Bei dem vorliegenden Beitrag handelt es sich um eine von den Autoren erstellte – inhaltsgemäße, aber nicht wortwörtliche – Übersetzung des Originalbeitrags *„Interpretation of empirical results in intervention studies: A commentary and kick-off for discussion“* (https://doi.org/10.1007/s12662-023-00915-5)

# Interpretation von empirischen Ergebnissen in Interventionsstudien: Ein Kommentar und Anpfiff zu weiterführenden Diskussionen

Dirk Büsch[1] & Florian Loffing[2]

[1]Carl von Ossietzky Universität Oldenburg

[2]Deutsche Sporthochschule Köln

**Zusammenfassung**

Sportwissenschaft als eine empirische Wissenschaft produziert Studienergebnisse, die hypothesenorientiert zu interpretieren sind. Die Validität der Interpretation von statistisch und praktisch bedeutsamen Ergebnissen in Interventionsstudien ist zum einen von der theoretisch-inhaltlichen Fundierung der Fragestellung und zum anderen vom konkreten methodologischen Vorgehen abhängig. Unter Berücksichtigung der empirisch-inhaltlichen und statistischen Hypothesenebene ergeben sich dabei wiederkehrende Interpretationsschwierigkeiten, wenn Zahlen in Worte bzw. Handlungsempfehlungen überführt werden. Anhand zweier Beispiele soll eine Diskussion in der Scientific Community initiiert werden, die bei entsprechendem Interesse der Kolleg:innen zu methodologischen Aspekten in dieser Zeitschrift fortgeführt werden könnte.

## Vorbemerkungen

Dieser Kommentar verzichtet bewusst auf Zahlen und Gleichungen, da es hierbei ausschließlich um einen Vorschlag gehen soll, wie bestimmte Werte, die in der empirischen Forschung mittlerweile zum Standardrepertoire gehören (sollten), sinnvoll eingesetzt und interpretiert werden können. Für die detaillierte mathematisch-statistische Herleitung und Begründung verweisen wir auf die verwendete Literatur. Wir haben uns bewusst für einen Kommentar entschieden, um eine offene Diskussion in der empirischen Sportwissenschaft bzw. in dieser Zeitschrift anzustoßen und nicht um auf der Grundlage eines eingereichten Beitrags mit den Gutachter:innen über die Einreichungsplattform eine interne Diskussion über unterschiedliche Positionen zu führen, zumal wir inhaltlich nichts wirklich Neues thematisieren, sondern einen Aspekt herausgreifen wollen, mit dem wir und wahrscheinlich die meisten Kolleg:innen im Bereich von Lehre und Forschung immer wieder konfrontiert sind. Es handelt sich hierbei um die Interpretation von Ergebnissen in Interventionsstudien (de Vet et al., 2011), die in Anlehnung an die differenzierte Validitätsdiskussion von Westermann (2000) auch als Verbesserung der Interpretationsvalidität bezeichnet werden kann. Wir wollen dabei zwei Beispiele thematisieren, die auf unterschiedlichen Hypothesenebenen (Hussy & Möller, 1994) angesiedelt sind, aber gleichermaßen verdeutlichen, dass bei der Interpretation von Ergebnissen in Interventionsstudien nicht nur zwischen „trifft zu“ oder „trifft nicht zu“, sondern unter bestimmten Voraussetzungen im Forschungs- und „Übersetzungskontext“ auch differenzierter interpretiert werden kann und sollte.

## Zielstellungen

Für inhaltliche Abstufungen einer differenzierten Interpretation bieten sich aus unserer Perspektive exemplarisch zunächst zwei aktuelle Diskussionspunkte an: (1) Man versucht die Frage zu beantworten, um wie viel Mal wahrscheinlicher ist bspw. die Annahme der Alternativhypothese gegenüber der Nullhypothese, um die Bestätigung/Beibehaltung oder Ablehnung einer Forschungshypothese besser absichern zu können. (2) Man versucht die Frage zu beantworten, ob der in der Studie oder für die Gesamtheit abgeschätzte Effekt mit einer bestimmten Sicherheit größer als ein begründet zu erwartender Mindesteffekt ausfällt.

Im ersten Fall bietet der so genannte Bayes-Faktor eine Lösung an, der mit frei verfügbaren Programmen wie R-Project oder JASP für eine Reihe gängiger inferenzstatistischer Verfahren bestimmt werden kann (van Doorn et al., 2021), und im zweiten Fall bedarf es einer Erweiterung des Ansatzes für die Bestimmung von Effektgrößen um einen

zufallsbereinigten Mindesteffekt sowie entsprechende Vertrauensbereiche (Cumming, 2014; Herbert, 2019; Lakens, 2013).

## Interpretationshilfe auf der Ebene der Testhypothesen

In der frequentistischen Statistik besagt die Nullhypothese ($H_0$) zumeist, dass es keinen statistisch bedeutsamen Unterschied oder keinen statistisch bedeutsamen Zusammenhang gibt, wohingegen die Alternativhypothese ($H_1$) zumeist einen statistisch bedeutsamen Unterschied oder einen statistisch bedeutsamen Zusammenhang postuliert. Man prüft in diesem Fall die empirischen Daten gegen ein theoretisches Modell, welches immer von der Gültigkeit der Nullhypothese ausgeht (siehe u. a. Lakens, 2017; Lakens, Scheel, et al., 2018; für Ausführungen zu Äquivalenztests mit dem Ziel der Prüfung, ob ein Effekt unterhalb einer a priori bestimmten Schwelle für das interessierende Phänomen liegt). Entsprechend drückt der „berüchtigte *p*-Wert“ (auch als empirische Irrtumswahrscheinlichkeit bezeichnet) die bedingte Wahrscheinlichkeit für das Vorliegen empirischer Daten oder davon noch deutlicher abweichenden empirischen Daten (D) unter der Annahme der Gültigkeit der Nullhypothese aus, d. h. $p = P(D|H_0)$. Fällt der *p*-Wert klein aus und unterschreitet er die „magische“ Grenze des vor der statistischen Analyse festgelegten Signifikanzniveaus α (auch als kritische Irrtumswahrscheinlichkeit bezeichnet, Lakens, Adolfi, et al., 2018), so wird klassischerweise geschlussfolgert, dass die empirischen Daten vermutlich nicht mit der Annahme der Nullhypothese vereinbar sind. Entsprechend wird die Nullhypothese vorläufig abgelehnt und die Entscheidung zugunsten der Alternativhypothese getroffen. Im umgekehrten Fall, wenn der *p*-Wert über dem Signifikanzniveau liegt, wird die Nullhypothese beibehalten, jedoch nicht bestätigt und die Alternativhypothese zurückgewiesen (Hussy & Jain, 2002; siehe Otte et al., 2022, für eine Analyse der in Fachartikeln vorzufindenden Formulierungen für statistisch nicht signifikante Ergebnisse). Wir haben es hierbei mit der wahrscheinlich am weitesten verbreiteten, dichotomen Entscheidungsregel in der empirischen (Sport-)Wissenschaft zu tun, die seit ihrer Einführung allerdings auch durchgehend kritisch diskutiert und wiederholt fehlinterpretiert wurde und wird (Nickerson, 2000). Dabei „verbietet“ es die Regel unter anderem, dass *p*-Werte dahingehend miteinander verglichen werden dürfen, dass ein kleinerer *p*-Wert als Beleg für die Größe eines Effekts der Alternativhypothese oder vice versa interpretiert werden darf (Büsch & Strauß, 2016).

Im Gegensatz zum frequentistischen Ansatz werden in der Bayes-Statistik Alternativ- ($H_1$) und Nullhypothese ($H_0$) als „konkurrierende“ Modelle betrachtet, wobei angenommen wird, dass empirische Daten (D) vorliegen. Der resultierende Bayes-Faktor (BF) quantifiziert

die Stärke der Evidenz für $H_1$ relativ zu $H_0$, und je nachdem, ob die Forschungshypothese der $H_1$ oder der $H_0$ entspricht, kann auch der BF entweder zu Gunsten der $H_1$ oder der $H_0$ berechnet werden (Kass & Raftery, 1995). Ein wesentlicher Vorteil des BF besteht darin, dass durch den Vergleich der beiden Hypothesen ausgedrückt werden kann, um wie viel Mal wahrscheinlicher die Daten unter der einen Hypothese, z. B. der $H_1$, gegenüber den Daten unter der anderen Hypothese, z. B. der $H_0$ sind. Unter der Annahme, dass wir den BF zu Gunsten der $H_1$ berechnen wollen [$BF_{10} = P(D|H_1) / P(D|H_0)$], ergeben sich die Interpretationsoptionen in Tabelle 1. Der BF kann alternativ auch für die $H_0$ ($BF_{01} = 1/BF_{10}$) angegeben werden, womit im Gegensatz zum frequentistischen Ansatz die statistische Evidenz zu Gunsten der $H_0$ (relativ zur $H_1$) abgeschätzt und quantifiziert werden kann (Dienes, 2014). Der Bayes-Faktor kann bei wiederholenden Messungen auch als „Aktualisierungsfaktor" interpretiert werden, der uns anzeigt, was wir aus neu gewonnenen Daten „lernen" können. Pointiert kann bei wiederholenden Messungen die Frage beantwortet werden, inwiefern neue Daten unser Vertrauen (*belief*) in Bezug auf z. B. die Gültigkeit zweier konkurrierender Hypothesen $H_1$ und $H_0$ verändern? Wenn wir beispielsweise davon ausgehen, dass vor einer Erhebung neuer Daten die Wahrscheinlichkeit für die $H_1$ mit 60 % [$P(H_1) = 0.6$] sowie die Wahrscheinlichkeit für die $H_0$ mit 40 % [$P(H_0) = 0.4$] angenommen wird und ergibt sich basierend auf den neuen Daten ein $BF_{10} = 1$, dann würden wir aus den Daten nichts Neues (*keine Evidenz*) lernen. Bei einem $BF_{10} = 1$ ist nämlich die Wahrscheinlichkeit, dass die vorliegenden Daten (D) unter Gültigkeit der $H_1$ beobachtet werden konnten, $P(D|H_1)$, identisch zu der Wahrscheinlichkeit, dass die Daten unter Gültigkeit der $H_0$ beobachtet werden konnten, $P(D|H_0)$. Umgekehrt steigt unser Vertrauen in die Gültigkeit der $H_1$ gegenüber der $H_0$ im Vorher-Nachher-Vergleich, je größer der $BF_{10}$ ausfällt (Tabelle 1). Um das aktualisierte Vertrauen in Bezug auf die Gültigkeit der $H_1$ relativ zur $H_0$ (*posterior beliefs*), d. h. das Verhältnis der Wahrscheinlichkeiten zugunsten der $H_1$ und der $H_0$ unter Vorliegen neuer Daten bestimmen zu können, wird das Verhältnis der Wahrscheinlichkeitsannahmen vor den neuen Daten (*prior beliefs*) mit dem Bayes-Faktor $BF_{10}$ in Beziehung gesetzt (zur mathematischen Begründung siehe z. B. Wagenmakers et al., 2016). Ein hoher Wert für $BF_{10}$ „verschiebt" demnach das anfängliche in den *prior beliefs* quantifizierte Wahrscheinlichkeitsverhältnis von $H_1$ zu $H_0$ zugunsten der $H_1$ in den *posterior beliefs*. Angenommen, der zu dem oben genannten Studienbeispiel mit *prior beliefs* = 0.6/0.4 = 1.5 gehörige Bayes-Faktor sei $BF_{10} = 10$, dann ergibt sich daraus für die *posterior beliefs* =1.5 x 10 = 15. Während vor den neuen Daten also die $H_1$ lediglich als 1.5-mal wahrscheinlicher eingestuft wurde als die $H_0$, haben wir nun, unter Berücksichtigung der neuen Daten, ein deutlich stärkeres Vertrauen in die Gültigkeit der $H_1$ gegenüber der $H_0$, da uns die

$H_1$ 15-mal wahrscheinlicher als die $H_0$ erscheint. Umgekehrt können neue Daten auch dazu führen, dass unser Vertrauen in die Gültigkeit der $H_0$ gegenüber der $H_1$ steigt. Dies ist immer dann der Fall, wenn der Wert für $BF_{10}$ kleiner ist als 1 und umso „überzeugender", je näher sich der Wert für $BF_{10}$ dem Wert von 0 annähert (Tabelle 1).

*Tabelle 1. Evidenzkategorien für den Bayes-Faktor in Anlehnung an Jeffreys (1961), entnommen aus (Wetzels et al., 2011).*

| **$BF_{10}$** | **Interpretation** |
|---|---|
| > 100 | entscheidende (*decisive*) Evidenz zugunsten von $H_1$ |
| 30 – 100 | sehr starke (*very strong*) Evidenz zugunsten von $H_1$ |
| 10 – 30 | starke (*strong*) Evidenz zugunsten von $H_1$ |
| 3 – 10 | substantielle (*substantial*) Evidenz zugunsten von $H_1$ |
| 1 – 3 | anekdotische (*anecdotal*) Evidenz zugunsten von $H_1$ |
| 1 | keine Evidenz |
| 1/3 – 1 | anekdotische Evidenz zugunsten von $H_0$ |
| 1/10 – 1/3 | substantielle Evidenz zugunsten von $H_0$ |
| 1/30 – 1/10 | starke Evidenz zugunsten von $H_0$ |
| 1/100 – 1/30 | sehr starke Evidenz zugunsten von $H_0$ |
| < 1/100 | entscheidende Evidenz zugunsten von $H_0$ |

BF = Bayes Faktor

Trotz der gravierenden Unterschiede zwischen *p*-Wert einerseits und Bayes-Faktor andererseits ist dennoch zu berücksichtigen, dass der Bayes-Faktor mit kleiner werdendem *p*-Wert ansteigt, obwohl der *p*-Wert nur eine Aussage gegen die Nullhypothese zulässt (Wetzels et al., 2011). Ein auf dem frequentistischen Ansatz beruhender *p*-Wert kann, mit oder ohne Einbezug des Stichprobenumfangs und ggf. unter Berücksichtigung von Einschränkungen in der Höhe des *p*-Wertes, in eine geschätzte Obergrenze für den Bayes-Faktor $BF_{10}$ (sog. *bayes factor bound*; BFB) überführt werden (für detaillierte Ausführungen siehe u. a. Benjamin & Berger, 2019; Held & Ott, 2016, 2018; Sellke et al., 2001). Die Bayes-Faktor-Obergrenze repräsentiert den höchstmöglichen Bayes-Faktor $BF_{10}$, der mit dem beobachteten *p*-Wert vereinbar ist, und liefert somit, z. B. für Gutachter:innen eines eingereichten Beitrags mit einem frequentistischem Datenanalyse-Ansatz ein hilfreiches Maß zur Einschätzung der statistischen Evidenz eines Resultats zugunsten der $H_1$ relativ zur $H_0$. Die Bayes-Faktor-Obergrenze liegt für $p = .05$ bei BFB = 2.46, für $p = .01$ bei BFB = 7.99 und für $p = .001$ bei BFB = 53.26. Insbesondere ein knapp unterhalb des oftmals verwendeten Signifikanzniveaus $\alpha = 0.05$ liegender *p*-Wert ist also nur mit einer geringfügigen, anekdotischen Evidenz assoziiert (siehe

Tabelle 1), welche mit den Worten Jeffreys (1961) ausgedrückt, „nicht mehr als eine bloße Erwähnung wert“ ist (*worth no more than a bare mention*). Auch wenn *p*-Wert und Bayes-Faktor korrespondieren, so erlaubt der *p*-Wert u. a. keine direkte Abschätzung der uns in der Wissenschaft wie auch im Alltag z. B. bei der Einschätzung der Vertrauenswürdigkeit fremder Personen zumeist interessierenden Wahrscheinlichkeiten zweier konkurrierender Annahmen (Tschirk, 2019; Wasserstein & Lazar, 2016). Allerdings wäre aus der Perspektive des Bayes-Faktors auch für den frequentistischen Ansatz zu überdenken, ob bspw. statt der „üblichen“ kritischen Irrtumswahrscheinlichkeit von $\alpha = .05$ konsequenterweise $\alpha = .01$ oder kleiner (zur Diskussion siehe z. B. Benjamin et al., 2018; Lakens, Adolfi, et al., 2018) als „Standardkriterium“ angestrebt werden sollte, wenn als Zielgröße substanzielle Evidenzen sichergestellt werden sollen (Benjamin & Berger, 2019; Cohen, 1994; Wetzels et al., 2011). Als „Nebeneffekt“ ist dabei u. a. zu bedenken, dass eine Verschärfung der Anforderungen an die statistische Evidenz, d. h. eine „strengere Prüfung“ auch unmittelbare Konsequenzen, z. B. für den Umfang der Mindeststichprobe nach sich ziehen würde (siehe Brysbaert, 2019, für ein Tutorial, das die frequentistische und Bayessche Perspektive berücksichtigt).

### Interpretationshilfe auf der Ebene der empirisch-inhaltlichen Hypothesen

Während auf der Ebene der Testhypothesen bzw. der statistischen Vorhersage die Interpretation des *p*-Werts von entscheidender Bedeutung ist, bedarf es für die Interpretation einer empirisch-inhaltlichen Hypothese eine Fokussierung auf die Effektgröße. Dabei werden trotz aller berechtigter Kritik in der Scientific Community die Konventionen von Cohen (1988) für eine Interpretation weitgehend bevorzugt, obwohl der Autor selbst auf die Kontextabhängigkeit seiner Vorschläge hingewiesen hat (z. B. Durlak, 2009; Mesquida et al., 2022). Ein kleiner Interventionseffekt bei Leistungssportler:innen, bei denen marginale Unterschiede über Erfolg und Misserfolg entscheiden können, ist eben anders zu interpretieren als ein kleiner Interventionseffekt bei Trainingsanfänger:innen, bei denen bereits Plan- und Regelmäßigkeit körperlicher Aktivität zu kurzfristigen, teilweise exponentiellen Leistungsverbesserungen führen können (Rhea, 2004); die viel zitierte Ausgangswertregel lässt grüßen (Cordes, 1989).

Effektgrößen bezeichnen, wie das Kompositum nahelegt, die Größe eines Populations- oder Stichprobeneffekts, der entweder als Unterschieds- oder Abstandsmaß, z. B. die Effektgröße *d* für den an der gemeinsamen Streuung relativierten Mittelwertsunterschied, oder als Zusammenhangsmaß zwischen zwei Variablen, z. B. die Effektgröße *r*, dargestellt werden

kann (Cohen, 1988). Da Abstandsmaße in Zusammenhangsmaße und vice versa überführt werden können, beschränken wir uns im Folgenden auf die am häufigsten verwendete Effektgröße *d* zur Interpretation der Bedeutsamkeit eines Unterschieds zwischen zwei Gruppen resp. der Veränderung in einer Stichprobe.

Im Gegensatz zu den Konventionen für *d*, d. h. kleiner Effekt $d \geq 0.2$, mittlerer Effekt $d \geq 0.5$ und großer Effekt $d \geq 0.8$, bedingt eine kontextabhängige Interpretation des Effekts die vor Untersuchungsbeginn zu stellende Frage: „Wie groß sollte ein potenzieller Effekt ausfallen, damit die Intervention als lohnenswert interpretiert werden könnte?“ oder anders herum formuliert: „Wie groß könnte ein potenzieller Effekt ausfallen, ohne dass die Intervention als lohnenswert zu interpretieren wäre?“ Beide Fragen lassen sich vereinfachend dahingehend zusammenfassen: „Wie groß sollte ein Effekt *mindestens* ausfallen, damit die Intervention als lohnenswert interpretiert werden könnte?“. Eine entsprechende Mindesteffektgröße kann entweder theoretisch-inhaltlich oder „messmethodisch“ bestimmt werden. Im ersten Fall ließe sich die Mindesteffektgröße aus der Theorie unter Berücksichtigung der thematisch relevanten Studien ableiten, wohingegen im zweiten Fall eine messmethodische Effektgröße zu definieren wäre, die zwar größer als ein Nulleffekt ausfiele (Cumming, 2014), aber für eine inhaltliche Interpretation zu klein bzw. trivial wäre. Mögliche Gründe für kleine bzw. triviale und somit zu vernachlässigende Effekte können bspw. aus der Zuverlässigkeit des Messverfahrens, der Homogenität der Varianzen der Differenzen usw. resultieren. Die Grundannahme dieser Vorgehensweise erinnert an die Differenzierung bei Einzelfallanalysen zwischen der kleinsten bedeutsamen Veränderung (*Minimal Important Change*, MIC), die bspw. auf einem Konsens von Expert:innen beruhen kann, und der kleinsten nachweisbaren Veränderung (*Minimal Detectable Change*, MDC), die im Wesentlichen von der Genauigkeit des Messverfahrens bzw. dem Standardmessfehler (*Standard Error of Measurement*, SEM) abhängig ist (De Vet et al., 2006; King, 2011; Terwee et al., 2021). Für eine Abgrenzung von Gruppenstudien gegenüber Einzelfallanalysen kann die Mindesteffektgröße bei Interventionsstudien auch als kleinste lohnenswerte Veränderung (*Minimal Meaningful Change*, MMC) bezeichnet werden.

Der Ansatz von Mindesteffektgrößen geht u. a. auf Überlegungen von Murphy und Myors (1999) zur Prüfung von Minimum-Effekt-Hypothesen zurück. Die Autoren thematisieren, dass das Nullhypothesen-Testen (im Sinne von Nulleffekten) zwar einfach umzusetzen ist und sich trotz aller Kritik und Missinterpretationen in der empirischen Forschung weitgehend durchgesetzt hat, jedoch „wahre Nulleffekte“ die Realität nicht widerspiegeln. Bei Minimum-Effekt-Hypothesen wird demgegenüber danach gefragt, ob ein

Effekt „gut genug“ ist, um bspw. eine Intervention als lohnenswert oder lohnenswerter gegenüber anderen Interventionen bezeichnen zu können. Diese Annahme entspricht den Überlegungen zum zuvor ausgeführten Bayes-Faktor in Bezug auf die Frage, ob neue Daten unser Vertrauen in das Wahrscheinlichkeitsverhältnis zweier konkurrierender Hypothesen verändern (siehe auch Rouanet, 1996). Die Entscheidung über Minimumeffekte orientiert sich bei Murphy und Myors (1999; 2023) im Rahmen der *F*-Statistik (u. a. für Varianzanalysen) daran, wieviel Prozent erklärter Varianz als vernachlässigbar angesehen werden können bzw. als äquivalent zum Nulleffekt begründet anzunehmen sind. Unter der Annahme, dass bspw. bis zu einem Prozent erklärter Varianz in der *F*-Statistik vernachlässigt werden könnte, entspräche dies einer Effektgröße von $\eta^2 = .01$. Übertragen auf die *t*-Statistik entspräche diese Annahme einem *d*-Wert von ungefähr 0.20 (zur Umrechnung von Effektgrößen siehe z. B. https://www.psychometrica.de/effektstaerke.html). In dem Fall, dass die Reliabilität des Messverfahrens als sehr hoch einzuschätzen ist, z. B. bei einem $ICC \geq .95$ oder $ICC \geq .99$ kann der zu vernachlässigende Effekt auch geringer angesetzt werden. Beispielsweise entspräche eine erklärte Varianz von 0.5 % ($\eta^2 = .005$) einem $d \cong 0.14$ und von 0.1 % ($\eta^2 = .001$) einem $d \cong 0.06$. Letztendlich ist zu entscheiden, welcher Minimumeffekt vernachlässigbar erscheint bzw. vice versa, ab welchem „Grenzwert“ von einem interessierenden bzw. lohnenswerten Effekt ausgegangen werden kann (siehe auch Minimum Effect Tests, Jovanovic et al., 2022).

Die Mindesteffektgröße kann auch als „smallest effect size of interest“ (SESOI) bezeichnet werden und dokumentiert den Grenzwert bzw. „Referenzpunkt“, der von der Effektgröße (ES) in der untersuchten Stichprobe einschließlich des Vertrauensintervalls der Effektgröße (CI) nicht unterschritten werden sollte (Anvari & Lakens, 2021). Die SESOI ist begründet und transparent zu dokumentieren. Dabei dienen die Vertrauensintervalle zur Abschätzung, ob ein Interventionseffekt so groß ist, dass er sich mit einer bestimmten Wahrscheinlichkeit resp. Sicherheit „reproduzieren“ lässt und dabei immer größer als der Mindesteffekt ist (Herbert, 2019; Herbert, 2000; Kamper, 2019). Man kann sich damit etwas sicherer sein, dass man nicht auf das falsche Pferd bzw. die falsche Intervention gesetzt hat und mit hoher Wahrscheinlichkeit auch von einer praktischen Bedeutsamkeit des Effekts auszugehen ist. Die Herangehensweise bei diesem effektgrößenorientierten Vorgehen (*Magnitude Based Approach*) lässt sich wiederum mithilfe eines Baumdiagramms (*Tree Plot*) illustrieren (siehe Abbildung 1).

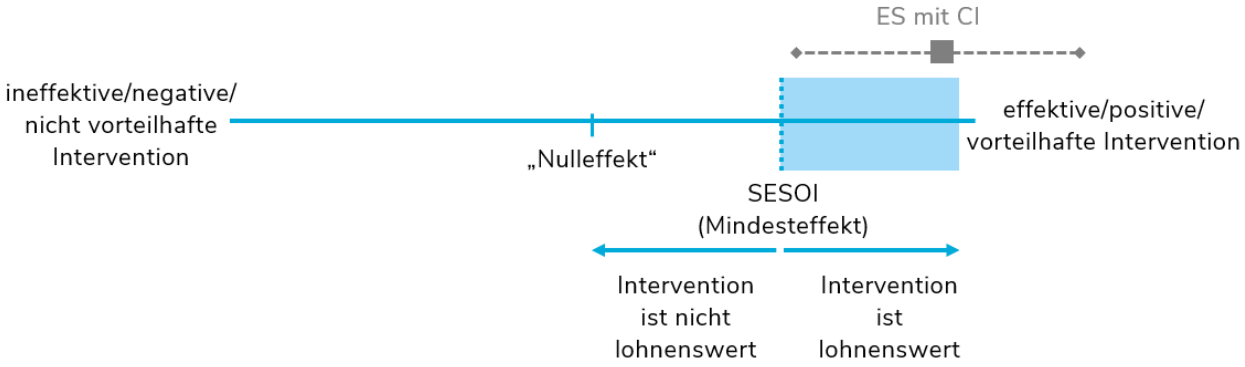


*Abbildung 1. Baumdiagramm für eine Effektgröße mit Vertrauensintervall und Mindesteffekt (in Anlehnung an Herbert, 2020). Anmerkung: ES: Effektgröße, CI: Vertrauensintervall, SESOI: Smallest Effect Size of Interest.*

Unter Berücksichtigung von Effektgröße und Vertrauensintervall sowie Mindesteffekt und „Nulleffekt“, d. h., dass es „keinen“ bzw. „keinen substanziellen“ Unterschied zwischen zwei Messzeitpunkten gibt, können die Interventionseffekte valide interpretiert werden. Liegen bspw. die Effektgröße und das Vertrauensintervall oberhalb des Mindesteffekts (SESOI, z. B. definiert über MMC), kann davon ausgegangen werden, dass die Intervention vorteilhaft bzw. praktisch bedeutsam ist. Liegen im Gegensatz dazu die Effektgröße und das Vertrauensintervall zwischen Nulleffekt und Mindesteffekt, so ist davon auszugehen, dass die Intervention zwar effektiv, aber nicht lohnenswert ist, wohingegen bei einer Effektgröße und einem Vertrauensintervall unterhalb des Nulleffekts eine Intervention als nicht vorteilhaft bzw. praktisch nicht bedeutsam, evtl. sogar als zunehmend schädlich einzuschätzen wäre. Die Validität der Interpretation ist jedoch immer dann eingeschränkt, wenn das Vertrauensintervall einen Referenzpunkt, d. h. entweder den Mindesteffekt oder den Nulleffekt einschließt (siehe Abbildung 2).

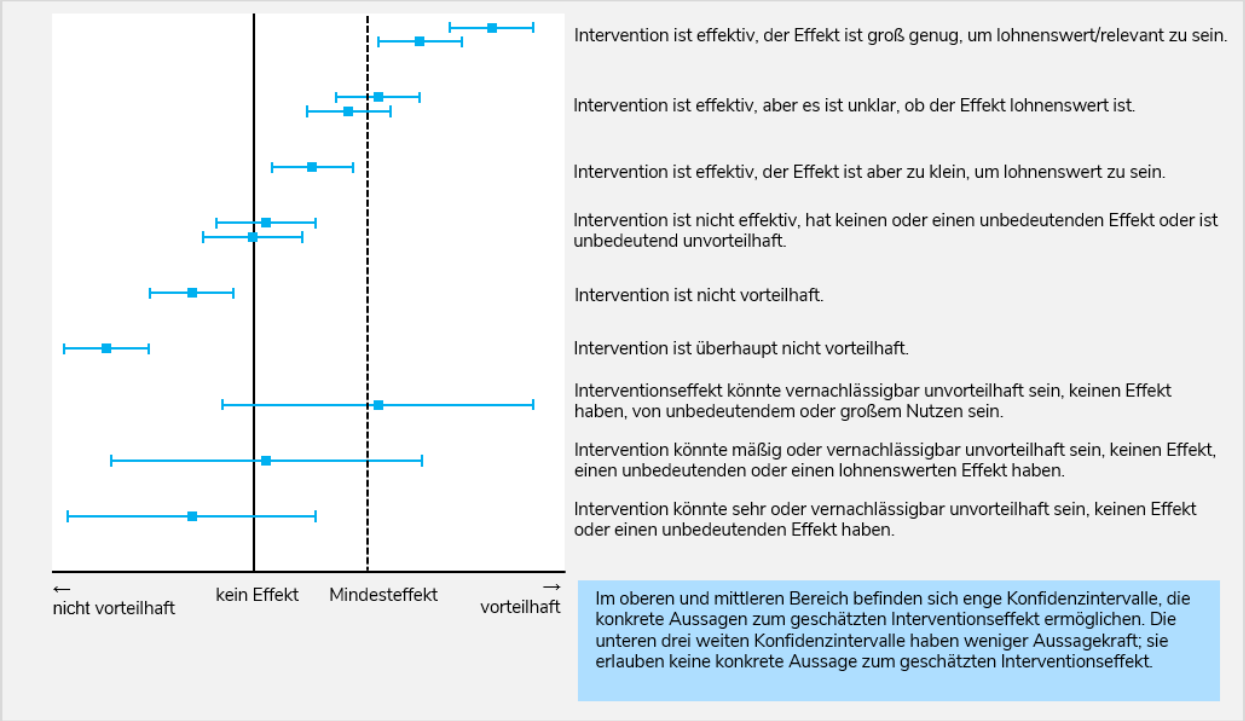


*Abbildung 2. Interpretation von Interventionseffekten unter Berücksichtigung einer Mindest-effektgröße sowie eines Nulleffekts (in Anlehnung an Kamper, 2019).*

## Abschließende Bemerkungen

Auch wenn in diesem Kommentar nicht darauf eingegangen wurde, gibt es selbstverständlich weitere Aspekte in der empirischen (Sport-)Forschung, die für eine hinreichende Interpretationsvalidität berücksichtigt werden müssen, z. B. eine theoretische Fundierung (Fiedler et al., 2021), eine Visualisierung der (Roh-)Daten (Loffing, 2022) usw. Die hier in den Blick genommenen Aspekte repräsentieren lediglich einen kleinen Ausschnitt eines insgesamt komplexen, oftmals auch in seinen Wechselwirkungen unterschätzten Forschungsprozesses, welcher an verschiedenen Stellen „Fallen“ bereit hält, in die jede:r von uns schon einmal getreten ist und leider Einschränkungen für die Aussagekraft von Studien beinhaltet. Den vorgestellten Ansätzen gemein ist das Streben nach einer Verbesserung der methodologischen Präzision, die auch für die sportwissenschaftliche Forschung einen substanziellen Mehrwert bietet, sodass dieser Kommentar als Anpfiff einer fortzusetzenden und sich ständig weiter entwickelnden Diskussion zu verstehen ist und nicht als Abpfiff einer vielleicht auch ermüdenden Diskussion missverstanden werden soll (Mesquida et al., 2022). Für den Anpfiff der zu initiierenden Diskussion formulieren wir abschließend zwei „provokante“ Vorschläge zur Interpretation von Studienergebnissen: (1) Berechne immer den Bayes-Faktor, um die Vertrauenswürdigkeit in deine interessierende Testhypothese (besser) abschätzen bzw. interpretieren zu können! (2) Bestimme im Vorhinein immer die

Mindesteffektgröße, um den substanziellen Nutzen deiner Intervention (besser) abschätzen bzw. interpretieren zu können!

## Literatur


Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Benjamin, D. J., & Berger, J. O. (2019). Three recommendations for improving the use of p-values. *The American Statistician*, *73*(sup1), 186-191. https://doi.org/10.1080/00031305.2018.1543135

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6-10. https://doi.org/10.1038/s41562-017-0189-z

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.72

Büsch, D., & Strauß, B. (2016). Wider die „Sternchenkunde"! *Sportwissenschaft*, *46*(2), 53-59. https://doi.org/10.1007/s12662-015-0376-x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997-1003.

Cordes, J. C. (1989). Die Ausgangswertregel. *Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin*, *41*(1), 47-58. https://doi.org/10.1055/s-2008-1065373

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29. https://doi.org/10.1177/0956797613504966

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge University Press. https://doi.org/10.1017/CBO9780511996214

De Vet, H. C. W., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes*, *4*(1), 54. https://doi.org/10.1186/1477-7525-4-54

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917-928. https://doi.org/10.1093/jpepsy/jsp004

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*(4), 816-826. https://doi.org/10.1177/1745691620970602

Held, L., & Ott, M. (2016). How the maximal evidence of p-values against point null hypotheses depends on sample size. *The American Statistician*, *70*(4), 335-341. https://doi.org/10.1080/00031305.2016.1209128

Held, L., & Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application*, *5*(1), 393-419. https://doi.org/10.1146/annurev-statistics-031017-100307

Herbert, R. (2019). Significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. *Journal of Physiotherapy*, *65*(3), 178-181. https://doi.org/10.1016/j.jphys.2019.05.001

Herbert, R. D. (2000). How to estimate treatment effects from reports of clinical trials. I: Continuous outcomes. *Australian Journal of Physiotherapy*, *46*(3), 229-235. https://doi.org/10.1016/S0004-9514(14)60334-2

Hussy, W., & Jain, A. (2002). *Experimentelle Hypothesenprüfung in der Psychologie*. Hogrefe.

Hussy, W., & Möller, H. (1994). Hypothesen. In T. Herrmann & W. Tack (Eds.), *Enzyklopädie der Psychologie: Themenbereich B Methodologie und Methoden, Serie I Forschungsmethoden der Psychologie. Band 1 Methodologische Grundlagen der Psychologie* (pp. 475-507). Hogrefe.

Jeffreys, H. (1961). *Theory of probability* (3. ed.). Oxford University Press.

Jovanovic, M., Torres Ronda, L., & French, D. N. (2022). Statistical modeling. In D. N. French & L. Torres Ronda (Eds.), *NSCA's essentials of sport science* (pp. 644-701). Human Kinetics.

Kamper, S. J. (2019). Confidence intervals: Linking evidence to practice. *Journal of Orthopaedic & Sports Physical Therapy*, *49*(10), 763-764. https://doi.org/10.2519/jospt.2019.0706

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795. https://doi.org/10.2307/2291091

King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(2), 171-184. https://doi.org/10.1586/erp.11.9

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863). https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168-171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269. https://doi.org/10.1177/2515245918770963

Loffing, F. (2022). Raw data visualization for common factorial designs using SPSS: A syntax collection and tutorial. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.808469

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports and exercise science: a narrative review of selected methodological issues in the field. *Royal Society Open Science*, *9*(12), 220946. https://doi.org/doi:10.1098/rsos.220946

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234-248. https://doi.org/10.1037/0021-9010.84.2.234

Murphy, K. R., & Myors, B. (2023). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (5 ed.). Routledge.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301. https://doi.org/10.1037/1082-989X.5.2.241

Otte, W. M., Vinkers, C. H., Habets, P. C., van Ijzendoorn, D. G. P., & Tijdink, J. K. (2022). Analysis of 567,758 randomized controlled trials published over 30 years reveals trends in phrases used to discuss

results that do not reach statistical significance. *PLoS Biology*, *20*(2), e3001562. https://doi.org/10.1371/journal.pbio.3001562

Rhea, M. R. (2004). Determining the magnitude of treatment effects in strength training research through the use of the effect size. *Journal of Strength and Conditioning Research*, *18*(4), 918-920. https://doi.org/10.1519/14403.1

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin*, *119*, 149-158. https://doi.org/10.1037/0033-2909.119.1.149

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62-71. https://doi.org/10.1198/000313001300339950

Terwee, C. B., Peipert, J. D., Chapman, R., Lai, J.-S., Terluin, B., Cella, D., Griffiths, P., & Mokkink, L. B. (2021). Minimal important change (MIC): A conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Quality of Life Research*, *30*(10), 2729-2754. https://doi.org/10.1007/s11136-021-02925-y

Tschirk, W. (2019). *Bayes-Statistik für Human- und Sozialwissenschaften*. Springer. https://doi.org/10.1007/978-3-662-56782-1

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813-826. https://doi.org/10.3758/s13423-020-01798-5

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169-176. https://doi.org/10.1177/0963721416643289

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

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

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291-298. https://doi.org/10.1177/1745691611406923