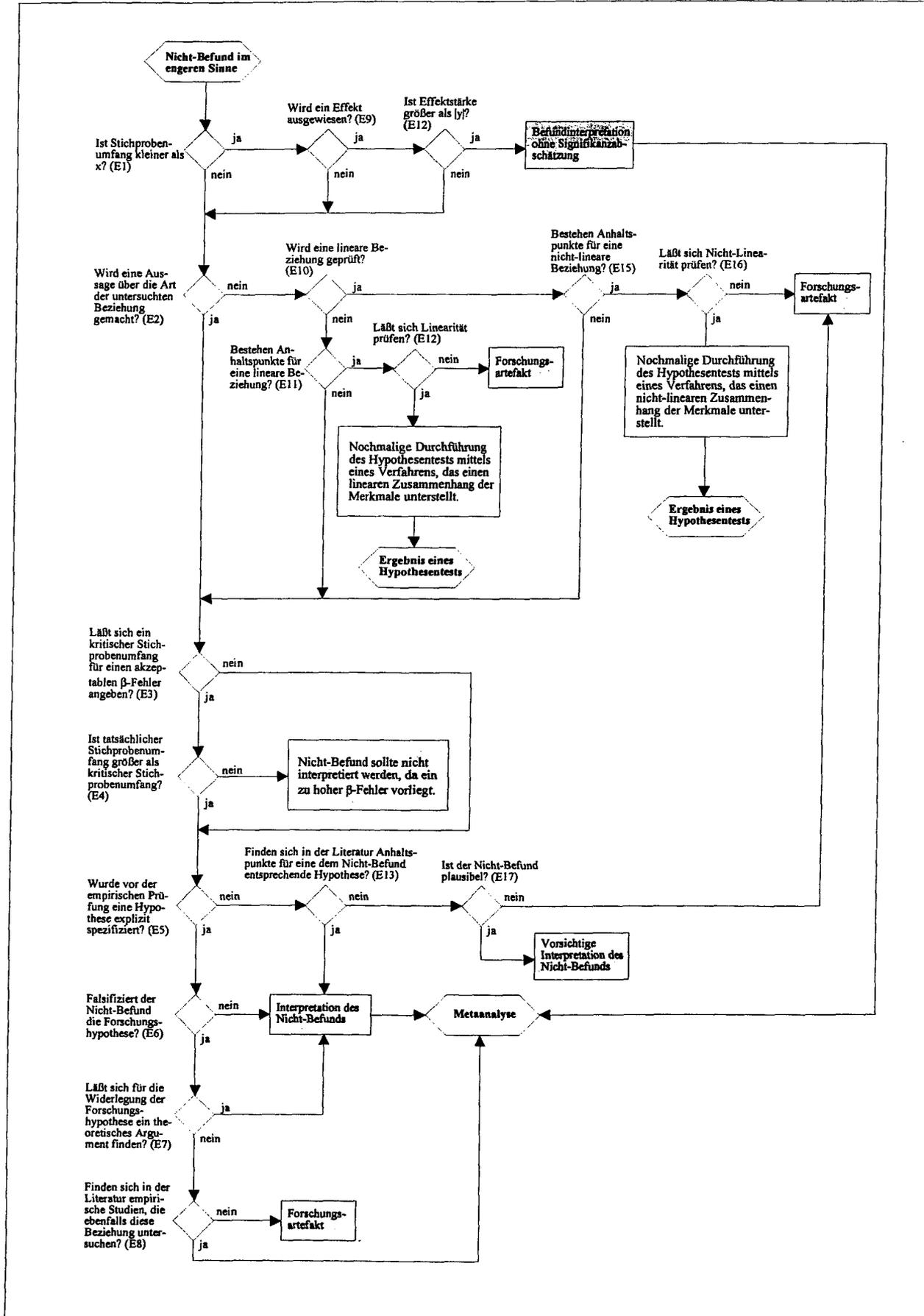


Abb. 7: Modul F: "Metaanalyse"

