Evidenz, Signifikanz und das kleine p

Anmerkungen zur statistischen Praxis (nicht nur) in der empirischen Unterrichtsforschung

Oliver Passon · Tassilo von der Twer

Eingegangen: 16. September 2020 / Überarbeitet: 2. Dezember 2020 / Angenommen: 3. Dezember 2020 /
Online publiziert: 22. Dezember 2020
© Der/die Autor(en) 2020

Zusammenfassung In den letzten Jahren hat die Forderung nach „Evidenzbasierung“ in den Bildungswissenschaften zu einer vermehrten Anzahl quantitativer empirischer Untersuchungen geführt – etwa Interventionsstudien im Kontrollgruppendesign. Seit vielen Jahrzehnten gibt es jedoch eine Kontroverse darüber, wie so gewonnene Daten statistisch zu deuten seien. Insbesondere unterliegt die Interpretation des sog. p -Wertes beim Hypothesentest zahlreichen Missverständnissen. Das führte jüngst zu einer Empfehlung der Amerikanischen Gesellschaft für Statistik, das Etikett „statistisch signifikant“ nicht mehr zu verwenden. Wir diskutieren einige dieser Schwierigkeiten bzw. Missverständnisse und geben Hinweise auf mögliche Lösungsansätze.

Schlüsselwörter Evidenzbasierung · Replikationskrise · Null-Hypothesen Signifikanztest · p -Wert · frequentistische Statistik · Bayessche Statistik

O. Passon (✉) · T. von der Twer
Bergische Universität Wuppertal, Wuppertal, Deutschland
E-Mail: passon@uni-wuppertal.de

T. von der Twer
E-Mail: twer@uni-wuppertal.de

Evidence, significance and the small p

Remarks on the statistical practice (not only) in the empirical classroom research

Abstract Recent years have seen the claim that also educational science should be “evidence based” and this lead to an increase in empirical studies, e.g. in the control group design. However, for many years there have been controversies on the interpretation of these data. Especially the meaning of the p -value of null-hypothesis significance testing is subject to many misconceptions. In response to these problems the American Statistical Association has recently recommended to abandon the term “statistical significant” in scientific publications entirely. We discuss some of these problems and misconceptions and provide some advice for possible solutions.

Keywords Evidence based · Replication crisis · Null-hypothesis significance testing · p -value · Frequentist statistics · Bayesian statistics

1 Einleitung

Die letzten Jahre haben die Forderung nach einer „evidenzbasierten Praxis“ in der Pädagogik (Tooley und Darby 1998; Coe 1999) und eine „empirische Wende“ der deutschen Bildungspolitik und Bildungsforschung (Buchhaas-Birkholz 2009) erlebt. Robert E. Slavin (2002) geht so weit, die Evidenzbasierung mit einer „wissenschaftlichen Revolution“ zu vergleichen, die andere Disziplinen bereits vor mehr als 100 Jahren vollzogen hätten.

Im Zuge dessen kommt es in der Bildungsforschung zu einer immer größeren Zahl von quantitativen Forschungsarbeiten. Diese folgen häufig dem Muster von Null-Hypothesen Signifikanztests; etwa bei Interventionsstudien im Kontroll- bzw. Vergleichsgruppendesign.¹

Gleichzeitig war und ist diese (quantitativ-)empirische Ausrichtung von einem kritischen Diskurs begleitet; siehe etwa Baumert und Tillmann (2016) und Ahrbeck et al. (2016).

Diese Debatte zu den konzeptionellen Grundlagen halten wir für äußerst wichtig – sie soll aber an dieser Stelle nicht weitergeführt werden. Stattdessen wollen wir die Diskussion um die Evidenzbasierung mit dem (etwas) weniger ideologisierten Problemkomplex der statistischen Methodik verknüpfen.

¹ Bellmann und Müller (2011, S. 21) weisen darauf hin, dass in dieser Debatte der Evidenzbegriff mehrdeutig sei. Während in der englischsprachigen Diskussion darunter im Wesentlichen nur Ergebnisse aus randomisierten kontrollierten Studien (randomized controlled trials, kurz: *RCT*) oder Metaanalysen verstanden würden, fasse die deutschsprachige Diskussion auch Untersuchungen mit weniger strengem Design als Beiträge zur Evidenzbasierung auf. Auf kuriose Weise würden in der deutschsprachigen Diskussion also die Versprechen der *evidence based education* mit (nach empirischen Maßstäben) schwächeren Methoden (etwa Gelegenheitsstichproben, fehlender Randomisierung) verknüpft. Für unsere Argumentation sind diese methodischen Fragen jedoch von untergeordneter Bedeutung. Wo das Forschungsdesign hinter dem „Goldstandard“ der RCTs zurückbleibt (etwa bei quasi-experimentellen Interventionsstudien) gelten unsere Einwände sogar mit noch größerem Recht.

Seit vielen Jahrzehnten gibt es eine anhaltende Diskussion über das methodische Vorgehen und die Ergebnisinterpretation von (Null-)Hypothesen Signifikanztests (siehe etwa: Nickerson (2000); Ziliak und McCloskey (2008); Kline (2013); Lakens (2019)). Das Ziel dieser Arbeit ist es, die Grundlagen dieser Debatte zu beleuchten und sie stärker in die Bildungsforschung zu tragen. Angesichts des erheblichen Aufschwungs der empirischen Bildungsforschung in den letzten Jahren und der damit verbundenen Hoffnung auf steuerungsrelevantes Wissen halten wir diese Debatte für besonders wichtig (vgl. Nagengast und Rose (2018, S. 686)).

Zusätzliche Aktualität hat dieser Diskurs durch die sog. „Replikationskrise“ gewonnen. Dieser Begriff – ursprünglich für die Psychologie geprägt – ist in der Zwischenzeit zum Synonym dafür geworden, dass sich in immer mehr hypothesentestenden Wissenschaften zahlreiche Forschungsergebnisse nicht replizieren lassen (Pashler und Wagenmakers 2012; Makel und Plucker 2014; Open Science Coll. 2015).

Noch früher erregte ein Beitrag von John P. A. Ioannidis Aufsehen. Dieser profilierte Vertreter der evidenzbasierten Medizin veröffentlichte bereits 2005 eine Arbeit mit dem provokanten Titel „Why Most Published Research Findings Are False“ (Ioannidis 2005). Dort führt Ioannidis im Kern den (eigentlich recht banalen) Nachweis, dass die Wahrscheinlichkeit für korrekte (positive) Forschungsergebnisse in den hypothesentestenden Wissenschaften nicht nur von der Irrtumswahrscheinlichkeiten abhängt, sondern natürlich auch von der Rate der „korrekten“ Forschungshypothesen (d. h. der sog. Prävalenz bzw. Vortestwahrscheinlichkeit).

Nimmt man z. B. an, dass 10% der (Forschungs-)Hypothesen zutreffen, die statistische Teststärke („*power*“) den für viele psychologische Studien typischen Wert von $1 - \beta \approx 50\%$ hat und auf einem Signifikanzniveau von 5% getestet wird, werden ca. 50% der „statistisch signifikanten“ Ergebnisse falsch sein (sog. *false discovery rate*).²

Vor diesem Hintergrund ging die *American Statistical Association* schließlich sogar so weit, den Verzicht auf die Verwendung des Begriffs „statistisch signifikant“ zu empfehlen (Wasserstein et al. 2019). Dabei zielt sie nicht auf eine bloße Sprachregelung, sondern sieht die Notwendigkeit, die Methodik der wissenschaftlichen Datenanalyse und Ergebniskommunikation grundlegend zu reformieren.

Nach einer knappen Rekapitulation des gewöhnlichen Hypothesentests (Abschn. 2.1) und seiner Geschichte (2.2) behandeln wir einige Fehlinterpretationen des dabei gewonnenen p -Werts (2.3) und die daraus folgenden unerwünschten Konsequenzen (2.4). In Abschn. 3 diskutieren wir, welche Alternativen zum Standardverfahren sich bieten. Mit der Bayesschen Statistik betrachten wir einen dieser Vorschläge in Abschn. 3.1 etwas genauer. Mit einer knappen Zusammenfassung und einigen praktischen Hinweisen für eine bessere statistische Praxis schließen wir in Abschn. 4.

² Siehe auch <http://www.shinyapps.org/> für instruktive numerische Simulationen.

2 Geschichte, Probleme und Praxis des Hypothesentests

2.1 Das Standardverfahren des Null-Hypothesen Signifikanztests

Bevor wir das Standardverfahren des Null-Hypothesen Signifikanztests diskutieren, müssen wir eine knappe Bemerkung zum Wahrscheinlichkeitsbegriff vorausschicken. In der üblichen (frequentistischen) Auffassung ist die Wahrscheinlichkeit eines Ereignisses A mit der relativen Häufigkeit seines Auftretens verknüpft. Dies setzt also einen (zumindest hypothetisch) beliebig oft wiederholbaren Prozess voraus, der verschiedene Ausgänge haben kann (und im interessanten Fall auch hat). Die relative Häufigkeit des Auftretens von A bei n Ereignissen ($h_n(A) = \frac{n(A)}{n}$) ist ein Schätzwert für seine Wahrscheinlichkeit. Die Genauigkeit dieser Schätzung wächst nach Wahrscheinlichkeit mit zunehmendem Wert von n .³

Diese frequentistische Deutung stellt den Standard in der angewandten statistischen Literatur dar, etwa Bortz und Döring (2006) oder das monumentale Werk von Sachs (2004).⁴ Aus ihr folgt, dass Wahrscheinlichkeiten nur für die Werte von „Zufallsvariablen“ erklärt sind – und nicht etwa für „Hypothesen“. Nach dieser frequentistischen Auffassung von Wahrscheinlichkeit sind Hypothesen (etwa: „Intervention A ist wirksamer als Intervention B“, oder auch „Es gibt eine durch Menschen verursachte Klimakrise“) entweder wahr oder falsch. Da sie keinem wiederholbaren Prozess zugeordnet werden können, sind Wahrscheinlichkeitsaussagen hier nicht bloß falsch, sondern sinnlos. Diese Eigenschaft hat für das Testen von Hypothesen in der frequentistischen Statistik offensichtlich wichtige Auswirkungen. Wenden wir uns nun dem „Standardverfahren“ des Hypothesentests zu, wie es in der gängigen Literatur (etwa Bortz und Döring (2006)) eingeführt wird.

Für die Anwendung einer statistischen Methode zum Test einer Hypothese muss zunächst ein Kennwert identifiziert werden, der die hypothesenrelevanten Informationen zusammenfasst. Möchte man etwa die erhöhte Lernwirksamkeit einer Intervention im Vergleich zu einer konventionell unterrichteten Vergleichsgruppe überprüfen, kann man sich für die Differenz der Mittelwerte μ_i eines Wissenstests zwischen den beiden Gruppen interessieren.

Die sog. Null-Hypothese (H_0) formuliert in der Regel die Annahme, dass die Intervention ohne Effekt ist ($\mu_1 = \mu_2$). Die alternative Hypothese ($\mu_1 \neq \mu_2$ bzw. $\mu_1 > \mu_2$) wird als H_1 bezeichnet. Um die Null-Hypothese zu testen, untersucht

³ Einige Lehrbücher (etwa Bortz et al. (2008, S. 6)) formulieren prägnant, dass sich die Wahrscheinlichkeit als Grenzwert der relativen Häufigkeit für $n \rightarrow \infty$ ergibt: $p_A = \lim h_n(A)$. Hier kann jedoch nicht der analytische Grenzwertbegriff gemeint sein, denn statistische Fluktuationen können die Differenz zwischen relativer Häufigkeit und dem „Grenzwert“ immer wieder vergrößern. Man spricht deshalb auch von einer „stochastischen Konvergenz“ (Klenke 2006, S. 125f). Diese besagt vereinfachend: Nicht die relative Häufigkeit konvergiert gegen einen Grenzwert p_A , sondern nur die *Wahrscheinlichkeit* einer Abweichung konvergiert gegen Null, oder noch stärker und formal: $P(\lim h_n(A) = p_A) = 1$. Dieser Zusammenhang wird „starkes Gesetz der großen Zahl“ genannt und stellt die *Konsistenz* der frequentistischen Wahrscheinlichkeitsauffassung sicher. Da der stochastische Grenzwert jedoch den Begriff „Wahrscheinlichkeit“ bereits voraussetzt, kann er nicht zu seiner Definition beitragen (Caticha 2008, S. 27). Anders formuliert: Wahrscheinlichkeit ist ein logisch einfacher Begriff.

⁴ Diese Auffassung ist so üblich, dass sich unserer Erfahrung nach viele Anwender gar nicht bewusst sind, einer bestimmten „Wahrscheinlichkeitsschule“ anzugehören.

man eine Stichprobe mit n_1 bzw. n_2 Teilnehmenden und betrachtet die folgende Zufallsvariable (in üblicher Notation):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_{(\bar{X}_1 - \bar{X}_2)}}. \quad (1)$$

Unter bestimmten Bedingungen (etwa der Varianzgleichheit)⁵ und bei Zutreffen der Null-Hypothese folgt diese Variable der *Student-t*-Verteilung (mit $n_1 + n_2 - 2$ Freiheitsgraden).

Ergibt die Untersuchung für die Mittelwertdifferenz den Ausfall t , berechnet sich der zugehörige p -Wert als $p = P(|T| \geq t)$. Es handelt sich also um die Wahrscheinlichkeit, das beobachtete Datum (oder noch extremere) zu erhalten, unter Voraussetzung der Nullhypothese.

Gilt nun $p < 0,05$ (bzw. $p < 0,01$) spricht man von einem „statistisch signifikanten“ (bzw. „sehr signifikanten“) Ergebnis, bei dem die Stichprobe schlecht mit der Null-Hypothese verträglich ist. In diesem Fall wird H_0 abgelehnt und die Alternativhypothese akzeptiert (Bortz und Döring 2006, S. 494). Dort lesen wir ferner über den p -Wert:

Diese Wahrscheinlichkeit heißt Irrtumswahrscheinlichkeit (als diejenige Wahrscheinlichkeit, mit der wir uns irren würden, wenn wir die H_0 fälschlicherweise zugunsten von H_1 verwerfen). (Bortz und Döring 2006, S. 494)

Hier wird der p -Wert also ausdrücklich mit der Wahrscheinlichkeit identifiziert, einen sog. α -Fehler zu begehen. Die ebenfalls wichtige Wahrscheinlichkeit dafür, die Null-Hypothese irrtümlich nicht zu verwerfen, wird üblicher Weise mit β bezeichnet. Diese beiden Varianten werden auch Fehler 1. und 2. Art genannt.

2.2 Zur Geschichte des Standard-Signifikanztests

Das Standardverfahren des Null-Hypothesen Signifikanztests kombiniert Verfahren und Begriffe, die ab der 1920er und 30er Jahre innerhalb der frequentistischen Statistik von Ronald A. Fisher sowie Jerzy Neyman und Egon Pearson entwickelt wurden. Auf Fisher geht etwa die Popularisierung des p -Wertes zurück (Fisher 1925).

Kurioser Weise wurde das Konzept einer alternativen Hypothese H_1 sowie die Unterscheidung von α - und β -Fehlern von Neyman und Pearson (1933) in *Abgrenzung* zu Fishers Methodologie entwickelt. Bei diesem Verfahren wird zudem darauf verzichtet, auf der Grundlage eines *einzelnen* Experiments auf die Gültigkeit bzw. Plausibilität einer Hypothese zu schließen. Stattdessen ist das „Signifikanzniveau“ α gar keine Eigenschaft der Daten, sondern des Versuchsdesigns. Anstatt (wie Fisher) auf ein „induktives Schließen“ im Einzelfall zielen Neyman und Pearson auf eine „Handlung“, die auf lange Sicht (d.h. im frequentistischen Sinne) vorher festgelegte geringe Fehlerraten α und β aufweist (Gigerenzer et al. 1989, Kap. 3.4).

⁵ Es existieren auch Verfahren, um auf diese meist unrealistische Voraussetzung verzichten zu können.

Man erkennt deutlich den Unterschied zwischen dem p -Wert (gewonnen aus einer einzelnen Stichprobe von Werten einer Zufallsvariablen und für die Evidenz gegen eine einschlägige Nullhypothese verwandt) und dem α -Niveau (einem im Untersuchungsdesign festgelegten Fehler-Niveau, das seine frequentistische Bedeutung erst bei häufiger Wiederholung des Testverfahrens erhält). Die im vorangegangenen Abschnitt zitierte Bemerkung aus Bortz und Döring (2006, S. 494) enthält nun eine typische Gleichsetzung bzw. Verwechslung beider Größen. Hier wird (wie in zahlreichen Lehrbüchern, vgl. die Analyse von Hubbard und Bayarri (2003)) der p -Wert gleichzeitig als Evidenz aus der Einzelmessung *und* als (quasi experimentelle) Fehlerrate α aufgefasst.

Zahlreiche Autor*innen haben darauf hingewiesen, dass diese inkohärente Kombination von Ideen rivalisierender Statistik-Schulen zu den Missverständnissen beigetragen hat, welche die Interpretation des p -Wertes begleiten (Gigerenzer et al. 1989; Goodman 1993; Lehmann 1993). Gigerenzer et al. (2004) bezeichnen das routinemäßige Testen der Null-Hypothese als „Null Ritual“ und formulieren sarkastisch:

[...] (T)he null ritual originated neither from Fisher nor from any other renowned statistician and does not exist in statistics proper. It was instead fabricated in the minds of statistical textbook writers in psychology and education.

Wenden wir uns nun also der Frage zu, welche Bedeutung der p -Wert hat und welche weiteren Interpretationen unzulässig sind.

2.3 Was bedeutet der p -Wert (nicht)?

Über die tatsächliche Bedeutung des p -Wertes lässt sich wenig sagen, das über seine technische Definition hinausgeht. Es handelt sich schlicht um die Wahrscheinlichkeit dafür, die beobachteten Daten oder noch extremere Ausfälle (D) zu messen, gegeben die Null-Hypothese (H_0). Symbolisch ausgedrückt:⁶

$$p = P(D|H_0). \quad (2)$$

Diese Kenngröße quantifiziert die Verträglichkeit der Daten mit der Null-Hypothese. Jede Interpretation des p -Wertes muss dabei berücksichtigen, dass die Gültigkeit von H_0 bei seiner Berechnung *vorausgesetzt* wurde, sowie, dass es sich um eine „Datenwahrscheinlichkeit“ (und keine „Hypothesenwahrscheinlichkeit“) handelt.

Betont werden sollte jedoch, dass der p -Wert diese Bedeutung nur besitzt, falls alle *Voraussetzungen* für seine Berechnung erfüllt sind. Dazu zählen an erster Stelle die Bedingungen für die Anwendung des zugrunde gelegten statistischen Modells, die Randomisierung der Stichprobe, die *score* Reliabilität, die Validität der Kon-

⁶ Um auszudrücken, dass auch extremere Ausfälle betrachtet werden, findet sich manchmal auch die sinnvolle Schreibweise „ D^+ “ statt „ D “. Die Hypothese als Argument einer bedingten Wahrscheinlichkeit wird im Bayesschen Rahmen (siehe Abschn. 3.1) eine präzise Bedeutung erhalten.

strukte etc. (Kline 2013, S. 13f). In der Praxis werden diese Voraussetzungen nie streng erfüllt sein, woraus bereits eine gewisse Variabilität des p -Wertes folgt.⁷

Wenden wir uns nun der Frage zu, welche irrigen Annahmen sich mit der Bedeutung des p -Werts verbinden. In der Literatur finden sich zahlreiche Listen mit verbreiteten Missverständnissen. Goodman (2008) identifiziert ein „dreieckiges Dutzend“ solcher Fehlschlüsse, und Kline (2013, S. 95) spielt auf die Großwildjagd an, wenn er seine Liste „the big five“ nennt. Unsere Aufzählung trifft hier eine Auswahl, die wir scherzhaft die „fantastischen Vier“ nennen wollen:

1. **Der inverse-Wahrscheinlichkeits-Fehlschluss:** „ $p < 0,05$ bedeutet, dass H_0 weniger als 5% Wahrscheinlichkeit besitzt, wahr zu sein.“ Da p jedoch unter der Annahme berechnet wurde, dass H_0 wahr ist, kann es nicht gleichzeitig die *Wahrscheinlichkeit* dafür sein, dass H_0 wahr ist. Der Fehler dieser Interpretation lässt sich auch wie folgt erläutern: Diese Deutung identifiziert den p -Wert mit der Wahrscheinlichkeit $P(H_0|D)$. Im Vergleich zu Gl. 2 haben hier die Daten D und H_0 die Plätze getauscht (hier auch „invertieren“ genannt). Aus zwei Gründen ist dies inkorrekt. Zum einen lehrt der Satz von Bayes, dass $P(A|B) \neq P(B|A)$ gilt. Zum anderen können innerhalb der frequentistischen Statistik keine Wahrscheinlichkeitsaussagen über das Zutreffen von Hypothesen getroffen werden.
2. **Der Zufall-Wahrscheinlichkeits-Fehlschluss:** „ $p < 0,05$ bedeutet, dass die Wahrscheinlichkeit des Messergebnisses, bloßer Zufall zu sein, weniger als 5% beträgt.“ Die nachvollziehbare Intuition hinter diesem Fehlschluss lautet, dass unter der verbreiteten Null-Hypothese („kein Effekt“) jeder Unterschied lediglich dem Stichprobenfehler (vulgo: dem „Zufall“) geschuldet ist. Bei der p -Wert Berechnung wird H_0 (d.h. eine zufällige Verursachung) jedoch *vorausgesetzt*. Wollte man tatsächlich quantifizieren, wie verträglich der Ausgang mit der Annahme einer zufälligen Verursachung ist, müsste man deshalb erneut die Hypothesenwahrscheinlichkeit $P(H_0|D)$ berechnen. Es handelt sich hier also um eine Variante des Fehlschlusses zur inversen Wahrscheinlichkeit (Carver 1978).⁸
3. **Der α -Fehler-Fehlschluss:** „ $p < 0,05$ bedeutet, dass die Wahrscheinlichkeit, H_0 in einer Studie irrtümlich zu verwerfen (also einen α -Fehler zu begehen), kleiner als 5% ist.“

Zu diesem Missverständnis wird man aufgrund der formalen Ähnlichkeit zwischen p -Wert und α -Niveau eingeladen. In Abschn. 2.2 haben wir bereits auf den Unterschied dieser beiden Größen hingewiesen. An dieser Stelle kann jedoch auch wie folgt argumentiert werden: Eine solcher Wahrscheinlichkeitsaus-

⁷ Bei der (im Wortsinn) Logik des Hypothesentests handelt es sich also um eine Abwandlung des Widerspruchsbeweises. In der Aussagenlogik gilt, dass aus „ $H \rightarrow \neg D$ “ sowie „ D “ auf „ $\neg H$ “ geschlossen werden kann („modus tollens“). Mit anderen Worten: Man verwirft eine Voraussetzung, wenn ihre Folgerung nicht zutrifft. Im Hypothesentest wird nun die logische Negation ($\neg D$) durch eine Wahrscheinlichkeitsaussage („ D ist sehr unwahrscheinlich“) ersetzt. Dies ist natürlich keine (logisch) korrekte Schlussfigur – aber die Hoffnung scheint zu sein, dass der Schluss wenigstens mit hoher Wahrscheinlichkeit zutrifft. In Cohen (1997, S. 23) wird das Problem dieser Argumentation genauer (und unterhaltsam) erläutert und als weiterer Einwand gegen den Hypothesentest verwendet.

⁸ Natürlich können Nullhypothesen auch die Anwesenheit eines bestimmten Effekts behaupten. In diesem Fall bestünde der analoge Fehlschluss darin, jede *Abweichung von dieser Nullhypothese* als bloß zufällig (mit Wahrscheinlichkeit p) anzusehen.

sage über einen „Fehler 1. Art“ wäre eine Aussage darüber, ob H_0 wahr ist, denn nur dann begeht man ja einen solchen Irrtum. Es handelt sich also erneut um eine Variante des Fehlschlusses zur inversen Wahrscheinlichkeit.

4. **Der Replikations-Fehlschluss:** „ $p < 0,05$ bedeutet, dass die Wahrscheinlichkeit, ein solches Ergebnis *nicht* replizieren zu können, bei unter 5% liegt.“ Hier wird also die Wahrscheinlichkeit der erfolglosen Replikation mit p identifiziert. Dies stellt eine Wahrscheinlichkeitsaussage für das Auftreten von bestimmten Daten dar – ohne jedoch an die Bedingung „ H_0 ist gültig“ geknüpft zu sein. Diese Interpretation kann also ebenfalls nicht zutreffen, obwohl tatsächlich unter bestimmten Modellannahmen ein indirekter Zusammenhang zwischen p -Wert, anderen Kenngrößen und der Replikationswahrscheinlichkeit formuliert werden kann (Greenwald et al. 1996).

Interessanterweise führen also all diese Fehlschlüsse zu einer Überschätzung der Aussagekraft des p -Werts. Wie verbreitet sind diese Missverständnisse, und welche Folgen haben sie? Wenden wir uns zunächst der ersten Frage zu. Eine Reihe von empirischen (*sic*) Untersuchungen habe diese Frage untersucht und sind zu einheitlichen Ergebnissen gelangt. So haben z. B. Haller und Krauss (2001) an sechs deutschen Universitäten eine Umfrage unter Studierenden und Lehrenden in der Psychologie durchgeführt. Sie haben gefunden, dass fast alle der befragten fortgeschrittenen Studierenden einer falschen p -Wert Interpretation anhängen, während bei den befragten Lehrpersonen immer noch 80-90% eine unzutreffende Interpretation vertraten (siehe dazu auch Gigerenzer et al. (2004)). Es kann kaum bezweifelt werden, dass solche Untersuchungen in der Bildungsforschung zu ähnlichen Ergebnissen führen würden.

So findet man in einem aktuellen Lehrbuch zu Forschungsmethoden der Naturwissenschaftsdidaktik zum Begriff „Signifikanz“ folgenden Eintrag im Glossar:

Ein Ergebnis ist statistisch signifikant (bedeutsam), wenn die Wahrscheinlichkeit, dass es zufällig zustande gekommen ist, klein ist. Das Signifikanzniveau muss definiert werden. Häufig gelten Aussagen, bei denen mit einem Signifikanztest eine Irrtumswahrscheinlichkeit unter 5% gefunden wird, als signifikant. Krüger, Parchmann und Schecker (2014, S. 403)

Die irreführende Gleichsetzung von „statistisch signifikant“ mit „bedeutsam“ wird hier explizit vorgenommen (vgl. Abschn. 2.4). Die Formulierung von der Wahrscheinlichkeit eines „zufälligen Zustandekommens“ legt nahe, dass hier der „Zufall-Wahrscheinlichkeits-Fehlschluss“ vorliegt (Punkt 2 in der Liste von Abschn. 2.3). Die Sprechweise von der „Irrtumswahrscheinlichkeit“ verleitet ebenfalls zu dem „ α -Fehler-Fehlschluss“ (Punkt 3 in der Liste).

Lehrbücher zur Statistik und Datenauswertung gehen auf die Problematik der p -Werte meist nur recht oberflächlich oder gar nicht ein. Das bereits zitierte Werk von Bortz und Döring (2006) streift diese Debatte und erwähnt die Forderung von Kline (2013) (Bortz und Döring beziehen sich auf die 1. Auflage von 2004), den Begriff „statistisch signifikant“ nicht mehr zu verwenden. Ihre Replik ist entwaffnend:

Auch wenn wir diese harsche Kritik im Wesentlichen nachvollziehen können, wird der Begriff der statistischen Signifikanz in diesem Buch nicht gestrichen, zumal so manche Human- oder Sozialwissenschaftler froh sind, diesen Begriff überhaupt erst einmal richtig verstanden zu haben.

Ihre Position sei stattdessen, den traditionellen Signifikanztest durch Betrachtungen von Teststärke, Effektgröße und Konfidenzintervalle zu ergänzen (*ibid.* S. 601).

Dabei ist auch für diese Kenngrößen die Interpretation problematisch (siehe Abschn. 3.1.2 für eine Diskussion von Vertrauensintervallen). Zahlreiche empirische Studien in der Bildungsforschung berechnen z.B. Cohens d als Maß für die Effektstärke und folgen der Konvention, nach der die Werte von 0,2, 0,5 und 0,8 zwischen kleinen, mittleren und großen Effekten unterscheiden. Dabei weisen auch die vielzitierten Bortz und Döring (2006, S. 626) darauf hin, dass es sich dabei lediglich um eine grobe Orientierungshilfe handelt, die dem jeweiligen Forschungsfeld angepasst werden muss. In Bakker et al. (2019) haben Mathematikdidaktiker*innen eine Liste von zwölf Hinweisen formuliert, die bei der Einschätzung von Effektstärkemaßen beachtet werden sollten. Der einfache Schluss vom numerischen Wert von z.B. d_{Cohen} auf die inhaltliche „Stärke“ des Effekts ist nicht zu rechtfertigen.⁹

2.4 Konsequenzen aus der Fehlinterpretation des p -Wertes

Unterliegen Autor*innen von quantitativen empirischen Studien einem oder mehreren der oben diskutierten Missverständnisse, setzt dies natürlich nicht automatisch den Wert ihrer Arbeit herab. Allerdings haben unter anderem Loftus (1996), Sedlmeier (1996) und Gigerenzer et al. (2004) die Frage aufgeworfen, in welchem Sinne die Fokussierung auf das Testen (und Verwerfen) von Null-Hypothesen zu einer unerwünschten Engführung von Forschungsprogrammen führen kann. Diesen Gefahren wollen wir uns nun zuwenden.

2.4.1 Die Verwechslung von „statistisch signifikant“ und „bedeutsam“

Ein Ergebnis mit $p < 0,05$ „statistisch signifikant“ zu nennen, folgt einer bloßen Konvention. Die Attraktivität der Formulierung rührt wohl auch daher, dass im alltäglichen Sprachgebrauch „signifikant“ mit „wesentlich“, „bedeutsam“ oder „wichtig“ konnotiert. Diese Identifikation ist jedoch ungerechtfertigt, denn die Relevanz eines Resultats hängt ganz wesentlich auch von anderen Faktoren ab. Einige davon lassen sich vielleicht sogar quantifizieren (etwa durch die Effektstärke). Andere können gar nicht durch statistische Kenngrößen ausgedrückt werden. Mit anderen Worten: „statistisch signifikant“ und „wissenschaftlich signifikant“ sind zwei unterschiedliche Kategorien.

⁹ So zeigen Cheung und Slavin (2016), dass z.B. große Studien (oder solchen mit standardisierten Testinstrumenten) im Mittel nur halb so große Effekte aufweisen wie kleine Studien (oder solche, die eigene Testwerkzeuge verwenden). Diese bemerkenswerte Korrelation zwischen Forschungsdesign und Effektstärke kompromittiert somit die Ergebnisse aller Metaanalysen, die über Studien mit unterschiedlichem Design mitteln.

Die Tatsache, dass nicht-signifikante Ergebnisse eine geringere Veröffentlichungswahrscheinlichkeit haben (Rosenthal 1979), schafft zudem unerwünschte Forschungsanreize. Auf diese Weise wird provoziert, Freiheiten in der Auswahl von Analyseverfahren so zu nutzen, dass ein signifikanter p -Wert erreicht wird. Die Grenze zum wissenschaftlichen Fehlverhalten ist hier fließend (Head et al. 2015). Werden solche Ergebnisse schließlich in Metaanalysen kombiniert, führt dies zu einer systematischen Überschätzung der Effektstärke (Simonsohn et al. 2014). Dabei werden gerade Metaanalysen als probates Mittel angesehen, die Ergebnisse kleinerer Studien zu einem aussagekräftigeren Resultat zu bündeln.

2.4.2 Stagnation in der Theorieentwicklung und Vernachlässigung von Kontext

Gigerenzer (1998) hat darauf hingewiesen, dass ritualisiertes Testen (und Ablehnen) von Null-Hypothesen einer Vernachlässigung der Theorieentwicklung Vorschub leisten kann. Dominiert das skizzierte Standardverfahren, reichen bereits relativ unspezifische Hypothesen für die Gewinnung von (lediglich) „statistisch signifikanten“ und damit insinuiert bedeutsamen Resultaten.

Goodman (1999, S. 1001) sieht ebenfalls eine Entwicklung zur Verarmung des Diskurses. Er berichtet von der Tendenz, dass Veröffentlichungen mit großen randomisierten und kontrollierten Studien in der Medizin den Forschungsstand kaum diskutieren. Er vermutet eine Ursache in dem Missverständnis, auf der Grundlage einzelner Studien bereits Aussagen mit geringer Fehlerwahrscheinlichkeit gewinnen zu können.

2.4.3 Das geringe Ansehen von Replikationsstudien

Unterliegt man dem „Replikations-Fehlschluss“, erscheint die Wiederholung einer Studie wenig sinnvoll oder notwendig. Dies könnte eine zusätzliche Erklärung für deren geringe Anzahl sein Kline (2013, S. 269).

Makel und Plucker (2014) haben ca. 160.000 Veröffentlichungen von 100 führenden Zeitschriften in den Erziehungswissenschaften seit 1990 analysiert. Lediglich 0,13% dieser Arbeiten haben eine Replikationsstudie zum Gegenstand.

Interessant ist hier natürlich die Frage, ob sich ebenfalls eine „Replikationskrise“ zeigt (siehe Abschn. 1). Die zitierte Arbeit von Makel und Plucker (2014) findet in den Erziehungswissenschaften, dass 67% der Untersuchungen den ursprünglichen Befund replizieren können. Dies ist eine viel höhere Rate als im medizinischen Forschungsfeld und in Teilen der Psychologie (Open Science Coll. 2015). Makel und Plucker (2014) bemerken jedoch, dass häufig die selben Autor*innen an einer Replikation beteiligt sind, die bereits die ursprüngliche Studie veröffentlicht haben. Betrachtet man lediglich Replikationen von anderen Autor*innen, sinkt die Quote der erfolgreichen Replikationen auf 54% – weit entfernt von einer (naiv angenommen) Replikationswahrscheinlichkeit von 95%.

3 Konsequenzen für die wissenschaftliche Praxis

Die Hoffnung, dass Mechanismen der wissenschaftlichen Selbstkorrektur die oben geschilderten Probleme (zumindest mittel- oder langfristig) lösen, scheint trügerisch (Pashler und Harris 2012; Ioannidis 2012). In der Psychologie werden stattdessen Reformen der wissenschaftlichen Praxis auf vielen Ebenen diskutiert (Pashler und Wagenmakers 2012). Die *Open Science Collaboration* stellt eine solche Initiative dar. Sie koordiniert Replikationsstudien und setzt sich für die Veröffentlichung der Rohdaten und Analyseprotokollen empirischer Untersuchungen ein (Carpenter 2012). Auf diese Weise werden empirische Befunde nachvollziehbarer und die Gefahr des wissenschaftlichen Fehlverhaltens geringer. Das Dokument „*A Manifesto for reproducible science*“ enthält ebenfalls zahlreiche Reformvorschläge. Einige Stichworte sind hier: Ausbildung, Anreizsysteme und *peer review* (Munafò et al. 2017).

Vor dem Hintergrund unserer Ausführungen erscheint es geboten, die Praxis der Datenauswertung und die Kommunikation der Resultate zu reformieren. Genau in diese Richtung zielt die Empfehlung der *American Statistical Association*, den Begriff „statistisch signifikant“ nicht mehr zu verwenden (Wasserstein et al. 2019). Sie erläutern:

For example, no p -value can reveal the plausibility, presence, truth, or importance of an association or effect. Therefore, a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant. Yet the dichotomization into „significant“ and „not significant“ is taken as an imprimatur of authority on these characteristics.

Offensichtlich geht es den Autor*innen dabei nicht um eine bloße Redeweise. Vielmehr weisen Wasserstein et al. (2019) darauf hin, dass die daraus abgeleitete Unterscheidung in „relevante“ und „irrelevante“ Ergebnisse zu einer fehlerhaften Veröffentlichungspraxis führt. Die Darstellung des Forschungsstandes, etwa in Übersichtsartikeln, wird auf diese Weise stark verzerrt. Mit drastischen Worten resümieren sie:

For the integrity of scientific publishing and research dissemination, therefore, whether a p -value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight.

Bei all dem stellt sich jedoch die Frage, welche konkreten Alternativen sich zur Bewertung von Hypothesen bieten. Die Arbeit von Wasserstein et al. (2019) ist das *Editorial* zu einem Sonderheft der Zeitschrift *The American Statistician*, das sich genau dieser Frage widmet. Es enthält 43 Beiträge, deren Gemeinsamkeit unter anderem darin liegt, auf das Fehlen einer „Patentlösung“ für dieses Problem hinzuweisen. Kein mechanisch anwendbares Schlussverfahren kann die inhaltliche – und damit immer auch subjektive – Auseinandersetzung mit dem Forschungsgegenstand ersetzen.

Unstrittig ist jedoch die Notwendigkeit einer ausreichenden Teststärke $1 - \beta$, da Effekte sonst systematisch überschätzt werden (Colquhoun 2014). Zahlreiche Arbeiten des Sonderhefts propagieren zudem die Bayessche Statistik als alternativen Theorierahmen, der in zahlreichen Forschungsbereichen bereits gut etabliert ist.

Diesen Ansatz wollen wir im Folgenden etwas genauer erläutern. Auch hier gilt natürlich, dass er nicht als neues „Standardverfahren“ aufgefasst werden sollte. Unsere Darstellung der Bayesschen Statistik hat die zusätzliche Funktion, die konzeptionellen Besonderheiten bzw. Schwierigkeiten der frequentistischen Statistik noch einmal von einer anderen Seite zu beleuchten.

3.1 Anmerkungen zur Bayesschen Alternative

Wir haben gesehen, dass fast sämtliche Fehlinterpretationen des p -Wertes ihre Wurzel darin haben, irrtümlich einer Hypothese eine Wahrscheinlichkeit zuschreiben zu wollen – und diese Wahrscheinlichkeit gar mit $1 - p$ (für die Alternative) zu quantifizieren. Selbstverständlich soll Forschung zu Urteilen über die Plausibilität bzw. fast sichere Gültigkeit von Aussagen führen. Aber dazu ist das schematische Hypothesentesten innerhalb des frequentistischen Paradigmas für sich allein genommen zu schwach.

Aus Arbeiten von Harold Jeffreys, Richard Cox und Edwin Jaynes (sowie auf der Grundlage anderer konzeptioneller Vorarbeiten) hat sich seit den 1950er Jahren ein alternativer Theorierahmen entwickelt, die sog. „Bayessche Statistik“. In ihr quantifiziert die Wahrscheinlichkeit den „Grad einer Plausibilität“. Der Slogan in der englischen Fachliteratur lautet „probability is degree of belief“ bzw. „degree of plausibility“ (Jaynes 2003, S. 17). In diesem Sinne vereinbart also der „Bayesianer“:¹⁰

$P(A)$ = Grad der Überzeugung, dass das Ereignis A eintritt

Dieser *subjektivistische* Wahrscheinlichkeitsbegriff ist (im Gegensatz zum frequentistischen Begriff) nicht an einen wiederholbaren Vorgang geknüpft und kann auch auf Hypothesen angewendet werden. Er bleibt jedoch völlig akademisch, solange man keine Berechnungsvorschrift für diese Wahrscheinlichkeit kennt. Tatsächlich kann in Abwesenheit von Kontextwissen bzw. von Beobachtungen, über deren Ausgang die Hypothese eine Aussage trifft, ihr „Plausibilitätsgrad“ lediglich geraten werden. In Anwesenheit von Daten oder sonstigem Hintergrundwissen kann die Berechnung dieser Wahrscheinlichkeit jedoch als Problem der bedingten Wahrscheinlichkeit aufgefasst werden. Zu ihrer Bestimmung verwendet man also den Satz von Bayes:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{\sum_i P(B|A_i) P(A_i)}. \quad (3)$$

¹⁰ In Tschirk (2019, S. 17) wird erläutert, wie aus sog. Plausibilitätsannahmen für $P(A)$ die Kolmogorow-Axiome der Wahrscheinlichkeitsrechnung zu motivieren sind. Wir folgen im Weiteren einer untechnischen Argumentation.

Die zweite Formulierung nutzt aus, dass für eine disjunkte Zerlegung A_i des Ergebnisraums Ω die Beziehung $P(B) = \sum_i P(B|A_i)P(A_i)$ gilt. Für unseren Fall gilt dann („ H “ bezeichnet hier die Hypothese, über die mit Hilfe der „Daten“ eine Aussage getroffen werden soll):

$$\underbrace{P(H|\text{Daten})}_{\text{A-posteriori-Wkt.}} \propto \underbrace{P(\text{Daten}|H)}_{\text{Likelihood}} \cdot \underbrace{P(H)}_{\text{A-priori-Wkt.}} \quad (4)$$

Die linke Seite der Gleichung wird A-posteriori Wahrscheinlichkeit genannt und setzt sich aus der *Likelihood* der konventionellen Statistik $P(\text{Daten}|H)$ ¹¹ und der A-priori-Wahrscheinlichkeit (in der englischsprachigen Literatur als *prior* bezeichnet) $P(H)$ zusammen. Letztere repräsentiert die Kenntnis *vor* der Messung. In diesem Konzept wird also formalisiert, wie neue Informationen den Kenntnisstand modifizieren.¹²

Das Theorem von Bayes und die subjektivistische Interpretation von Wahrscheinlichkeit sind also unmittelbar miteinander verknüpft, da, im Gegensatz zur frequentistischen Deutung, hier keine „Berechnungsvorschrift“ für Wahrscheinlichkeiten unmittelbar aus der Interpretation folgt. Dies erklärt den Namen „Bayessche Statistik“, obwohl bedingte Wahrscheinlichkeiten und der Satz von Bayes natürlich ebenfalls etablierte Inhalte der konventionellen (d.h. frequentistischen) Statistik darstellen.¹³

Offensichtlich kann unsere Darstellung der Bayesschen Statistik nur sehr skizzenhaft sein, und wir verweisen etwa auf Tschirk (2019) für eine systematische Behandlung. Eine glänzende Einführung bieten auch die kommentierte Leseliste von Etz et al. (2018) sowie Gelman et al. (2014). Ein Softwarepaket, das auch Bayessche Verfahren unterstützt, wurde an der Universität Amsterdam entwickelt: <https://jasp-stats.org/>.

3.1.1 Der Hypothesentest in der Bayesschen Statistik

Unter der Annahme, dass der p -Wert einer Irrtumswahrscheinlichkeit entspricht (in Abschn. 2.3 als „ α -Fehler-Fehlschluss bezeichnet“), wird man zu der Vorstellung verleitet, dass bei einem „statistisch signifikanten“ Ergebnis mindestens eine 19 : 1

¹¹ Es sollte darauf hingewiesen werden, dass der Ausdruck $P(D|H)$ (*Likelihood*) nicht mit dem p -Wert verwechselt werden darf, da letzterer nicht nur vom beobachteten Ereignis abhängt, sondern auch „extremere“ Ausfälle berücksichtigt. Ferner sei bemerkt, dass man im Fall einer stetigen Verteilung hier mit Dichtefunktionen arbeitet.

¹² Die Auswahl der A-priori-Verteilung stellt ein viel diskutiertes Problem der Bayesschen Statistik dar. Die Subjektivität, die auf diese Weise Einzug hält, zählt vermutlich zu den häufigsten Kritikpunkten an diesem Ansatz (Efron 1986). Gleichzeitig betonen Anhänger der Bayesschen Statistik den Vorteil, dass durch die A-priori-Verteilung das Vorwissen auf kohärente Art in die Analyse einbezogen werden kann (vgl. auch Tschirk (2019, Kap. 6.2)).

¹³ Für das Verhältnis zwischen Bayesscher und frequentistischer Statistik gilt ganz grundsätzlich, dass die Wahrscheinlichkeitsrechnung dieselbe bleibt. Neu ist die Ausdehnung des (subjektiven) Wahrscheinlichkeitsbegriffes auf Hypothesen; anders ist vor allem das induktive „Lernen aus Daten“ mit dem expliziten Einbau von A-priori-Verteilungen, die das Vorwissen oder auch dessen Fehlen spiegeln. Aus der Kombination dieser Elemente folgt schließlich die gesuchte A-posteriori-Verteilung. Es werden dabei unbekannte Parameter als „zufällig“ betrachtet und nicht die Daten.

Wahrscheinlichkeit für die Richtigkeit der alternativen Hypothese im Vergleich zur Null-Hypothese besteht. Warum dies unzutreffend ist, haben wir bereits ausführlich begründet. Innerhalb der Bayesschen Statistik kann eine solche Quote $P(H_1|D)$: $P(H_0|D)$ jedoch direkt angegeben werden:

$$\underbrace{\frac{P(H_1|D)}{P(H_0|D)}}_{\text{A-posteriori Quote}} = \underbrace{\frac{P(D|H_1)}{P(D|H_0)}}_{\text{Bayes-Faktor}} \cdot \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{A-priori Quote}}. \quad (5)$$

Ein solches Verhältnis zu betrachten kann unter Umständen auch inhaltlich sinnvoller sein. Schließlich mögen im Einzelfall die Daten weder durch H_0 noch durch H_1 beschreibbar sein.

Das in Gl. 5 auftretende Verhältnis der *Likelihoods* der Daten unter der Bedingung von H_1 bzw. H_0 wird auch als Bayes-Faktor (BF₁₀, sprich „BF Eins Null“, da die mit „1“ indizierten Größen im Zähler auftreten) bezeichnet.

Von der knappen Notation lasse man sich aber nicht täuschen: Die Berechnung des Zählers von BF₁₀ ($P(D|H_1)$) ist i. allg. aufwendig und mehrdeutig, wenn die Alternativhypothese die unspezifische Form $\mu_1 - \mu_2 \neq 0$ hat (Kass und Raftery 1995). In diesem Fall muss schließlich eine Verteilung der Modellparameter angenommen werden.¹⁴ Dies entspricht der Wahl einer A-priori-Wahrscheinlichkeitsverteilung für H_1 . Dennoch konnten Bayarri et al. (2016) zeigen, in welchem Sinne diese Größe auch eine rein frequentistische Interpretation besitzt. Auf diese Weise ist die Verwendung des Bayes-Faktors auch im Rahmen der frequentistischen Statistik akzeptabel und liefert ein alternatives „Evidenzmaß“ für die Entscheidung, seine Vorannahmen über die Hypothesen im Licht der Daten zu revidieren.¹⁵ Bei einem Wert < 1 drückt der Bayes-Faktor BF₁₀ Evidenz *zugunsten* der Null-Hypothese aus.

Die Berechnung des Bayes-Faktors bleibt allerdings aus den angedeuteten Gründen aufwendig (Kass und Raftery 1995), und dies stellt vermutlich ein ganz praktisches Hemmnis für seine weitere Verbreitung dar. Es gibt jedoch die Möglichkeit, aus dem p -Wert der konventionellen Statistik *obere Schranken* für den Bayes-Faktor abzuleiten (Held und Ott 2018).

Thomas J. Faulkenberry hat ein besonders einfaches Näherungsverfahren für ANOVA und t -Test Anwendungen entwickelt, das auf dem *Bayesian Information Criterion* (Raftery 1995) basiert. Als Anwendung diskutiert Faulkenberry (2018) eine Studie, bei der die Wirkung von Koffein auf die Merkfähigkeit untersucht wurde. Versuchs- und (Placebo behandelte) Kontrollgruppe bestanden aus zusammen $n = 73$ Probanden. Der t -Test ergab hier $t(71) = 2,0$, was (zweiseitig) einem Wert

¹⁴ Es gilt dann: $P(D|H_1) = \int P(D|\mu)P(\mu|H_1)d\mu$. Man beachte, dass für die Berechnung der *Likelihood* bezüglich der Null-Hypothese ($\mu_1 = \mu_2$) keine A-priori-Verteilung spezifiziert werden muss.

¹⁵ Die Literatur enthält verschiedene Vorschläge für die Interpretation von Bayes-Faktoren. So schlagen z. B. Kass und Raftery (1995) vier Kategorien vor: BF₁₀ von 1 bis 3 („Not worth more than a bare mention“), 3–20 („Positive“), 20–150 („Strong“) und > 150 („Very strong“). Offensichtlich sind solche Kategorien ebenfalls der Kritik ausgesetzt, die im Falle des p -Wertes erhoben wird, nämlich eine willkürliche Klassifizierung von Forschungsergebnissen vorzunehmen. Kass und Raftery (1995, S. 777) schränken deshalb ein: „[...] these categories are [...] a rough descriptive statement about standards of evidence in scientific investigation.“

von $p = 0,049$ entspricht. Auf dieser Grundlage kann also eine „signifikante“ Wirkung von Koffein behauptet werden. Die Näherungsformel für den Bayes-Faktor ergibt jedoch (Faulkenberry 2018, S. 38):

$$\text{BF}_{10} \approx \sqrt{\frac{\left(1 + \frac{t^2}{n-2}\right)^n}{n}} \approx 0,87 \quad (6)$$

Ein Bayes-Faktor $\text{BF}_{10} < 1$ liefert jedoch keinen Hinweis *gegen*, sondern ganz im Gegenteil *für* die Null-Hypothese. Dieses Resultat kann intuitiv verstanden werden, obwohl es in der Literatur als „Lindley Paradoxie“ bezeichnet wird (Lindley 1957). Falls die p -Wert-Verteilung unter der *alternativen* Hypothese bei kleinen p -Werten ($p \ll 0,05$) konzentriert ist, ist die Wahrscheinlichkeit für $p \approx 0,05$ bei Zutreffen der Null-Hypothese größer als bei Zutreffen von H_1 . Der Bayes-Faktor analysiert genau dieses Verhältnis, während der p -Wert des konventionellen Tests nur von der Null-Hypothese abhängt. Dieses Beispiel illustriert somit ein weiteres Problem bei der Interpretation des p -Werts.

In der Regel führt die Bayes-Faktor Analyse jedoch lediglich zu einer *konservativeren* Einschätzung der jeweiligen „Evidenz“ für die Ablehnung der Null-Hypothese. Aus dieser Tatsache könnte also gefolgert werden, dass eine Absenkung der Signifikanzschwelle bereits einen wertvollen Beitrag zu einer verbesserten statistischen Praxis liefere. In der Tat ist dieser Vorschlag in der Vergangenheit immer wieder diskutiert worden. Zuletzt haben Benjamin et al. (2018) die Forderung der Herabsetzung der Signifikanz-Schwelle auf $p < 0,005$ erhoben. Diese Forderung bedeutet im Übrigen nicht, dass die Ergebnisse identischer Studiendesigns nur strenger bewertet werden. Die Herabsetzung der Signifikanzschwelle bei unveränderter Stichprobengröße führt nämlich zu einer drastischen Reduzierung der Teststärke. Soll diese konstant gehalten werden, muss der Stichprobenumfang beträchtlich vergrößert werden – im konkreten Bsp. um ca. 70%. In Trafimow et al. (2018) findet sich Kritik an diesem Vorschlag.

Man beachte zudem, dass für die Angabe des Verhältnisses der A-posteriori-Wahrscheinlichkeiten von H_1 und H_0 (also innerhalb der Bayesschen Statistik) der Bayes-Faktor mit dem Verhältnis der A-priori-Wahrscheinlichkeiten multipliziert werden muss. Mit anderen Worten: Der gesamte Kontext, vorherige Studienergebnisse und theoretische Argumente müssen immer in die Bewertung der Daten einfließen.

3.1.2 Bayessche und frequentistische Vertrauensintervalle

In der Literatur findet sich häufig der Hinweis, dass die Angabe von Vertrauens- bzw. Konfidenzintervallen einen wichtigen Beitrag zur Verbesserung der statistischen Praxis darstelle. Aber auch die Interpretation von derartigen Intervallen ist in der frequentistischen Statistik nicht frei von Missverständnissen.

Weit verbreitet ist etwa die Vorstellung, das Konfidenzniveau gebe die Wahrscheinlichkeit dafür an, dass sich der Wert des untersuchten Parameters innerhalb des beobachteten Konfidenzintervalls befindet (Hoekstra et al. 2014). In der fre-

quentistischen Statistik lässt sich jedoch eine solche Wahrscheinlichkeitsaussage gar nicht formulieren, da der Parameter (eine unbekannte Größe mit festem Wert) keine Zufallsvariable ist. Deshalb sagt ein Vertrauensintervall auf z. B. dem 95%-Niveau auch nicht, dass 95% der zukünftigen Messungen des Parameters in das beobachtete Intervall fallen (Morey et al. 2016). Was das Konfidenzniveau stattdessen angibt, ist die Rate, mit der bei zukünftigen Messungen die *dann* berechneten Vertrauensintervalle den unbekannten Wert des Parameters enthalten. Die Wurzeln dieses Missverständnisses sind also dieselben, die bereits die *p*-Wert Interpretation beim Null-Hypothesen Signifikanztest erschwert haben.

Es gibt jedoch eine naheliegende Bayessche Version der Intervallschätzer, die als *credential intervals* bezeichnet werden (ebenfalls gebräuchliche Bezeichnungen sind *credibility interval* oder *credible interval*). Im Rahmen der A-posteriori-Verteilung kann man Intervalle bilden, auf die eine gewisse Wahrscheinlichkeit entfällt (etwa in der Form eines *highest density interval*). Die oben erwähnten Interpretationen, die bei frequentistischen Vertrauensintervallen nicht zutreffen, sind bei Bayesschen *credential intervals* also gerade zulässig. Eine genauere Darstellung findet man bei Tschirk (2019, Kap. 7.2) und Jaynes (1976).

Wir schlagen zwar vor, die Bayessche Alternative stärker zu berücksichtigen, wollen aber nicht verhehlen, dass Fehlinterpretationen auch in diesem Rahmen auftreten (Hoijsink et al. 2016; Herrera-Bennett et al. 2020). Ebenfalls sei angemerkt, dass einige Vertreter der Bayesschen Schule durch ihre Verbissenheit der Sache nicht unbedingt gedient haben (siehe hierzu etwa Jaynes (1976)).

4 Zusammenfassung

Das Paradigma der „Evidenzbasierung“ in den Bildungswissenschaften verlangt empirische Wirksamkeitsnachweise für Programme und Interventionen. Studien im Kontrollgruppendesign (experimentell oder quasi-experimentell) stellen vor allem im Bereich der Unterrichts- und Lehr-Lern-Forschung die Methode der Wahl dar (Nagengast und Rose 2018, S. 675f). Der in diesen Designs gewonnene *p*-Wert eines Null-Hypothesen Signifikanztests ist jedoch ein wenig geeignetes und dazu mit fehlerhaften Interpretationen behaftetes Maß für diese „Evidenz“.

Problematische Eigenschaften des Null-Hypothesen Signifikanztests sind seit langem bekannt, haben jedoch durch die sog. „Replikationskrise“ eine erneute Aufmerksamkeit erfahren. Diese Krise hat in der gängigen Praxis der Datenanalyse sicherlich *eine* Ursache, aber diese Praxis muss in einem größeren Zusammenhang gesehen werden. Zahlreiche Reformvorschläge zielen deshalb auf die institutionelle Verfasstheit des gesamten Wissenschaftssystems (Munafò et al. 2017).

Einige Maßnahmen sind jedoch unmittelbar umsetzbar. Wir schließen uns gerne der Empfehlung an, den Begriff „statistisch signifikant“ nicht mehr zu verwenden (Wasserstein et al. 2019), da er zu einer unangemessenen Dichotomisierung der Forschungsergebnisse führt. Aus den Befunden leitet sich ebenfalls die Forderung ab, zumindest zentrale Resultate zum Gegenstand von Replikationsstudien zu machen. Wir glauben ebenfalls, dass der Bayessche Theorierahmen eine sinnvolle Ergänzung des Methodenrepertoires der quantitativen empirischen Bildungsforschung darstellt.

In erster Linie verstehen wir diese Arbeit jedoch als Beitrag zu einer hoffentlich breiten Diskussion zur weiteren Steigerung der Qualität in der Bildungsforschung.

Danksagung Unser besonderer Dank geht an Thomas Zügge (U Wuppertal), der mit dem Hinweis auf die Veröffentlichung Wasserstein et al. (2019) die gesamte Arbeit angestoßen hat. Wir danken ebenfalls Michael Rochnia, Johannes Grebe-Ellis (beide U Wuppertal), Andy Field (University of Sussex), Ron Wasserstein (American Statistical Association), Nicole Lazar (University of Georgia), Jörn-Steffen Pischke (London School of Economics), Christof Schuster (U Giessen), Thomas J. Faulkenberry (Tarleton State University), Christoph Leuenberger (Université de Fribourg), Andreas Müller (Université de Genève) und Gerd Gigerenzer (MPI für Bildungsforschung, Berlin) für hilfreiche Anmerkungen zu einzelnen Fragen dieser Arbeit.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- Ahrbeck, B., Ellinger, S., Hechler, O., Koch, K., & Schad, G. (2016). *Evidenzbasierte Pädagogik – Sonderpädagogische Einwände*. Stuttgart: Kohlhammer.
- Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, 102, 1–8.
- Baumert, J., & Tillmann, K.-J. (Hrsg.). (2016). *Empirische Bildungsforschung – Der kritische Blick und die Antwort auf die Kritiker*. Zeitschrift für Erziehungswissenschaften – Sonderheft, Bd. 31.
- Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: a proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72, 90–103.
- Bellmann, J. (2016). Datengetrieben und/oder evidenzbasiert? *Zeitschrift für Erziehungswissenschaften*, 19(Sonderheft 31), 147–161.
- Bellmann, J., & Müller, T. (2011). Evidenzbasierte Pädagogik – ein Déjà-vu? In J. Bellmann & T. Müller (Hrsg.), *Wissen, was wirkt – Kritik evidenzbasierter Pädagogik* (S. 9–32). Wiesbaden: VS.
- Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Bd. 4. Heidelberg: Springer.
- Bortz, J., Lienert, G. A., Barskova, T., Leitner, K., & Oesterreich, R. (2008). *Kurzgefasste Statistik für die klinische Forschung*. Bd. 3. Heidelberg: Springer.
- Buchhaas-Birkholz, D. (2008). Die ‚empirische Wende‘ in der Bildungspolitik und in der Bildungsforschung: Zum Paradigmenwechsel des BMBF im Bereich der Forschungsförderung. *Erziehungswissenschaft*, 20(39), 27–33.
- Carpenter, S. (2012). Psychology’s bold initiative. *Science*, 335, 1558–1560.

- Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48(3), 378–399.
- Caticha, A. (2008). *Lectures on probability, entropy, and statistical physics*. MaxEnt 2008, the 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Boracéia Beach, São Paulo, Brazil, July 8–13, 2008. <https://arxiv.org/abs/0808.0012>
- Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283–292.
- Coe, R. (1999). Manifesto for evidence-based education. <http://www.cem.org/attachments/ebe/manifesto-for-ebe.pdf>. Zugegriffen: 3. Mai 2019.
- Cohen, R. (1997). The earth is round ($p < .05$). In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Hrsg.), *What if there were no significance tests?* (S. 21–35). New York, London: Taylor and Francis.
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p -values. *Royal Society Open Science*, 1(3), 1–16.
- Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, 40(1), 1–5.
- Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: an annotated reading list. *Psychonomic Bulletin & Review*, 25(1), 219–234.
- Faulkenberry, T. J. (2018). Computing Bayes factors to measure evidence from experiments: an extension of the BIC approximation. *Biometrical Letters*, 55(1), 31–43.
- Fisher, R. A. (1990). Statistical methods for research workers. In R. A. Fisher (Hrsg.), *Statistical methods, experimental design and scientific inference*. Oxford: Oxford University Press.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3. Aufl.). Boca Raton: Taylor & Francis. Siehe <http://www.stat.columbia.edu/~gelman/book/> für die Homepage zu diesem Werk
- Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, 8(2), 195–204.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krueger, L. (1989). *The empire of chance: how probability changed science and everyday life*. Cambridge: Cambridge University Press.
- Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual – what you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Hrsg.), *The Sage handbook of quantitative methodology for the social sciences* (S. 391–408). Thousand Oaks: SAGE.
- Goodman, S. N. (1993). p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137(5), 485–496.
- Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130, 995–1004.
- Goodman, S. N. (2008). A dirty dozen: twelve P-value misconceptions. *Seminars in Hematology*, 45(3), 135–140.
- Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, 33(2), 175–183.
- Haller, H., & Krauss, S. (2001). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research*, 7(1), 1–20.
- Herrera-Bennett, A. C., Heene, M., Lakens, D., & Ufer, S. (2020). Improving statistical inferences: Can a MOOC reduce statistical misconceptions about p -values, confidence intervals, and Bayes factors? *PsyArXiv*. <https://doi.org/10.31234/osf.io/zt3g9>.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of P-hacking in science. *PLoS Biology*, 13(3), e1002106.
- Held, L., & Ott, M. (2018). On p -values and Bayes factors. *Annual Review of Statistics and Its Application*, 5(1), 393–419.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157–1164.
- Hofmann, H., van Kooten, P., & Hulscher, K. (2016). Why Bayesian psychologists should change the way they use the Bayes factor. *Multivariate Behavioral Research*, 51(1), 2–10.
- Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p 's) versus errors (α 's) in classical statistical testing. *The American Statistician*, 57, 171–182.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), 124.
- Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6), 645–654.
- Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Hrsg.), *Foundation of probability theory, statistical inference, and statistical theories of science*. Dordrecht: Reidel.
- Jaynes, E. T. (2003). *Probability theory. The logic of science*. New York: Cambridge University Press.

- Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.
- Klenke, A. (2006). *Wahrscheinlichkeitstheorie*. Berlin, Heidelberg: Springer.
- Kline, R.B. (2013). *Beyond significance testing – statistics reform in the behavioral sciences*. Baltimore: United Book Press.
- Krell, M., & Vierarm, A. (2016). Analyse schwierigkeitserzeugender Aufgabenmerkmale bei einem Multiple-Choice-Test zum Experimentieren. In M. Hammann & U. Gebhard (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik* (Bd. 7, S. 283–298). Innsbruck: Studienverlag.
- Krüger, D., Parchmann, I., & Schecker, H. (Hrsg.). (2014). *Methoden in der naturwissenschaftsdidaktischen Forschung*. Heidelberg: Springer.
- Lakens, D. (2019). The practical alternative to the P-value is the correctly used P-value. *PsyArXiv*. <https://doi.org/10.31234/osf.io/shm8v>.
- Lehmann, E.L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association*, 88(424), 1242–1249.
- Lindley, D.V. (1957). A statistical paradox. *Biometrika*, 44(1–2), 187–192.
- Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5(6), 161–171.
- Makel, M.C., & Plucker, J.A. (2014). Facts are more important than novelty: replication in the education sciences. *Educational Researcher*, 43(6), 304–316.
- Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23, 103–123.
- Munafo, M., Nosek, B., Bishop, D., Button, K., Chambers, C., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 21.
- Nagengast, B., & Rose, N. (2018). Quantitative Bildungsforschung und Assessments. In R. Tippelt & B. Schmidt-Hertha (Hrsg.), *Handbuch Bildungsforschung* (Bd. 4, S. 669–688). Wiesbaden: Springer.
- Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society Series A*, 231, 289–337.
- Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods*, 5(2), 241–301.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Pashler, H., & Harris, C.R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531–536.
- Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530.
- Raftery, A.E. (1995). Bayesian model selection in social research. In P.V. Marsden (Hrsg.), *Sociological methodology* (S. 111–196). Cambridge: Blackwell.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.
- Sachs, L. (2004). *Angewandte Statistik* (11. Aufl.). Heidelberg: Springer.
- Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research*, 1(4), 41–63.
- Simonsohn, U., Nelson, L.D., & Simmons, J.P. (2014). P-curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6), 666–681.
- Slavin, R.E. (2002). Evidence-based education policies: transforming educational practice and research. *Educational Researcher*, 31(7), 15–21.
- Trafimow, D., et al. (2018). Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology*, 9, 699.
- Tschirk, W. (2019). *Bayes-Statistik für Human und Sozialwissenschaften*. Berlin, Heidelberg: Springer.
- Tooley, J., & Darby, D. (1998). *Educational research: a critique*. London: Office for Standards in Education.
- Wasserstein, R.L., Schirm, A.L., & Lazar, N.A. (2019). Moving to a world beyond „ $p < 0.05$ “. *The American Statistician*, 73(sup1), 1–19.
- 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.