SCM | Studies in Communication | Media

FULL PAPER

Führt „Ho“ auf die falsche Spur?
Die Kontroverse über Signifikanztests und ihre Relevanz für die Kommunikationswissenschaft

Is “Ho” leading down the wrong track?
The controversy about significance testing and its relevance to communication science

Andreas Vlašić

Andreas Vlašić
Medien Institut, Rheinuferstr. 9, D-67061 Ludwigshafen; Kontakt: vlasic(at)medien-institut.de

FULL PAPER

Führt „Ho“ auf die falsche Spur?
Die Kontroverse über Signifikanztests und ihre Relevanz für die Kommunikationswissenschaft

Is “Ho” leading down the wrong track?
The controversy about significance testing and its relevance to communication science

Andreas Vlašić

Zusammenfassung: Signifikanztests haben sich seit der ersten Hälfte des 20. Jahrhunderts in vielen Bereichen der Wissenschaft etabliert. Fast genauso lang werden ihre Beschränkungen und Probleme diskutiert, allerdings scheint dieser Diskurs fast unbeachtet von der Forschungspraxis zu verlaufen. Der Beitrag stellt überblicksartig die umstrittene Entwicklung des Nullhypothesentestens und die nachfolgend daran formulierte Kritik dar. Auf Basis einer Inhaltsanalyse der Forschungsberichte in den Fachzeitschriften *Medien & Kommunikationswissenschaft* sowie *Publizistik* aus den Jahren 2002 bis 2011 wird die Verwendung und Dokumentation von Signifikanztests in der Kommunikationswissenschaft betrachtet. Abschließend wird diskutiert, welche Konsequenzen sich daraus sowohl für die Forschungs- und Publikationspraxis als auch die methodische Ausbildung von Wissenschaftlern ergeben könnten.

Schlagwörter: Signifikanztest, Nullhypothesentest, Signifikanztest-Kontroverse, Qualitätsstandard, interdisziplinärer Methodendiskurs

Abstract: Since the first half of the 20th century, significance testing (NHST) has established itself in many fields of science, despite the fact that its limitations and problems have been discussed ever since then. However, the methodological debate seems to pass mostly unnoticed by research practice. This paper presents an overview of the controversial development of NHST and the criticism directed towards it. Furthermore, we investigate the use and documentation of NHST in German communication research, based on a content analysis of research reports published in the German-language journals *Medien & Kommunikationswissenschaft* and *Publizistik* during the years 2002 to 2011. Finally, we discuss implications for research and publication practice as well as possible consequences for academic training.

Keywords: statistical significance testing, null hypothesis significance testing (NHST), statistical significance controversy, standards for reporting on empirical research, interdisciplinary research methodology

1. Einführung[1]

Signifikanztests[2] gehören in vielen Bereichen der Wissenschaft zum methodischen Standard-Repertoire. Ausgehend von der Psychologie etablierten sie sich in der ersten Hälfte des 20. Jahrhunderts in zahlreichen wissenschaftlichen Disziplinen (Gigerenzer, Swijtnik, Porter, Daston, Beatty, & Krüger, 1999, S. 227-233).[3] Gleichsam parallel zu dieser Entwicklung entwickelte sich eine methodische Kontroverse über die Logik des Konzepts und die Aussagefähigkeit der Ergebnisse von Signifikanztests. Intensiv und wiederkehrend wurde der Diskurs in der Psychologie geführt (vgl. bspw. Cohen, 1994; Harlow, Mulaik, & Steiger, 1997; Nickerson, 2000), daneben aber auch in der Medizin (Sterne & Smith, 2001; Dubben & Beck-Bornholt, 2004; Moran & Solomon, 2004), der Ökonomie (Ziliak & McCloskey, 2008; Krämer, 2011), der Soziologie (Morrison & Henkel, 1970) oder der Politikwissenschaft (Gill, 1999; Behnke, 2005, 2007; Broscheid & Gschwend, 2005).[4] In der Kommunikationswissenschaft hat eine vergleichbare Diskussion über Signifikanztests und die Probleme ihrer Anwendung bislang nicht stattgefunden, dies gilt ungeachtet verschiedener Publikationen in englischsprachigen Fachzeitschriften (Katzer & Sodt, 1973; Chase & Simpson, 1979; Boster, 2002; Levine, Weber, Hullett, Park, & Lindsey, 2008; Levine, Weber, Park, & Hullett, 2008) und der Darstellung in Lehrbüchern (Schnell, Hill, & Esser, 1999; Hayes, 2005; Bortz & Schuster, 2010; Weber & Fuller, 2012).

Eine intensivere Thematisierung des Diskurses erscheint aus mehreren Gründen wünschenswert: Signifikanztests sind ein wichtiger Bestandteil des Instrumentariums der quantitativen Sozialforschung, sie dienen nicht zuletzt dazu, die Relevanz von Befunden zu untermauern. Allerdings wird die Aussagekraft solcher Tests häufig miss- bzw. überinterpretiert wird (vgl. hierzu die Ausführungen in Abschnitt 2). Darüber hinaus spielt das Konzept nicht nur für die Auswertung von Daten eine Rolle, sondern auch bei der Erstellung von Forschungsdesigns, etwa wenn Untersuchungseinheiten definiert oder die zu prüfenden Hypothesen formuliert werden. Hinzu kommt, dass der Diskurs um Signifikanztests in einer breiteren Perspektive die Frage aufwirft, welche praktische Bedeutsamkeit empirische Ergebnisse vor dem Hintergrund eines prinzipiell unsicheren Wissens haben, und in welchem Maß diese Unsicherheit durch statistische Analysen reduziert werden kann.[5]

---

1 Erste Überlegungen zu diesem Thema wurden bereits im Rahmen der Jahrestagung der Fachgruppe Methoden der DGPuK im Jahr 2005 vorgestellt, der Verfasser dankt den damaligen Gutachtern sowie den Gutachtern des vorliegenden Beitrags für ihre hilfreichen Anmerkungen.

2 Im Folgenden werden die Begriffe „Signifikanztest“ und „Nullhypothesentest“ synonym gebraucht.

3 Gigerenzer & Murray (1987) prägten für diese Entwicklung den Begriff der „probabilistischen Revolution“ bzw. der „Inferenzrevolution“ (Gigerenzer, Swijtnik, Porter, Daston, Beatty, & Krüger, 1999, S. 218).

4 Auch in stark anwendungsorientierten Disziplinen wie der Wildtierkunde (Anderson, Burnham, & Thompson, 2000) oder der Marktforschung (Sawyer & Peter, 1983) finden sich Diskussionsbeiträge.

5 Die zum Teil sehr hart geführte Kontroverse um Signifikanztests mag letztlich darin begründet sein, dass der Streit um Methoden auch die dahinter stehenden Theorien berührt.

Der vorliegende Beitrag will daher die Kontroverse über Signifikanztests aufgreifen und für die Forschungs- und Publikationspraxis in der Kommunikations- und Medienwissenschaft fruchtbar machen. Im Folgenden wird zunächst ein kurzer Überblick über die historische Entwicklung des Konzepts des Signifikanztestens sowie den aktuellen Stand der Diskussion in verschiedenen Disziplinen gegeben (Abschnitt 2). Dabei geht es weniger um eine extensive Darstellung der Positionen (hierzu gibt es mittlerweile zahlreiche Überblicksbeiträge, vgl. etwa Nickerson, 2000 oder Levine, Weber, Hullett, Park, & Lindsey, 2008), als vielmehr um die Frage, welche Informationen notwendig sind, um die Aussagekraft und Reichweite von Signifikanztests beurteilen zu können. Daran schließt sich eine Inhaltsanalyse der in den Fachzeitschriften *Medien & Kommunikationswissenschaft* sowie *Publizistik* publizierten Forschungsberichte der vergangenen 10 Jahre an (Abschnitt 3). Sie fokussiert die Frage, wie häufig inferenzstatistische Methoden in der kommunikationswissenschaftlichen Forschungspraxis angewendet werden und wie (ausführlich) die Ergebnisse dokumentiert sind. Abschließend wird diskutiert, welche Konsequenzen sich daraus einerseits für die Forschungs- und Publikationspraxis, andererseits für die methodische Ausbildung ergeben könnten.

2. Entstehung, Weiterentwicklung und Kritik an der Hybridtheorie des Signifikanztestens

Die Grundidee des Signifikanztestens findet sich erstmals im 18. Jahrhundert bei John Arbuthnot, der bereits 1710 anhand der Geburtsraten von Männern und Frauen in London zu beweisen versuchte, dass der Überhang von männlichen Neugeborenen kein Zufall sondern als Ausdruck des Wirkens Gottes zu verstehen sei (Gigerenzer 2004, S. 5-6). Der Statistiker Ronald A. Fisher entwickelte diese Idee rund 200 Jahre später zum Konzept des Nullhypothesentestens weiter.[6] Es beinhaltet im Wesentlichen 1) das Aufstellen einer Nullhypothese (d. h. der Hypothese, dass ein oder mehrere – empirisch realisierte – Stichproben aus derselben – hypothetischen und unendlichen – Population stammen, deren Stichprobenverteilung bekannt ist) und 2) das Verwerfen dieser Nullhypothese, wenn die Mittelwerte der Stichprobe bedeutsam (definiert durch das Signifikanzniveau α) vom Mittelwert der Stichprobenverteilung abweichen.[7] Dieses Konzept wurde in der Folge entscheidend von Jerzy Neyman und Egon S. Pearson erweitert. Abweichend zu Fisher werden in ihrer Konzeption zwei Hypothesen aufgestellt, Null- und Alternativhypothese, wobei auch die Alternativhypothese – ebenso wie die zugrundeliegende Verteilungsform – spezifiziert werden muss. Dadurch wird es möglich, sowohl den Fehler erster Art (fälschliches Zurückweisen der Nullhypo-

6 Wir folgen bei der Darstellung der Entstehung, Weiterentwicklung und Kritik des Konzepts den Ausführungen in Gigerenzer, Swijtnik, Porter, Daston, Beatty und Krüger (1999, S. 114-128).

7 Fisher suchte ursprünglich nach einem Weg, um anhand von ermittelten empirischen Daten beurteilen zu können, wie zutreffend Hypothesen sind. Wie die Diskussion über Signifikanztests verdeutlicht, gibt der Fishersche Ansatz darauf allerdings keine Antwort, vielmehr wird die Wahrscheinlichkeit betrachtet, bestimmte Daten zu erhalten, unter der Annahme, *dass* die Nullhypothese zutrifft.

these) als auch den Fehler zweiter Art (fälschliches Zurückweisen der Alternativhypothese) zu kontrollieren.

Zwischen Fisher auf der einen Seite und Neyman und Pearson auf der anderen Seite entbrannte in der Folge eine leidenschaftliche Diskussion über die richtige Interpretation und Anwendung von Hypothesentests.[8] Die Kontroverse wurde zu Lebzeiten der Kontrahenten nicht abschließend geklärt, ungeachtet dessen etablierte sich nach ihrem Tod in den Lehrbüchern und der Forschung eine Praxis, die beide Konzepte miteinander verbindet (vgl. die ausführliche Darstellung in Gigerenzer, Swijtnik, Porter, Daston, Beatty, & Krüger, 1999, S. 128-131). Das Grundprinzip dieser „Hybridtheorie" des statistischen Signifikanztestens lässt sich in drei Schritten zusammenfassen: 1) Formulierung zweier statistischer Hypothesen H_0 und H_1, 2) Festlegung eines Signifikanzniveaus α als der Wahrscheinlichkeit für einen Fehler erster Art, 3) Erhebung der Daten und Bestimmung der Wahrscheinlichkeit p dieser Daten unter der Annahme H_0, wobei H_0 zurückgewiesen wird, wenn p kleiner oder gleich dem festgelegten Signifikanzniveau α ist. Als „Reject-support" wird bezeichnet, wenn eine Zurückweisung von H_0 die Theorie des Forschers stützt, im Vergleich dazu seltener ist der „Accept-support", also die Konzeption, dass eine nicht mögliche Zurückweisung von H_0 die Theorie des Forschers stützt (Nickerson, 2000). Zudem sind zwei Formen von Nullhypothesen zu unterscheiden: „Nil-Nullhypothesen" gehen von einem Populationseffekt bzw. -unterschied von 0 aus, im Gegensatz ist mit „Non-nil-Nullhypothesen" die Annahme verbunden, dass ein spezifizierter (von 0 verschiedener) Populationseffekt vorliegt.

Mit dem ‚klassischen Modell' des Signifikanztestens sind verschiedene Probleme bzw. diskussionswürdige Annahmen verbunden, die mittlerweile in zahlreichen Publikationen diskutiert wurden.[9] So gibt es eine Reihe von Missverständnissen hinsichtlich der Logik von Signifikanztests: Etwa die Annahme, p drücke die Wahrscheinlichkeit aus, dass die Nullhypothese wahr sei (und 1-p die Wahrscheinlichkeit für die Alternativhypothese), oder die Interpretation, dass ein niedriger p-Wert einen Hinweis auf die Replizierbarkeit von Ergebnissen gibt. Zu solchen inhaltlichen Missverständnissen, die nicht zuletzt aus der Schwierigkeit des Denkens in (bedingten) Wahrscheinlichkeiten resultieren, treten mathematische Beschränkungen: Signifikanztests weisen eine hohe Sensitivität für die Stichpro-

8 Die Kontroverse brachte eine bemerkenswerte Polemik mit sich, so verglich Fisher seinen Kontrahenten Neyman mit „Russen, [die] mit dem Ideal vertraut gemacht werden, daß die Forschung in der reinen Wissenschaft auf die technologische Anwendung hin ausgerichtet werden könne und solle, im Rahmen der umfassend organisierten Anstrengung eines Fünfjahresplans für die ganze Nation" und nannte seine Position „entsetzlich [für] die intellektuelle Freiheit des Westens" (Gigerenzer, Swijtnik, Porter, Daston, Beatty, & Krüger, 1999, S. 127). Neyman wiederum bezeichnete einige von Fishers Tests als „in einem mathematischen Sinne ‚schlimmer als nutzlos'" (ebd.).

9 Exemplarisch hierfür sei Nickerson (2000) angeführt, der den methodischen Diskurs und die jeweiligen Kritikpunkte detailliert herausgearbeitet hat. Für weitere Kritik am Konzept des Signifikanztestens vgl. Witte (1980), Harlow et al. (1997), Gigerenzer, Swijtnik, Porter, Daston, Beatty und Krüger (1999), Kline (2004, S. 61-91), Gigerenzer (2004a, 2004b), Cohen (1990, 1994), Ziliak und McCloskey (2008); zur Verteidigung des Konzepts vgl. bspw. Chow (1996), Abelson (1997) oder Wainer und Robinson (2003). Für die Kommunikationswissenschaft haben Levine, Weber, Hullett, Park und Lindsey (2008) und Levine, Weber, Park und Hullett (2008) die Kritik am Nullhypothesentesten sowie mögliche Verbesserungen bzw. Alternativen zusammengefasst.

bengröße auf, zudem unterliegt die Festlegung zentraler Testkriterien (Definition des Signifikanzniveaus, Formulierung der Hypothesen) den subjektiven Entscheidungen des Forschers. Die häufig praktizierte Formulierung und Zurückweisung von Nil-Nullhypothesen etwa erlaubt in vielen Fällen keinen substanziellen Erkenntnisgewinn, größere Aussagekraft erhält man durch die Definition eines Bereichs, der als relevanter Unterschied zu 0 verstanden werden soll (also die Prüfung von Non-nil-Nullhypothesen). Ein dritter Komplex der Kritik schließlich bezieht sich auf die langfristigen Folgen der Anwendung von Signifikanztests für den wissenschaftlichen Fortschritt: So kann es problematisch sein, wenn Forscher sich an das mechanische Formulieren inhaltsarmer Nullhypothesen gewöhnen und über eine ritualisierte Verwendung von statistischen Werkzeugen die Analyse komplexer Probleme vernachlässigen (vgl. dazu auch Gigerenzer, 2004a). Die durch Signifikanztests errechneten p-Werte geben keinen Hinweis auf die praktische Bedeutsamkeit von Befunden, dennoch scheinen Studien mit statistisch signifikanten Befunden häufiger eingereicht bzw. publiziert zu werden, wodurch letztlich der Fehler erster Art in der Forschungsliteratur steigt.

Obwohl das Konzept des Signifikanztestens bereits früh (Berkson, 1942) und bis heute wiederkehrend kritisch diskutiert wurde, scheint dieser Diskurs lange Zeit ‚abgekoppelt‘ von der Forschungs- und Publikationspraxis und weitgehend ohne Folgen geblieben zu sein (McLean & Ernest, 1998; Sedlmeier, 2009). In den 90er-Jahren gewann die Diskussion wieder an Intensität, nicht zuletzt ausgehend von einer pointierten Kritik Jacob Cohens (1994):

> What’s wrong with significance testing? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe in that it does! What we want to know is ‘given these data, what is the probability that H0 is true?’ But as most of us know, what it tells us is ‘given that H0 is true, what is the probability of these (or more extreme) data?’. (S. 997)

Es entwickelte sich eine Methodendiskussion (vgl. bspw. Chow, 1996; Sedlmeier, 1996, 1998; Abelson, 1997; Hunter, 1997; Gigerenzer, 1998; Daniel, 1998a, 1998b), im Verlauf derer verschiedene Autoren den Nutzen von Signifikanztests grundsätzlich in Frage stellten oder gar deren Verbot forderten. Als Reaktion setzte die American Psychological Association (APA) im Jahr 1999 eine Arbeitsgruppe ein, die einen Konsens im Hinblick auf die Praxis der Beurteilung und Veröffentlichung von Signifikanztest in Forschungsberichten definieren sollte. Die Ergebnisse des Abschlussberichts (Wilkinson, 1999) schlugen sich in den nachfolgenden Ausgaben des APA Publication Manual nieder,[10] so wird in der aktuellen sechsten Auflage einleitend festgestellt, dass statistische Signifikanztests – trotz ihrer langjährigen Dominanz in der psychologischen Forschung – lediglich

10 Die intensive Thematisierung im Manual erfolgte allerdings erst mit einer deutlichen zeitlichen Verzögerung (Fidler, 2010). Eine ähnliche ‚Trägheit‘ zeigt sich an verschiedenen Stellen der methodischen Diskussion über die Begrenzungen von Signifikanztests: Trotz der über Jahrzehnte wiederholt vorgetragenen Empfehlung, Effektgrößen und Teststärken zu berichten, scheint sich die Forschungs- und Publikationspraxis in der Psychologie nicht oder nur wenig zu verändern (Sedlmeier & Gigerenzer, 1989; Cohen, 1990, S. 1310; McLean & Ernest, 1998, S. 18; Sedlmeier, 2009).

ein Ausgangspunkt für die Datenanalyse seien (APA, 1999, S. 33). Hypothesentests werden nicht grundsätzlich als untauglich abgelehnt, allerdings formulieren die Verfasser Empfehlungen, wie die Ergebnisse dokumentiert sein sollten, um eine Evaluation ihrer Aussagekraft und Reichweite durch Dritte zu ermöglichen. Dabei erscheinen vier Aspekte zentral:

- Ein wesentlicher Bestimmungsfaktor jedes Signifikanztests ist die *Größe der Stichprobe*: Signifikanztests reagieren sensibel auf die jeweils zugrundeliegende Fallzahl, je nach Größe der Stichprobe und damit der Varianz interessierender Merkmale kann beispielsweise der gleiche Mittelwertunterschied statistisch signifikant ausfallen oder nicht (Daniel, 1998, S. 25-27). Zudem kann es von der Fallzahl einer Erhebung bzw. den realisierten Zellbesetzungen abhängen, welches Maß für die Effektgröße berichtet werden sollte (Levine, Weber, Park, & Hullett, 2008, S. 190).
- In engem Zusammenhang mit der Größe der Stichprobe steht die Abschätzung der *Stärke* bzw. der *Power* des durchgeführten Tests (Rosnow & Rosenthal, 1989, S. 1277-1278). Das APA Publication Manual empfiehlt, die bei der Planung einer Analyse die relevanten Annahmen hinsichtlich der erwarteten Größe von Effekten, der geplanten bzw. realisierten Größe der Stichprobe und daraus resultierend der Power des Tests exakt zu dokumentieren (APA, 1999, S. 30). Dadurch lässt sich die Wahrscheinlichkeit angeben, inwiefern durch ein gegebenes Forschungsdesign die vermuteten Effekte tatsächlich entdeckt werden können.[11]
- Grundsätzlich legt das APA Publication Manual den Forschern eine extensive Dokumentation der Ergebnisse von statistischen Signifikanztests nahe. Hierzu gehört etwa die Angabe der *exakten p-Werte*, abweichend zur häufig geübten Praxis der Kennzeichnung von signifikanten Unterschieden auf aggregiertem Niveau.[12] Die Angabe exakter p-Werte ermöglicht es, zu einem späteren Zeitpunkt Meta-Analysen durchzuführen und die Ergebnisse einer Studie/eines Tests vor dem Kontext breiterer Forschung einzuordnen. Als weitere sinnvolle Angaben im Zusammenhang mit Punktschätzungen (Mittelwerte, Regressionskoeffizienten) werden Informationen über ihre Genauigkeit bzw. Variabilität angeführt. Darüber hinaus empfehlen die Verfasser des Manuals nachdrücklich die Dokumentation von Konfidenzintervallen, da diese eine „extrem wirkungsvolle Art der Berichterstattung“ (APA, 1999, S. 34, Überset-

11 Aus forschungspraktischer Sicht ist darüber hinaus anzunehmen, dass die Auseinandersetzung mit der Power eines Tests auch zu einer höheren Sensibilisierung des Forschers gegenüber den Ergebnissen führt. Cohen (1990, S. 1304) zeigte in einer Sekundäranalyse der Teststärke von psychologischen Studien, dass in vielen Fällen die durchgeführten Signifikanztests lediglich so aussagekräftig wie ein Münzwurf waren (vgl. dazu auch Levine, Weber, Hullett, Park, & Lindsey, 2008, S. 177-178).

12 Fisher verwarf die von ihm eingeführte Konvention der Kennzeichnung von signifikanten Ergebnissen auf 0,01- bzw. 0,05-Niveau einige Zeit später wieder und empfahl bereits 1956 die Angabe von exakten p-Werten (Gigerenzer, 1998, S. 199). Durch den exakten Ausweis lässt sich auch die Dichotomie der zu einem guten Teil willkürlichen Unterscheidung zwischen ‚signifikant vs. nicht signifikant‘ abmildern, Rosnow und Rosenthal fassen dies pointiert in der Formulierung zusammen: „[S]urely, God loves the .06 nearly as much as the .05.“ (Rosnow & Rosenthal, 1989, S. 1277)

zung d. Verf.) darstellten (vgl. dazu auch Meehl, 1997, S. 421 und Levine, Weber, Park, & Hullett, 2008, S. 189-190 sowie 192-194).

- Statistische Signifikanz drückt nicht gleichzeitig auch die praktische Bedeutsamkeit eines Ergebnisses aus, je nach Größe der Stichproben und Varianz der analysierten Parameter können auch kleine Unterschiede bzw. Effekte signifikant sein. Daher empfiehlt das APA Publication Manual, in jedem Fall zusammen mit den Ergebnissen eines Hypothesentests auch ein Maß für die *Effektgröße* zu berichten (APA, 1999, S. 30-31; für einen Überblick über mögliche Maße vgl. Grissom & Kim, 2005 und Levine, Weber, Park, & Hullett, 2008, S. 190-192).

3. Verwendung und Dokumentation von Signifikanztests in der kommunikationswissenschaftlichen Forschung

3.1 Forschungsfrage und Forschungsdesign

In der Kommunikationswissenschaft hat sich die Kontroverse um Signifikanztests bislang kaum niedergeschlagen (Katzer & Sodt, 1973; Chase & Simpson, 1979; Boster, 2002). In jüngerer Vergangenheit veröffentlichte ein Autorenteam um Timothy R. Levine und René Weber in der amerikanischen Fachzeitschrift Human Communication Research zwei Beiträge, in denen die Probleme und insbesondere mögliche Optimierungen bzw. Alternativen zum Nullhypothesentesten dargestellt werden (Levine, Weber, Hullett, Park, & Lindsey, 2008; Levine, Weber, Park, & Hullett, 2008). Daneben finden sich in den deutschen Fachzeitschriften vereinzelt Beiträge, die zwar thematisch im Kontext stehen (Stichprobentheorie, Auswahlverfahren), jedoch nicht dezidiert Signifikanztests fokussieren.[13] In einschlägigen Lehrbüchern wird die Problematik der Anwendung von Signifikanztests thematisiert (vgl. etwa Schnell, Hill, & Esser, 1999, S. 416-418; Hayes, 2005, S. 158-182; Bortz & Schuster, 2010, S. 112), allerdings scheinen die Komplexität und Begrenzungen des Konzepts dabei häufig eine nachrangige Rolle zu spielen.[14] Insgesamt gesehen entsteht der Eindruck, dass Signifikanztests zwar einen hohen Stellenwert in der empirischen Forschung haben, der Diskurs über ihre Beschränkungen aber bislang noch nicht in der nötigen Breite im Fach rezipiert wurde. Durch eine empirische Analyse sollten daher die folgenden explorativen Fragen beantwortet werden:

- Wie häufig werden statistische Signifikanztests im Kontext empirisch ausgerichteter kommunikationswissenschaftlicher Forschungsarbeiten eingesetzt?

13 Lauf (2001) diskutiert Anforderungen an die Dokumentation von Reliabilitätstests in Inhaltsanalysen, Gehrau und Fretwurst (2005) sowie Fretwurst, Gehrau und Weber (2005) beschäftigen sich mit der Dokumentation von Auswahlverfahren in kommunikationswissenschaftlichen Untersuchungen.

14 Dies lässt sich etwa anhand des Lehrbuchs von Bortz und Schuster (2012, S. 106-113) beobachten: Darin finden sich auf rund acht Seiten die Beschränkungen von Signifikanztests und die Bedeutung von Aspekten wie der Effektgröße oder der Teststärke zusammengefasst; dem gegenüber steht insgesamt ein Umfang von 523 Seiten mit der Darstellung von Analysenmethoden, in denen zu einem großen Teil Signifikanztests eingesetzt werden.

- Wie ausführlich werden die durchgeführten Signifikanztests dokumentiert (hier insbesondere mit einem Fokus auf den Aspekten Fallzahl, p-Werte, Effektgrößen und Teststärke)?
- Werden Beschränkungen des gewählten Analysemodells (und mögliche Alternativen) diskutiert?

Zur Beantwortung dieser Fragen wurde eine Inhaltsanalyse der Forschungsberichte in den Fachzeitschriften *Medien- und Kommunikationswissenschaft* sowie *Publizistik* aus den vergangenen zehn Jahren durchgeführt.[15] Die Konzentration auf diese beiden Zeitschriften erschien sinnvoll, da Fachzeitschriften einen Überblick über den Status quo einer Disziplin hinsichtlich der Forschungsagenda sowie der dominierenden methodischen Zugänge erlauben (Brosius & Haas, 2009, S. 170; Möhring & Scherer, 2011, S. 58-59). Für die Kommunikations- und Medienwissenschaft haben die beiden ausgewählten Fachpublikationen eine herausragende Bedeutung, was nicht zuletzt daran deutlich wird, dass die Mitgliedschaft in der Fachgesellschaft DGPuK ein Abonnement beider Zeitschriften einschließt.[16] Aus allen Forschungsberichten im Zeitraum von 2002 bis 2011 wurden zunächst diejenigen Beiträge identifiziert, in denen (zumindest auch) eine statistische Auswertung quantitativ erhobener empirischer Daten enthalten ist. Dabei war es unerheblich, ob die analysierten Daten aus einer eigenen Primärerhebung stammten oder eine Re-Analyse von bestehenden Datensätzen durchgeführt wurde. Sofern bei der Analyse und Darstellung der Daten inferenzstatistische Methoden Anwendung fanden, wurde weiter erfasst, welche Verfahren zur Prüfung der Forschungshypothesen eingesetzt wurden, und wie umfassend die Ergebnisse des Verfahrens im Bericht dokumentiert sind. Dabei standen vier Aspekte im Vordergrund:

1. Werden die die Fallzahlen, die den einzelnen Analysen zugrunde liegen, dokumentiert?
2. Werden die exakten Signifikanzwerte dokumentiert oder lediglich auf aggregiertem Niveau ($\leq 0{,}05$, $\leq 0{,}01$ und $\leq 0{,}001$) ausgewiesen?
3. Wird ein Maß für die Effektgröße bzw. -stärke berechnet?
4. Wird die Teststärke (β-Fehler) angegeben oder thematisiert (entweder in der Untersuchungsplanung zur Bestimmung der notwendigen Stichprobengröße oder bei der Auswertung zur Bestimmung der realisierten Teststärke nach der Erhebung)?

Angesichts der Komplexität der Frage nach der ‚richtigen' Verwendung von Signifikanztests erfolgte die Codierung vergleichsweise wenig restriktiv. So gibt es etwa verschiedene Maße für Effektstärke, entscheidend war nicht, ob ein spezifi-

15 Der Verfasser dankt Sabine Bock, Milena Martin und insbesondere Sandra Kokott und Ina Hohenegger für die Codierung der Beiträge.

16 Weitere Argumente – insbesondere in Abgrenzung der beiden genannten Fachzeitschriften zu anderen Publikationen – finden sich bei Brosius und Haas (2009, S.172).

scher Parameter angegeben, sondern *dass* überhaupt ein Maß berichtet wurde.[17] Die Codierung der Beiträge erfolgte in enger Absprache mit dem Verfasser, aus diesem Grund kann von einer hohen instrumentellen Reliabilität (Kolb, 2004, S. 337) ausgegangen werden. Nach Abschluss der Erfassung wurde ein Intercoderreliabilitätstest durchgeführt, der eine hohe Übereinstimmung der Codierungen zeigt (für alle Variablen zur Dokumentation von Signifikanztests beträgt der Koeffizient nach Holsti 0,951).[18] Grund hierfür dürfte sein, dass bei der Konzeption des Kategoriensystems bewusst Merkmale im Vordergrund standen, die einen vergleichsweise manifesten/expliziten Charakter haben (bspw. Einsatz inferenzstatistischer Analysen, Angabe von Fallzahlen, exakte Ausweisung von p-Werten).

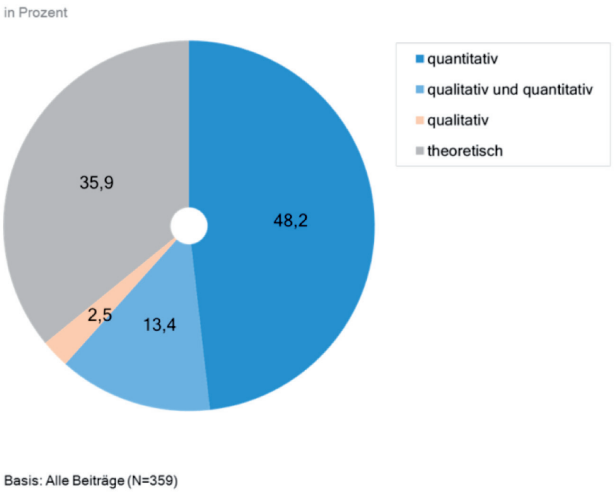
3.2 Ergebnisse

Insgesamt wurden im Untersuchungszeitraum in beiden Fachzeitschriften N=359 Forschungsberichte identifiziert. Die Mehrzahl dieser Beiträge ist empirisch ausgerichtet und beinhaltet qualitativ und/oder quantitativ erhobene Daten, lediglich rund 36 Prozent der Artikel beschäftigen sich aus einer rein theoretischen Perspektive mit ihrer Forschungsfrage (vgl. Abb. 1). Rund in der Hälfte aller Fälle basieren die Analysen auf einer ausschließlich quantitativen Datenbasis, Methodenkombinationen mit qualitativ und quantitativ erhobenen Daten machen rund 13 Prozent der Stichprobe aus, ausschließlich qualitativ orientierte Studien haben einen Anteil von rund drei Prozent. Die vergleichsweise deutliche Akzentuierung von quantitativen Forschungsmethoden in beiden Fachzeitschriften entspricht den Befunden von Möhring und Scherer (2011, S. 64-65), wobei die Frage offen bleiben muss, welchen Anteil daran die Präferenzen der Forscher auf der einen Seite und die Richtlinien der Herausgeber sowie Urteile der externen Gutachter auf der anderen Seite haben.

17 Gängige Maße hierfür wären etwa Cohens d, Cramer´s V, Konfidenzintervalle oder η^2, der Ausweis der erklärten Varianz einer Regressionsanalyse durch das Bestimmtheitsmaß R^2 oder die Stärke einer Korrelation durch den Korrelationskoeffizienten r wurden ebenfalls als mögliche Maße für die Effektstärke gewertet.

18 Der Reliabilitätstest wurde anhand einer Stichprobe von N=45 zufällig ausgewählten Beiträgen der Stichprobe, in denen sich Signifikanztests fanden, durchgeführt. Die Erfassung erfolgte durch eine Mitarbeiterin, die bis zu diesem Zeitpunkt nicht in die Codierung eingebunden war. Da die verschiedenen Aspekte zur Dokumentation von Signifikanztests in unterschiedlicher Kombination vorliegen konnten, wurden sie als nominale bzw. dichotome Variablen vercodet. Hinsichtlich der zentralen Aspekte der Analyse zeigt sich eine insgesamt vergleichbar hohe Reliabilität der Codierungen (Angabe von Fallzahlen=0,978, exakte p-Werte=0,936, Effektgröße/-stärke=0,919, Teststärke=0,972).

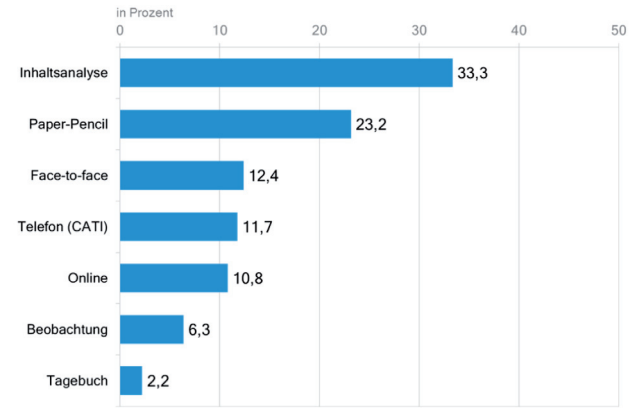
Abbildung 1: Theoretische vs. empirische Ausrichtung der Forschungsberichte



Betrachtet man die empirisch ausgerichteten Beiträge (N=221) hinsichtlich der darin eingesetzten Erhebungsmethoden, so zeigt sich, dass am häufigsten Inhaltsanalysen (rund 33 %, vgl. Abbildung 2) und schriftliche Befragungen (rund 23 %) durchgeführt werden.[19] Vergleichsweise selten sind computergestützte Telefon-Interviews (CATI), im Unterschied etwa zum Bereich der angewandten (Markt-)Forschung, in dem dieser Methode eine hohe Bedeutung zukommt. Beobachtungen und Tagebuchstudien haben den geringsten Anteil am Spektrum der eingesetzten Forschungsmethoden.

19 Mehrfachnennungen waren möglich, bei multimethodalen Forschungsdesigns wurden alle eingesetzten Erhebungsmethoden und durchgeführten statistischen Analysen erfasst. Experimentelle Anordnungen wurden nicht als eigene Kategorie definiert, die darin eingesetzten Erhebungsmethoden wurden jeweils einzeln codiert.

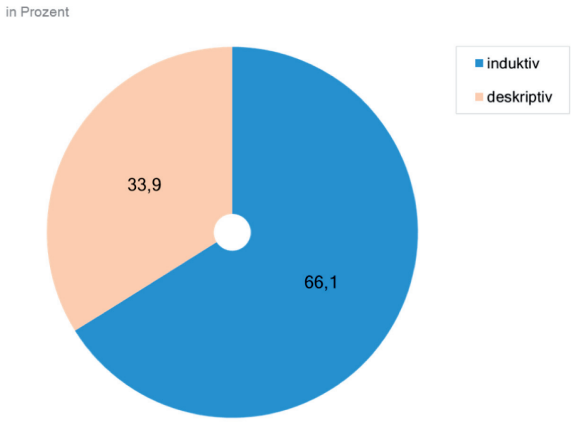
Abbildung 2: Eingesetzte Erhebungsmethoden in empirisch ausgerichteten Forschungsberichten



Basis: Alle eingesetzten Erhebungsmethoden (N=315)

Wie verbreitet ist nun der Einsatz von Signifikanztests in Beiträgen, in denen quantitativ erhobene Daten berichtet werden? In deutlich mehr als der Hälfte dieser Fälle (rund 66 %) kommen Methoden der schließenden Statistik zum Einsatz (vgl. Abb. 3). Dies bestätigt die Vermutung, dass Signifikanztests auch in der empirischen Kommunikations- und Medienwissenschaft eine wichtige Rolle spielen. Folgt man der Annahme, dass die Publikationslage in den beiden führenden Fachzeitschriften – zumindest näherungsweise – Aufschluss über die in der Fachgemeinschaft anerkannten inhaltlichen und methodischen Zugänge zum Forschungsfeld gibt, so lässt sich festhalten, dass Signifikanztests zum etablierten ‚Mainstream‘ gezählt werden können.

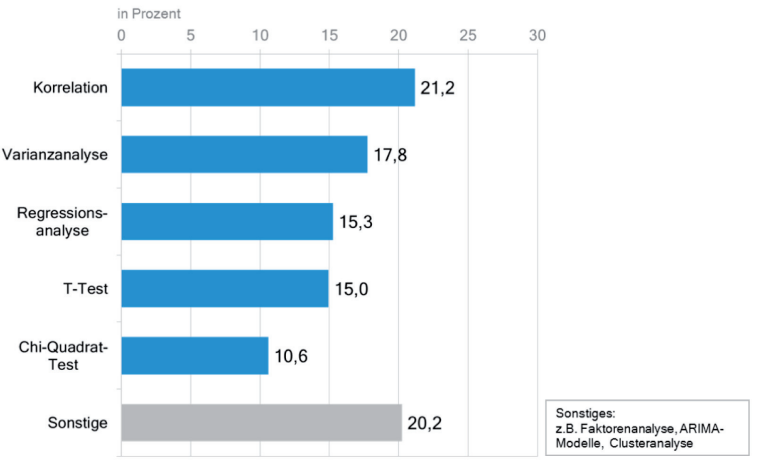
Abbildung 3: Häufigkeit des Einsatzes von Inferenzstatistik in empirisch ausgerichteten Forschungsberichten



Basis: Alle empirisch ausgerichteten Beiträge, in denen (auch) quantitative Daten erhoben wurden (N=221)

Um die Dokumentation der durchgeführten Signifikanztests näher zu beleuchten, werden im Folgenden die fünf am häufigsten eingesetzten Tests betrachtet. Es handelt sich dabei um die Korrelations-, Varianz- und Regressionsanalyse sowie den T-Test und den χ^2-Test (vgl. Abb. 4).

Abbildung 4: Häufigkeit verschiedener inferenzstatistischer Analysen



Basis: Alle durchgeführten inferenzstatistischen Analysen (N=321)

Die Ergebnisse der durchgeführten Signifikanztests werden fast durchweg mit expliziten Angaben über die jeweils zugrunde liegende Anzahl von Fällen versehen (vgl. Tab. 1). Während hier also eine Art ‚Deckeneffekt' zu beobachten ist, stellt

sich das Bild bezüglich der Dokumentation der resultierenden p-Werte nahezu umgekehrt dar: Bei der Dokumentation des ‚typischen Kriteriums‘ für statistische Signifikanz folgen die Forscher in aller Regel der Konvention, p-Werte aggregiert für die etablierten Signifikanz-Niveaus auszuweisen (bspw. über eine Kennzeichnung der Koeffizienten durch Asterisken oder Indizes). Eine exakte Angabe von p-Werten findet sich am vergleichsweise häufigsten für die analysierten T-Tests sowie die Varianzanalysen, insgesamt gesehen ist eine genaue Dokumentation dieses Parameters jedoch eher selten.

Hinsichtlich der Angabe von Maßen für die Effektgröße zeigt sich je nach Analyseform ein heterogenes Bild: Bei den ausgewerteten Regressions- und Korrelationsanalysen findet sich fast durchweg ein entsprechender Parameter (in aller Regel ist dies das Bestimmtheitsmaß R^2 bzw. der Korrelationskoeffizient r[20]). Für die durchgeführten Varianzanalysen hingegen wird deutlich seltener eine Effektgröße berichtet: Lediglich rund die Hälfte der analysierten Tests weist einen entsprechenden Parameter (üblicherweise η^2) aus. Bei χ^2-Tests liegt der Anteil mit rund 18 Prozent nochmals niedriger, in der Darstellung der Ergebnisse von T-Tests schließlich spielt die Effektgröße so gut wie keine Rolle. Ein ‚Bodeneffekt‘ zeigt sich abschließend hinsichtlich der Teststärke: In keinem der analysierten Beiträge findet sich eine explizite Dokumentation oder gar Diskussion dieses Aspekts (etwa im Rahmen der Dokumentation der Versuchsplanung).

Tabelle 1: Dokumentation der häufigsten inferenzstatistischen Analysen

	χ2-Test (N=34)	*T-Test (N=48)*	*Korrelations-analyse (N=68)*	*Varianz-analyse (N=57)*	*Regressions-analyse (N=49)*	*Gesamt (N=256)*
genaue Fallzahl	97,1	95,8	98,5	100,0	98,0	97,9
exakte p-Werte	5,9	25,0	4,5	15,8	6,1	11,5
Effektgröße	17,6	4,2	95,5	49,1	100,0	53,3
Teststärke	-	-	-	-	-	-

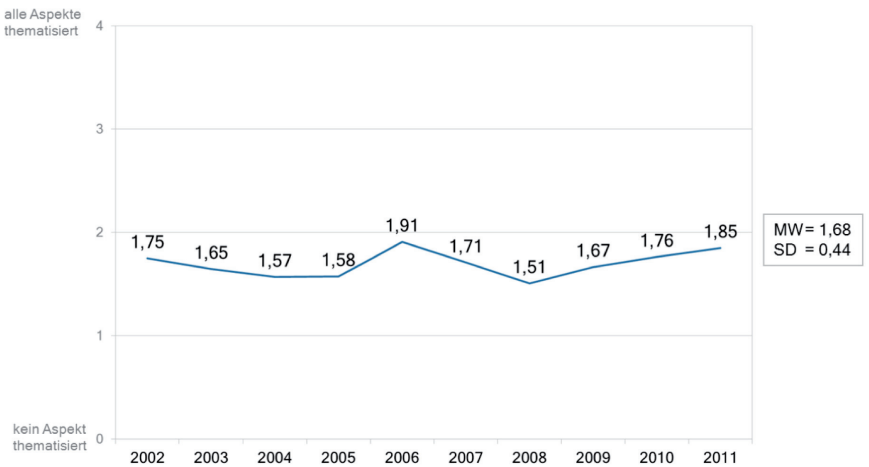
Basis: Alle durchgeführten inferenzstatistischen Analysen (in Prozent)

Um mögliche Entwicklungen über die Zeit hinweg analysieren zu können, wurde ein Index gebildet, der für jeden Beitrag erfasst, ob in den darin beschriebenen Signifikanztests die vier Aspekte – explizite Angabe von Fallzahlen, exakte p-Werte, Maße für die Effektgröße sowie Teststärke – angegeben/thematisiert werden oder nicht. Der Index kann somit Werte zwischen 0 (=kein Aspekt thematisiert) und 4 (=alle Aspekte thematisiert) annehmen. Betrachtet man nun die Entwicklung dieses Index‘ über die untersuchten Jahre hinweg, so zeigt sich eine große Kontinuität: Im Mittel werden im Zeitverlauf bei der Dokumentation von Signifikanztests annähernd zwei Parameter ausgewiesen (vgl. Abb. 5), die Schwankun-

20 Da der Ausweis des Korrelationseffizienten r ein genuiner Bestandteil der Korrelationsanalyse ist, scheint das Ergebnis nicht unbedingt geeignet, um daraus auf eine hohe Sensibilisierung für die Probleme des Signifikanztestens zu schließen.

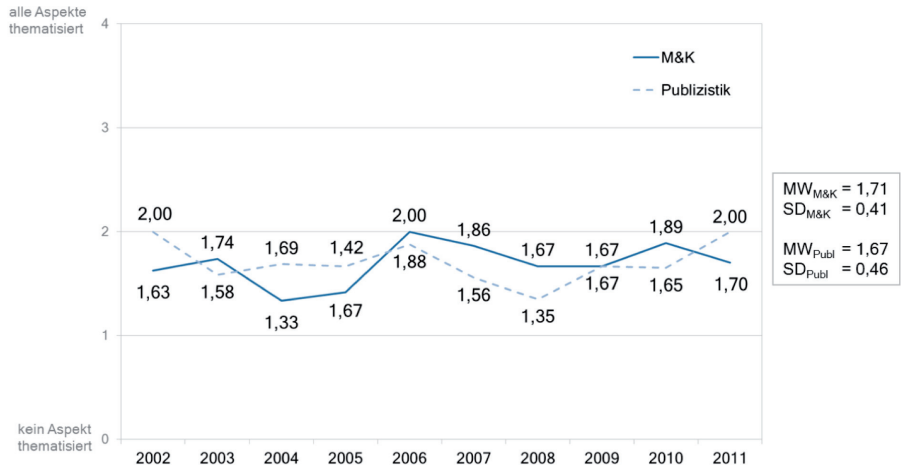
gen bleiben in einem geringen Bereich. Eine detaillierte Betrachtung zeigt, dass am häufigsten die Fallzahl sowie ein Maß für die Effektstärke ausgewiesen werden, für letztere scheinen nicht zuletzt die Konventionen der Dokumentation von Korrelations- und Regressionsanalysen ausschlaggebend zu sein. Zwischen den beiden Fachzeitschriften finden sich weder summativ betrachtet noch in den einzelnen Jahren wesentliche Unterschiede im Hinblick auf die Verwendung und Dokumentation von Signifikanztests (vgl. Abb. 6). Dies lässt darauf schließen, dass die Publikationspraxis im Hinblick auf Nullhypothesentests tatsächlich den Status Quo im Fach abbildet (und nicht etwa auf Schwerpunktsetzungen durch Herausgebergremien o. Ä. zurückzuführen ist). Das über die Jahre und beide Fachzeitschriften hinweg stabile Bild könnte zudem als indirekter Hinweis darauf interpretiert werden, dass die Debatte um Signifikanztests und die daraus resultierenden Empfehlungen bislang nicht in der nötigen Breite im Fach rezipiert wurden. Daran scheinen auch einschlägige Publikationen im Fach (Hayes, 2005; Levine, Weber, Hullett, Park, & Lindsey, 2008; Levine, Weber, Park, & Hullett, 2008) bislang nichts geändert zu haben.

Abbildung 5: Dokumentation von Signifikanztests im Zeitverlauf



Basis: Alle Beiträge, in denen mindestens eine der fokussierten Analysen (Chi-Quadrat-Test, T-Test, Korrelationsanalyse, Varianzanalyse, Regressionsanalyse) durchgeführt wurde (N=137)

Abbildung 6: Dokumentation von Signifikanztests im Zeitverlauf – nach Fachzeitschrift



Basis: Alle Beiträge, in denen mindestens eine der fokussierten Analysen (Chi-Quadrat-Test, T-Test, Korrelationsanalyse, Varianzanalyse, Regressionsanalyse) durchgeführt wurde (N=137)

4. Zusammenfassung und Diskussion

Wie die Inhaltsanalyse der Forschungsberichte in den Fachzeitschriften *Medien & Kommunikationswissenschaft* und *Publizistik* zeigt, haben sich Signifikanztests auch in der Kommunikations- und Medienwissenschaft auf breiter Basis etabliert: In rund zwei Drittel aller Beiträge, in denen quantitative Daten ausgewertet werden, kommen sie zum Einsatz. Allein diese häufige Verwendung spricht dafür, die Diskussion über Einsatzmöglichkeiten und Aussagekraft von Nullhypothesentests auch im kommunikationswissenschaftlichen Fachdiskurs stärker als bisher aufzugreifen. Zudem zeigt sich durchaus Optimierungspotenzial hinsichtlich der Dokumentation von Signifikanztests. So werden in aller Regel die errechneten p-Werte nicht exakt, sondern auf aggregierten Signifikanzniveaus ausgewiesen, was eine nachfolgende Durchführung von Meta-Analysen erschwert. Ebenfalls ausbaufähig erscheint die Angabe von Parametern zur Einschätzung der Effektgrößen, die vor allem bei χ^2-Quadrat- und T-Tests nur selten dokumentiert werden. Die Notwendigkeit einer stärkeren Sensibilisierung für die konzeptionellen Beschränkungen von Signifikanztests tritt deutlich beim Aspekt der Teststärke von statistischen Analysen zutage: Sie hat hohe Relevanz sowohl für die Beurteilung der ‚praktischen Bedeutung‘ von signifikanten Ergebnissen – insbesondere bei Studien mit kleinen Fallzahlen oder geringen Unterschieden – als auch für die Bewertung von nicht-signifikanten Befunden. In den inhaltsanalytisch ausgewerteten Forschungsbeiträgen aus den Jahren 2002 bis 2011 wird sie jedoch durchweg nicht thematisiert.

Insgesamt entsteht der Eindruck, dass die Forschungs- bzw. Publikationspraxis der Medien- und Kommunikationswissenschaft der Komplexität von Signifikanztests nur eingeschränkt Rechnung trägt. Wie ließe sich dies künftig möglicherwei-

se ändern? Erstens erscheint eine stärkere Standardisierung der Dokumentation von Forschungsergebnissen hinsichtlich der diskutierten Aspekte sinnvoll. Allein die Vorgabe, für jede statistische Analyse auch ein Maß der Effektgröße anzugeben bzw. die Teststärke der Analyse zu berechnen, könnte zur Sensibilisierung der Forscher beitragen. Eine Beschäftigung mit der Teststärke von Signifikanztests bringt geradezu zwangsläufig einen höheren Aufwand für die Planung von empirischen Erhebungen mit sich: Der Forscher ist gehalten, Vermutungen hinsichtlich der (theoretisch) zu erwartenden Größe von Effekten anzustellen, und muss weiterhin – über die per Konvention festgelegte Wahrscheinlichkeit eines Fehlers erster Art hinaus – auch bestimmen, in welchem Ausmaß das Risiko des Nicht-Entdeckens möglicher Zusammenhänge (Fehler zweiter Art) für ihn akzeptabel ist. Letztlich berührt die Frage nach der Teststärke die Herausforderung, statistische Hypothesen möglichst präzise und ‚riskant' zu formulieren, um inhaltlich aussagekräftige Forschung zu ermöglichen. Levine, Weber, Park und Hullett (2008) zeigen bspw. am Beispiel der Korrelation, wie Non-nil-Nullhypothesen auf Basis bisheriger Forschungsergebnisse spezifiziert und geprüft werden können (S. 196-199; vgl. hierzu und für weitere Analysen auch Bortz & Döring, 2006, S. 635-669).

Mindestens genauso wichtig erscheint es zweitens, dass ein entsprechender Common Sense unter den Herausgebern und Gutachtern in der Fachgemeinschaft entsteht. Mithilfe der computergestützten Datenanalyse lassen sich viele der genannten Parameter problemlos berechnen, deutlich schwieriger hingegen ist es, die etablierten Konventionen der Darstellung zu verändern. Entscheidende Bedeutung kommt dabei der Publikationspraxis und damit den Richtlinien zu, die für die Einreichung von Beiträgen bei Fachzeitschriften oder wissenschaftlichen Tagungen gelten. Sie beeinflussen die ‚Sichtbarkeit' und damit Akzeptanz von theoretischen und methodischen Zugängen und tragen dadurch wesentlich zur Durchsetzung entsprechender Standards bei.[21] Dabei erscheint die im Rahmen der Signifikanztest-Kontroverse wiederholt vorgetragene Forderung, Nullhypothesentests sollten aufgrund ihrer Defizite nicht mehr durchgeführt werden, zu weitreichend. Letztlich muss jeder Forscher im Einzelfall entscheiden, ob ein Signifikanztest das geeignete statistische Werkzeug für die Beantwortung seiner Forschungsfrage ist oder nicht, und wie er ggf. die Hypothesen formuliert, um das Erkenntnispotenzial auszuschöpfen. Umso mehr sollten diese Entscheidungen jedoch theoretisch begründet und die resultierenden Ergebnisse ausführlich dokumentiert werden. Vogelgesang und Scharkow (2012) werfen hinsichtlich der Dokumentationspraxis von Inhaltsanalysen die Frage auf, ob die Herausgeber der Fachzeitschriften angesichts der uneinheitlichen Handhabung von Reliabilitätstests nicht konkrete Mindestanforderungen definieren sollten, die bei der Einreichung von Manuskripten zu erfüllen sind. Dieser Vorschlag erscheint auch für die Dokumentation von Signifikanztests bedenkenswert, wobei die Empfehlungen des APA Publication Manuals eine Orientierung geben könnten.

21 Orlitzky (2012, S. 200) vertritt die Ansicht, dass die notwendige „De-Institutionalisierung" des Signifikanztestens nicht allein durch die Individuen zu leisten ist, vielmehr müssten (neue) institutionelle Regeln und Anreize die Verwendung alternativer Methoden fördern.

So wünschenswert eine stärkere Sensibilisierung für die Problematik und die Veränderung (bzw. Erweiterung) der Konventionen für die Dokumentation von Signifikanztests auch sein mag, so sehr sollte man sich gleichzeitig darüber bewusst sein, dass jede ‚Standardisierung‘ an Grenzen stößt. In der Kontroverse um das Nullhypothesentesten wird deutlich, dass statistische Analysen auf einer intensiven theoretischen Fundierung basieren (und diese folglich nicht ersetzen können).[22] Gerade weil verschiedene Testkonzepte je nach Fragestellung, Forschungsdesign, formulierten Hypothesen, empirischen Verteilungen der Merkmale etc. besser oder schlechter geeignet sein können, bleiben wichtige Entscheidungen in das Ermessen des einzelnen Forschers gestellt. Weder Signifikanztests noch die dazu diskutierten Alternativen bieten eine einfache, ‚universale Technik‘ des statistischen Schließens. Cohen fasst die Situation daher pointiert zusammen: „(D)on’t look for the magic alternative to NHST [Null Hypothesis Significance Testing, d. Verf.], some other objective mechanical ritual to replace it. It doesn´t exist.“ (Cohen, 2004, S. 1001)

Damit richtet sich drittens der Blick auf die akademische Ausbildung, die mit darüber entscheidet, welche theoretischen und methodischen Zugänge sich in einem Fach etablieren und halten können.[23] Für die Lehre stellt die Komplexität des Signifikanztestens (und nicht zuletzt auch die Ungereimtheiten der Hybridtheorie) eine veritable Herausforderung dar. In einer Befragung unter Psychologie-Studenten fanden Haller und Krauss (2002), dass die Mehrheit der Befragten das Konzept des Signifikanztestens nicht bzw. falsch verstanden hatten (sie replizierten damit die Befunde einer Studie unter amerikanischen Studenten, vgl. Oakes, 1986 zitiert nach Haller & Krauss, 2002, S. 3). Inkorrekte Interpretationen und Missverständnisse zeigen sich darüber hinaus auch bei Postgraduierten (Nelson, Rosenthal, & Rosnow, 1986; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993; Windisch, Huot, & Green, 2007). Zu dieser Unsicherheit mag allerdings auch beitragen, dass das Nullhypothesentesten in einigen Büchern nicht korrekt oder zumindest unvollständig beschrieben wird (Gliner, Leech, & Morgan, 2002, S. 90-91; Krämer & Gigerenzer, 2005, S. 225).[24] Ein Aufgreifen des Diskurses über Signifikanztests im Fach könnte dazu beitragen, dass in Zukunft Lehrbücher und Curricula entsprechend angepasst werden. Grundsätzlich wäre auch eine intensivere Beschäftigung mit der Thematik im Rahmen der Ausbildung von Me-

22 Verschiedene Autoren weisen darauf hin, dass die Diskussion über Signifikanztests letztlich die Frage berührt, wie präzise Theorien formuliert werden (Witte, 1989; Cohen, 1990; Behnke, 2007).

23 Gigerenzer, Swijtnik, Porter, Daston, Beatty und Krüger (1999, S.128-131) zufolge war die Synthese der verschiedenen Ansätze zu einer Hybridtheorie (die Autoren sprechen auch von einem „Theorieeintopf“) vor allem durch wissenschaftshistorische bzw. wissenssoziologische Faktoren bedingt (Abkoppelung der statistisch-mathematischen Diskussion von der empirischen Forschungspraxis, redaktionelle Entscheidungen der Herausgeber von Fachzeitschriften, Kanonisierung in Lehrbüchern).

24 Wie unterschiedlich die Behandlung des Themas erfolgen kann, zeigt ein Vergleich von zwei aktuellen Lehrbüchern: Während Bortz und Döring (2006, S. 600-669) die Diskussion um Signifikanztests und deren Konsequenzen auf rund 70 Seiten sehr ausführlich darstellen und dabei u. a. als Modifikation die Prüfung von Non-Nil-Nullhypothesen („Minimum-Effekt-Nullhypothesen“, Bortz & Döring, 2006, S. 635) vorschlagen, findet sich im Lehrbuch von Bortz und Schuster (2010, S.106-113), das 2010 als Neuauflage von Bortz (2005) erschienenen ist, lediglich eine vergleichsweise knappe Abhandlung.

dien- und Kommunikationswissenschaftlern wünschenswert, dies erscheint angesichts der knappen zeitlichen Ressourcen in vielen Bachelor- und Masterstudiengängen derzeit allerdings kaum zu realisieren. Zudem vermutet der Autor aus eigener Erfahrung, dass sich ein tiefergehendes Verständnis für (die Beschränkungen von) Signifikanztests erst auf Basis eigener Forschungsarbeiten und damit über einen längeren Zeitraum entwickelt. Daher scheint es sinnvoll, in dieser Hinsicht keine ‚Maximalziele' zu formulieren (also „noch mehr Statistik im Stundenplan"), sondern als Minimalziel anzustreben, dass Studierende im Rahmen der vorgesehenen Statistikausbildung stärker für die Grenzen von Signifikanztests und mögliche Alternativen hierzu sensibilisiert werden. Dabei könnte die Darstellung der Kontroverse bzw. die Entstehungsgeschichte der Hybridtheorie als ein Ansatzpunkt für ein besseres Verständnis hilfreich sein: Die umstrittene Entwicklung des Konzepts durch die Statistiker Fisher, Neyman und Pearson sowie die nachfolgenden Kontroversen machen deutlich, wie voraussetzungsreich die theoretischen Annahmen sind, auf denen Signifikanztests beruhen. Ein zentraler Baustein der methodischen Ausbildung von empirisch arbeitenden Wissenschaftlern sollte daher das Bewusstsein dafür sein, dass „mechanisches Rezeptwissen" (Gigerenzer, 1998, S. 200) nicht das statistische Denken ersetzen kann. Wenn dies dazu führt, dass angehende Wissenschaftler künftig weniger häufig Hypothesentests einsetzen und anstatt dessen etwa ihre Forschungsergebnisse intensiver deskriptiv beschreiben, dann wäre dies alles andere als ein Verlust.

Viertens folgt aus der Kontroverse um Signifikanztests in anderen wissenschaftlichen Disziplinen eine forschungsstrategische Überlegung: Mit Blick auf die Idee eines kumulativen Erkenntnisfortschritts in der Wissenschaft wäre es interessant zu erforschen, welche Folgen eine (auch für die Medien- und Kommunikationswissenschaft denkbare) Bevorzugung oder Überbetonung von statistisch signifikanten Ergebnissen hat (Witte, 1980, S. 57; Nickerson, 2000, S. 270-272). Für die Überprüfung eines derartigen „Publication Bias" sind verschiedene Ansätze denkbar, beispielsweise die Re-Analyse von Studien im Hinblick auf ihre Effektstärke (Cohen, 1962, zit. nach Cohen, 1994, S. 1002; Sedlmeier & Gigerenzer, 1989) oder intensivere Versuche zur Replikation von Befunden. Die Herausgeber von Fachzeitschriften können diesen Prozess unterstützen, indem sie die Einreichung entsprechender Beiträge anregen und sich – inhaltliche Relevanz vorausgesetzt – für die Publikation von Studien mit nicht-signifikanten Ergebnissen einsetzen.[25]

Der Diskurs über Signifikanztests hatte in der Vergangenheit oftmals einen dogmatischen und stellenweise hitzigen Charakter,[26] in einem bemerkenswerten Gegensatz dazu steht seine bislang begrenzte Rezeption. Handelt es sich also lediglich um ein randständiges Thema, einen ‚Sturm im Wasserglas'? In einem größeren Rahmen betrachtet gibt es in der Tat viele weitere Faktoren, die bei der

25 Seit 2002 publiziert das Online-Journal „Journal of Articles in Support of the Null Hypothesis" (http://www.jasnh.com) dezidiert psychologische Studien, in denen keine signifikanten Ergebnisse gefunden wurden.

26 Dies ließ sich schon bei der Kontroverse zwischen Fisher und Neyman/Pearson beobachten, ein eindrucksvolles Beispiel aus jüngerer Vergangenheit ist die Diskussion unter den Wirtschaftswissenschaftlern Hoover und Ziegler (2008a, 2008b), McCloskey und Ziliak (2008a, 2008b) sowie Spanos (2008).

quantitativen empirischen Forschung eine wichtige Rolle spielen, beispielsweise die Operationalisierung der Forschungsfrage, die Wahl der Erhebungsmethode, das Design des Erhebungsinstruments, die Auswahl der Untersuchungsobjekte u. a. m. Die statistische Datenanalyse – und damit das Konzept des Signifikanztestens – ist so gesehen ‚nur‘ ein Teilbereich des Forschungsprozesses. Allerdings ist er mit den anderen Bereichen eng verzahnt und kann für sie weitreichende Folgen haben, etwa wenn die Logik des Nullhypothesentestens beeinflusst, in welcher Weise Forschungsdesigns konzipiert oder Forschungsfragen formuliert werden,[27] oder wenn die Fehlinterpretation der Bedeutung von signifikanten Ergebnissen zu einer Inflation des Fehlers erster Art in der Forschungsliteratur führt. In der Kontroverse über Signifikanztests wurden diese und weitere Argumente intensiv diskutiert, so dass es für empirische Kommunikationsforscher gute Gründe gibt, sich intensiv mit dieser Thematik zu befassen. Solange sich die Ergebnisse der Kontroverse und die daraus resultierenden Empfehlungen allerdings nicht in der Forschungs- und Publikationspraxis niederschlagen, wird der Diskurs ein ‚Sturm im Wasserglas‘ bleiben. Dies jedoch nicht, weil die darin diskutierten Probleme wissenschaftstheoretisch irrelevant wären, sondern vielmehr, weil die (auch in der Wissenschaft zu beobachtende und soziologisch plausible) ‚Macht des Faktischen‘ den Status Quo erhält.

An der Diskussion über Signifikanztests lässt sich somit auch etwas über den wissenschaftlichen Diskurs lernen. Poppers Erkenntnistheorie zufolge bleibt jedes Wissen prinzipiell unsicher, dies gilt auch für wissenschaftliche Theorien und Methoden. Spätestens seit Kuhns Beschreibung von Paradigmenwechseln in der Wissenschaft wurde zudem deutlich, dass dabei auch soziale Faktoren eine Rolle spielen können. Umso wichtiger ist es daher, dass die Wissenschaft ihren Blick auf ‚die Realität‘ und deren empirische Erfassung immer wieder kritisch reflektiert. Die Diskussion über Signifikanztests – die zugegebenermaßen zuweilen eher spezialisiert und kleinteilig anmutet – ist auf diese Weise in letzter Konsequenz mit den ‚großen Fragen‘ nach der angemessenen Beschreibung und Erfassung von gesellschaftlicher Wirklichkeit verknüpft. Eine intensivere Rezeption der methodischen Diskussion über Signifikanztests und damit einhergehend ein besseres Verständnis darüber, welche Aussagen sie ermöglichen – und besonders auch: welche nicht – könnte dazu beitragen, dass der wissenschaftliche Diskurs eine ausgewogene Balance hält zwischen der ‚Macht der Zahlen‘ und dem ‚Zwang des besseren Arguments‘.

27 Kaplans „law of the instrument“ (Kaplan, 1964, S. 28) beschreibt die Beobachtung, dass die zur Verfügung stehenden Methoden bzw. ‚Werkzeuge‘ die Wahl der zu lösenden Probleme durch Wissenschaftler beeinflussen können.

Literatur

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would have to be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Hrsg.), *What if there were no significance tests?* (S. 117-141). Mahwah, NJ: Erlbaum.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *Journal of Wildlife Management, 64*(4), 912-923.

Behnke, J. (2005). Lassen sich Signifikanztests auf Vollerhebungen anwenden? Einige essayistische Anmerkungen. *Politische Vierteljahresschrift, 46*, S. O1-O15.

Behnke, J. (2007). Kausalprozesse und Identität. Über den Sinn von Signifikanztests und Konfidenzintervallen bei Vollerhebungen. *Beiträge zu empirischen Methoden der Politikwissenschaft, 2*(3). Abgerufen von http://www.wiso.uni-hamburg.de/fileadmin/sowi/ak_methoden/Behnke_-_Kausalprozesse_und_Identitaet.pdf

Berkson, J. (1942). Tests of Significance Considered as Evidence. *Journal of the American Statistical Association, 37*(219), 325-335. Abgerufen von http://www.botany.wisc.edu/courses/botany_940/06EvidEvol/papers/1Berkson.pdf

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

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

Boster, F. J. (2002). On making progress in communication science. *Human Communication Research, 28*, 473-490.

Brandstätter, E. (1999). Konfidenzintervalle als Alternative zu Signifikanztests. *Methods of Psychological Research Online, 4*(2). Abgerufen von http://www.dgps.de/fachgruppen/methoden/mpr-online/issue7/art3/brandstaetter.pdf

Broscheid, A., & Gschwend, T. (2005). Zur statistischen Analyse von Vollerhebungen. *Politische Vierteljahresschrift 46*, S. O-16–O-26.

Brosius, H.-B., & Haas, A. (2009). Auf dem Weg zur Normalwissenschaft. Themen und Herkunft der Beiträge in Publizistik und Medien & Kommunikationswissenschaft. *Publizistik, 54*, 169-190.

Chase, L. J. & Simpson, T. J. (1979). Significance and substance: An examination of experimental effects. *Human Communication Research 5*, 351-354.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.

Chow, S. L. (2002). Issues in Statistical Inference. *History and Philosophy of Psychology Bulletin, 14*(1), 30-41. Abgerufen von http://cogprints.org/2892/1/CritAPA.pdf

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Daniel, L. G. (1998a). Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals. *Research in the Schools, 5*(2), S. 23-32.

Daniel, L. G. (1998b). The Statistical Significance Controversy Is Definitely Not Over: A Rejoinder to Responses by Thompson, Knapp, and Levin. *Research in the Schools, 5*(2), 63-65.

Dubben, H.-H., & Beck-Bornholdt, H.-P. (2004). *Unausgewogene Berichterstattung in der medizinischen Wissenschaft - publication bias-*. Hamburg: Institut für Allgemeinmedi-

zin des Universitätsklinikums Hamburg-Eppendorf. Abgerufen von http://www.uke.de/institute/allgemeinmedizin/downloads/institut-allgemeinmedizin/BROSCHUERE_-_Publication_bias.pdf

Fretwurst, B., Gehrau, V., & Weber, R. (2005). Notwendige Angaben zu Auswahlverfahren. Theoretische Überlegungen und eine empirische Auswertung der Dokumentationspraxis in der KW. In V. Gehrau, B. Fretwurst, B. Krause, & G. Daschmann (Hrsg.), *Auswahlverfahren in der Kommunikationswissenschaft* (S. 32-51). Köln: von Halem.

Fidler, F. (2010). *The American Psychological Association Publication Manual Sixth Edition: Implication for Statistics Education* (Arbeitspapier). Abgerufen von http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_C156_FIDLER.pdf

Gehrau, V., & Fretwurst, B. (2005). Auswahlverfahren in der Kommunikationswissenschaft. Eine Inhaltsanalyse aktueller Veröffentlichung über empirische Studien in der Kommunikationswissenschaft. In V. Gehrau, B. Fretwurst, B. Krause, & G. Daschmann (Hrsg.), *Auswahlverfahren in der Kommunikationswissenschaft* (S. 13-31). Köln: von Halem.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*(2), 199-200.

Gigerenzer, G. (2004a). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Gigerenzer, G. (2004b). Die Evolution des statistischen Denkens. *Unterrichtswissenschaft – Zeitschrift für Lernforschung, 32*(1), 4-22.

Gigerenzer, G., Swijtnik, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1998). *Das Reich des Zufalls. Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Ursachen*. Heidelberg, Berlin: Spektrum Akademischer Verlag.

Gill, J. (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly, 52*(3), 647-674.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems With Null Hypothesis Significance Testing (NHST). What Do the Textbooks Say? *The Journal of Experimental Education, 71*(1), 83-92.

Haller, H., & Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online 2002, 7*(1), 1-20. Abgerufen von http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Hrsg.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hayes, A. F. (2005). *Statistical Methods for Communication Science*. Mahwah, NJ: Lawrence Erlbaum.

Hoover, K. D., & Siegler, M. (2008a). Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology, 15*(1), 1-37.

Hoover, K. D., & Siegler, M. (2008b). “The rhetoric of ‘Signifying nothing’: a rejoinder to Ziliak and McCloskey”. *Journal of Economic Methodology 15*, 57-68.

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

Kaplan, A. (1964). *The conduct of inquiry. Methodology for behavioral science*. San Francisco, CA: Chandler.

Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. *Journal of Communication, 23*, 251-265.

Kline, R. B. (2004). *Beyond significance testing: reforming data analysis methods in behavioral research*. Washington D.C.: American Psychological Association.

Krämer, W. (2011). *Das Signifikanztest-Ritual und andere Sackgassen des Fortschritts in der Statistik* (SFB 823 Discussion Paper 32). Abgerufen von https://eldorado.tu-dortmund.de/bitstream/2003/29073/1/DP_3211_SFB823_Kr%c3%a4mer.pdf

Krämer, W., & Gigerenzer, G. (2005). How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities. *Statistical Science, 20*(3), 223-230. Abgerufen von http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdfview_1&handle=euclid.ss/1124891288

Kolb, S. (2004). Verlässlichkeit von Inhaltsanalysedaten. Reliabilitätstest, Errechnen und Interpretieren von Reliabilitätskoeffizienten für mehr als zwei Codierer. *Medien & Kommunikationswissenschaft, 52*(3), 335–354.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research, 34*, 171-187.

Levine, T. R., Weber, R., Park, H. S. & Hullett, C. (2008). A Communication Researchers' Guide to Null Hypothesis Significance Testing and Alternatives. *Human Communication Research*, *34*, 188-209.

McCloskey, D. N., & Ziliak, S. T. (2008a). Science is judgment, not only calculation: a reply to Aris Spanos's review of The cult of statistical significance. *Erasmus Journal for Philosophy and Economics, 1*(1), 165-170.

McCloskey, D. N., & Ziliak, S. T. (2008b). Signifying nothing: reply to Hoover and Siegler. *Journal of Economic Methodology, 15*(1), 39-55.

McLean, J. E., & Ernest, J. M. (1998). The Role of Statistical Significance Testing In Educational Research. *Research in the Schools, 5*(2),15-22.

Meehl, P. E. (1997) The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Hrsg.), *What if there were no significance tests?* (S. 117-141). Mahwah, NJ: Erlbaum.

Möhring, Wiebke, & Scherer, Helmut (2011). Eine Frage des Themas? Einsatzfelder qualitativer und quantitativer Verfahren in den letzten Jahrzehnten. In A. Fahr (Hrsg.), *Zählen oder Verstehen? Diskussion um die Verwendung quantitativer und qualitativer Methoden in der empirischen Kommunikationswissenschaft* (S. 57-71). Köln: Herbert von Halem Verlag.

Moran, J. L., & Solomon, P. J. (2004). A farewell to P-values? *Critical Care & Resuscitation, 6*, 139-138.

Morrison, D. E., & Henkel, R. E. (Hrsg.). (1970) *The Significance Test Controversy*. Aldine: Chicago.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, *41*, 1299-1301.

Nickerson, R. S. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods, 5*(2), 241-301.

Orlitzky, M. (2012). How Can Significance Tests Be Deinstitutionalized? *Organizational Research Methods*, *15*(2), 199-228.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist, 44*(10), 1276-1284. Abgerufen von http://wiki.ubc.ca/images/3/3c/Rosnow_%26_Rosenthal._1998._Statistical_Procedures_%28aspirin_example%29.pdf

Sawyer, A. G., & Peter, J. P. (1983). The Significance of Statistical Sicnificance Tests in Marketing Research. *Journal of Marketing Research, 10,* 122-133.

Schnell, R., Hill, P. B., & Esser, E. (1999). *Methoden der empirischen Sozialforschung*. München, Wien: Oldenburg.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online 1996, 1*(4), 41-63.

Sedlmeier, P. (1998). Was sind die guten Gründe für Signifikanztests? *Methods of Psychological Research Online 1998, 3*(1), 39-42.

Sedlmeier, P. (2009). Beyond the Significance Test Ritual: What Is There? *Journal of Psychology, 217*(1), 1-5.

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power have an Effect on the Power of Studies? *Psychological Bulletin*, *105*(2), 309-316.

Vogelgesang, J., & Scharkow, M. (2012). Reliabilitätstests in Inhaltsanalysen. Eine Analyse der Dokumentationspraxis in Publizistik und Medien & Kommunikationswissenschaft. *Publizistik, 57*, 333–335.

Schweiger, W., Rademacher, P., & Grabmüller, B. (2009). Womit befassen sich kommunikationswissenschaftliche Abschlussarbeiten? *Publizistik, 54*, 533–552.

Spanos, A. (2008). Review of S. T. Ziliak and D. N. McCloskey's The Cult of Statistical Significance. *Erasmus Journal for Philosophy and Economics, 1*(1), 154-164. Abgerufen von http://ejpe.org/pdf/1-1-br-2.pdf

Sterne, J. A. C., & Smith, G. D. (2001). Sifting the evidence - what's wrong with significance tests? *Physical Therapy, 81*(8), 1464-1469.

Thompson, B. (1998). Statistical Significance and Effect Size Reporting: Portrait of a Possible Future. *Research in the Schools, 5*(2), 15-22.

Wainer, H., & Robinson, D. H. (2003). Shaping Up the Practice of Null Hypothesis Significance Testing. *Educational Researcher*, 22-30. Abgerufen von http://legacy.aera.net/uploadedFiles/Journals_and_Publications/Journals/Educational_Researcher/3207/3207_ResNewsComment.pdf

Weber, R., & Fuller, R. (2012). *Statistical Methods for Communication Researchers and Professionals*. Dubuque, IA: Kendall Hunt.

Windish, D. M., Huot, S. J., & Green, M. L. (2007). Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature. *Journal of the American Medical Association*, *298*(9), 010-1022. Abgerufen von http://jama.jamanetwork.com/article.aspx?articleid=208638

Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations (Task Force on Statistical Inference). *American Psychologist, 54*(8), 594-604.

Witte, E. H. (1980). *Signifikanztest und statistische Interferenz: Analysen, Probleme, Alternativen*. Stuttgart: Enke.

Witte, E. H. (1989). Die „letzte“ Signifikanztestkontroverse und daraus abzuleitende Konsequenzen. *Psychologische Rundschau*, *40*, 76-84.

Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance - How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary Issues in the Analysis of Data: A Survey of 551 Psychologists. *Psychological Science*, *4*(1), 49-53. Abgerufen von http://homepage.psy.utexas.edu/homepage/class/Psy391P/Josephs%20PDF%20files/Zuckerman%20et%20al..PDF

Is "Ho" leading down the wrong track? The controversy about significance testing and its relevance to communication science

Andreas Vlašić

The ongoing controversy about statistical significance testing (NHST)

Since the first half of the 20th century, significance testing (NHST) has established itself in many fields of science. Starting from psychology, the "inference revolution" (Gigerenzer, Swijtnik, Porter, Daston, Beatty, & Krüger, 1999) took place place e.g. in medicine, economics, sociology, or political science. Even at an early stage of this process methodological discussion arose concerning the implications and limitations of the so-called "hybrid theory" (Berkson, 1942; Harlow, Mulaik, & Staiger, 1997). As a result, many authors stretch the fact that NHST is just one of several possible strategies for analysis of quantitative data; if conducted, results should be documented in detail in order to allow for an assessment and further analysis of the collected data.

However, the methodological debate and its conclusions appear to pass mostly unnoticed by research practice (Sedlmeier, 2009). This seems to be true for communication science, where there has been only little discussion on this topic (see as an exception Levine, Weber, Hullett, Park, & Lindsey, 2008; Levine, Weber, Park, & Hullett, 2008). Therefore, this paper addresses the following exploratory questions: 1) How frequently do communication researchers use NHST in the context of empirically oriented studies? 2) When using NHST, do researchers report and interpret central parameters (e.g. effect size, power)? 3) Is there a discussion of limitations of NHST (and possible alternatives)?

Method and Procedure

To answer the research questions we conducted a quantitative content analysis of research reports published in the German-language journals *Medien & Kommunikationswissenschaft* and *Publizistik* during the years 2002 to 2011. According to the recommendations of the APA Publication manual, the analysis focused four key aspects: 1) Documentation of sample size, 2) specification of exact p-values, 3) estimation of effect sizes and 4) statistical power.

Results

The results show that significance tests play an important role in German-language empirical communication research: In two thirds of the studies reviewed, researchers use methods of inferential statistics. While almost all of the studies indicate sample sizes, much less frequently exact p-values are given. When it

comes to dimensions for effect size, the results depend on the chosen analysis: looking at correlations and regression analyses, most often coefficients for effect size are specified; however, the reporting of ANOVA, χ^2 test or Student´s T frequently lacks adequate estimates. Finally, none of the studies in our sample discussed the power of the conducted tests. The findings remain widely constant over the years, there is no systematic difference between the two journals.

Discussion

Considering the frequent use of NHST in communication research, the outcome of this study suggests that the scientific community could benefit from an increased awareness for the (appropriate) usage of significance tests in research and publication practice. For this purpose various approaches could be useful: 1) Establishing a higher degree of standardization for the reporting on research results; 2) developing a common sense among editors and reviewers (and therefore in the scientific community); 3) increasing attention for the issue within the scope of academic training; 4) conducting further studies to analyze feasible publication biases in communication research.

References

Berkson, J. (1942). Tests of Significance Considered as Evidence. *Journal of the American Statistical Association, 37*(219), 325 – 335.

Gigerenzer, G., Swijtnik, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1998). *Das Reich des Zufalls. Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Ursachen*. Heidelberg, Berlin: Spektrum Akademischer Verlag.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Hrsg.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Sedlmeier, P. (2009). Beyond the Significance Test Ritual: What Is There? *Journal of Psychology, 2009, 217*(1), 1-5.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research, 34*, 171-187.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. (2008). A Communication Researchers' Guide to Null Hypothesis Significance Testing and Alternatives. *Human Communication Research*, *34*, 188-209.