Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Dissertationsschrift zur Erlangung des

akademischen Grades Doctor rerum naturalium

(Dr. rer. nat.)

ANALYSE VON STATISTISCHEN QUALITÄTSINDIKATOREN PSYCHOLOGISCHER STUDIEN DER JAHRE 2010 – 2021 MIT DEM R-PAKET JATSdecoder

vorgelegt von:

Dipl.-Psych. Ingmar Böschen

Hamburg, 2023

Tag der Disputation: 8.11.2023

Promotionsprüfungsausschuss:

Vorsitzender: Prof. Ulf Liszkowski

1. Dissertationsgutachter: Prof. Martin Spieß
2. Dissertationsgutachter: Prof. Jan Wacker

1. Disputationsgutachterin: Prof. Tania Lincoln
2. Disputationsgutachterin: Prof. Juliane Degner

Inhaltsverzeichnis

1 Zusammenfassung

Einhergehend mit der Digitalisierung und immer größer werdenden Bildungs- und Forschungsetats, steigt die Anzahl jährlich publizierter Forschungsberichte kontinuierlich. Allein in dem Zeitraum, in dem diese Arbeit entstand (2019-2023), ist die Anzahl frei zugänglicher Open-Access Dokumente, die in der PubMed Central Datenbank verlinkt sind, von 2,8 auf 5 Mio. gestiegen. Die Identifikation von relevanten und hochwertigen Studien anhand von Metadaten (meist Titel, Zusammenfassung und Schlüsselwörter) wird mittlerweile durch die schiere Anzahl an Studien erschwert. Eine Option, Studien zusätzlich anhand von methodischen Eigenschaften zu identifizieren, ist daher von immer höherer Relevanz und ermöglicht ebenfalls die Beantwortung von Meta-Fragestellungen zum Wandel dieser Eigenschaften über die Zeit.

In den meisten empirischen Studien ist die Verwendung frequentistischer Verfahren ein allgemeiner Standard. Seit ihrer Entwicklung zu Beginn des letzten Jahrhunderts wird die praktische Anwendung dieser Verfahren jedoch kontrovers diskutiert und kritisiert. In der Literatur sind diverse problematische bzw. fehlerhafte Anwendungsszenarien und weit verbreitete, problematische Forschungspraktiken aufgeführt, die zu starken Einschränkungen der Validität der Ergebnisse und den damit verbundenen Schlussfolgerungen führen. Ein Indiz für die praktische Relevanz dieser Kritikpunkte in der Psychologie liefert die im Jahr 2015 publizierte Replikationsstudie der Open Science Collaboration. In den 100 wiederholten Experimenten, deren Ergebnisse in hochrangigen Zeitschriften verschiedener Disziplinen veröffentlicht wurden, konnte der veröffentlichte Effekt nur in einem geringen Anteil der Replikationen bestätigt werden (je nach Auslegung und Journal zwischen 23% und 62%).

In dieser Arbeit wird das neu entwickelte R-Paket *JATSdecoder* als hilfreiches Werkzeug der Wissenschaftsforschung vorgestellt (Publikation 1) und evaluiert. Das in *JATSdecoder* enthaltene Modul *get.stats()* wird hinsichtlich seiner Eigenschaft, im Text aufgeführte statistische Ergebnisangaben auszulesen, mit der bereits etablierten Software *statcheck* verglichen (Publikation 2). Die Präzision der Extraktion von ausgewählten methodischen Studienmerkmalen mit dem Modul *study.character()* wird durch einen Vergleich mit

einem händisch kodierten Datensatz der Eigenschaften realisiert (Publikation 3). In einem Anwendungsszenario werden die Veränderungen von ausgewählten, mit *JATSdecoder* extrahierten, methodischen Studieneigenschaften in $N = 57.909$ psychologischen Forschungspapieren aus 12 Zeitschriften, die innerhalb der letzten 12 Jahre erschienen sind, analysiert (Publikation 4).

JATSdecoder extrahiert verlässlich die Metadaten und Textelemente aus NISO-JATS kodierten Dokumenten und stellt ein allgemeines Werkzeug für die Wissenschaftsforschung dar. Bezüglich der Fähigkeit statistische Testergebnisse verlässlich und korrekt aus Texten zu extrahieren ist die Funktion *get.stats()* den Algorithmen von *statcheck* überlegen. Die entwickelten Algorithmen zur Extraktion methodischer Studienmerkmale stellen zusätzlich ein hilfreiches Instrument zur Identifikation von Studien mit spezifischen methodischen Eigenschaften dar und können eine händische Kodierung ersetzen.

Im globalen Vergleich von Studien, die zwischen 2010 und 2015 veröffentlicht wurden und Studien, die zwischen 2016 und 2021 erschienen, hat sich in den hier analysierten Forschungspapieren der Median der ermittelten Stichprobengrößen von 105 auf 190 erhöht. Der Median der Anzahl berichteter p-Werte je Studie sank von 14 auf 12. Der Anteil an Artikeln mit berichteten Korrekturverfahren für multiples Testen hat sich von 27% auf 23% verringert. Berichte von Konfidenzintervallen erhöhten sich von 21% auf 32%, Angaben zu Poweranalysen von 5% auf 11%. In 85% der Artikel wurden Ergebnisse von Nil-Nullhypothesentests mit p-Werten berichtet, von denen lediglich 2% mit einem α-Fehlerniveau unterhalb von $0,05$ durchgeführt wurden.

Die Forschungsartikel der 12 Zeitschriften, die hier mit *JATSdecoder* analysiert wurden, stellen eine große, wenngleich selektive Stichprobe aller psychologischen Forschungsberichte dar, weshalb eine Generalisierung der Ergebnisse auf die Psychologie als Ganzes, nur eingeschränkt möglich ist. Weiterhin ist zu beachten, dass die globalen Effekte von den vergleichsweise vielen Open-Access Artikel überlagert sind.

2 Einleitung

Als empirische Wissenschaft verfolgt die Psychologie das Ziel, das Erleben und das Verhalten des menschlichen Organismus zu beschreiben und allgemeine Gesetzmäßigkeiten abzuleiten. In der Praxis bedient sie sich dabei meist mathematischer, bzw. statistischer Konzepte und Formalismen, um objektive Schlussfolgerungen in Bezug zu den jeweiligen Forschungsfragestellungen zu begründen.

Das Publizieren von Studienergebnissen ist dabei ein elementarer Bestandteil der Arbeit von WissenschaftlerInnen und ein gewichtiges Bewertungsinstrument ihrer Arbeit. Durch hohe Förderetats und digitale Infrastrukturen beschleunigt, steigt die jährliche Anzahl an Veröffentlichungen von psychologischen Forschungsergebnissen stetig an. Neben dem klassischen Vertriebsmodell werden seit einigen Jahren immer mehr wissenschaftliche Beiträge auch über Open-Access-Modelle, sowie Hybridlösungen angeboten. Bei Open-Access Zeitschriften (z.B. *PlosOne* oder *Frontiers in Psychology*) fallen meist Verwaltungsgebühren für die AutorInnen an, die Leserschaft erhält freien Zugang zu den publizierten Inhalten und darf diese, je nach Lizensierung, weiterverarbeiten. Allein in der PubMed Central Datenbank (PMC) (PubMed-Central, 2020), einem der größten Repositorien für frei zugängliche Forschungsliteratur der Biologie und Gesundheitswissenschaften mit mehr als 5 Millionen verlinkten Dokumenten, stehen bereits mehr als 200.000 Artikel mit psychologischem Forschungskontext kostenfrei abrufbar zur Verfügung.

Innerhalb der Psychologie findet seit Jahren eine kritische Auseinandersetzung über den erarbeiteten Wissensstand und das methodische Vorgehen statt. Dabei stehen etablierte Standards der Konzeption von Studien, sowie der Aus- und Bewertung von Daten, zur Diskussion, die die Validität der zu treffenden Schlussfolgerungen mindern können. Zahlreiche WissenschaftlerInnen äußerten fundierte Kritik an den gegenwärtigen Anwendungspraktiken/-ritualen statistischer Methoden (Meehl, 1967; Berger & Sellke, 1987; Gigerenzer, Krauss & Vitouch, 2004; Gelman & Loken, 2014), Ergebnisinterpretationen (Haller & Krauss, 2002; Gelman & Stern, 2006) und Anreizsystemen (Rosenthal, 1979; Francis, 2012). Weiterhin fördern die äußeren

Rahmenbedingungen einer wissenschaftlichen Tätigkeit die Bereitstellung von neuem Wissen durch Veröffentlichungen, gleichzeitig erzeugen sie einen Veröffentlichungsdruck, der negative Auswirkungen auf die Qualität bzw. Replizierbarkeit von Forschungsergebnissen haben kann. Eine begründet pessimistische Perspektive nimmt Ioannidis (2005) ein und postuliert, dass die meisten veröffentlichten Forschungsergebnisse - und dies betrifft nicht nur die Psychologie - falsch sind.

Durch die massenhafte Bereitstellung frei zugänglicher Forschungsarbeiten und -ergebnisse über Open-Access Vertriebswege eröffnen sich neue Perspektiven für die Wissenschaftsforschung/Meta-Science. Ob mit dem stetigen Anstieg jährlich veröffentlichter Studienergebnisse der Psychologie auch eine Veränderung methodenspezifischer Qualitätskriterien einhergeht, wird in dieser Arbeit mit Hilfe von selbst entwickelten Textextraktions- und Textmanipulationsalgorithmen untersucht, die als frei verfügbares R-Paket *JATSdecoder* (Böschen, 2022) bereitgestellt werden. Um die allgemeine Einsatzfähigkeit und Präzision der in *JATSdecoder* enthaltenen Algorithmen zu bewerten, werden vor der praktischen Anwendung drei Evaluationsstudien durchgeführt.

3 Die Replikationskrise der Psychologie und mögliche Ursachen

In der von der Open Science Collaboration (2015, OSC) administrierten, groß angelegten Replikationsstudie psychologischer Studienergebnisse ($N = 100$ Studien), replizierten, je nach Teilgebiet und Auslegung von *erfolgreich repliziert*, zwischen 23% und 53% (Replikation ebenfalls signifikant), bzw. zwischen 34% und 62% (Effektgröße der Originalstudie im 95% Konfidenzintervall der Replikation), der von internationalen Autorengruppen wiederholten Studien. Unabhängig von der Definition, war der Anteil an erfolgreichen Replikationen bei den sozialpsychologischen Studien am niedrigsten, bei den kognitionspsychologischen Studien am höchsten. Durch diese praktische Überprüfung einiger psychologischer Forschungsergebnisse stellt das Ergebnis des OSC-Berichts eindrücklich die Glaubwürdigkeit der Psychologie als Wissenschaft infrage. Vor dem

Hintergrund, dass in den letzten Jahrzehnten zahlreiche Artikel erschienen sind, die sich kritisch mit der aktuellen Forschungspraxis der Psychologie auseinandersetzen, erscheint das Ergebnis der Replikationsstudie wenig überraschend.

Die in der Replikationsstudie der OSC wiederholten Experimente wurden alle in renommierten Zeitschriften mit Peer-Review-Verfahren veröffentlicht. Beim Peer-Review-Verfahren können von den EditorInnen ausgewählte ExpertInnen den Artikel vor Veröffentlichung bewerten, kommentieren und auch zurückweisen. Trotzt der mangelnden Reproduzierbarkeit und der methodischen Schwächen/Mängel vieler Studien stellt es einen wichtigen Eckpfeiler in der Qualitätssicherung wissenschaftlicher Artikel dar und kommt sowohl in Closed- als auch in Open-Access Zeitschriften zur Anwendung.

In den nächsten Abschnitten werden einige allgemeine, in der Literatur aufgeführte, Kritikpunkte zusammengefasst, die in der späteren Analyse als Qualitätsindikatoren dienen werden, um zu bewerten, ob sich die methodischen Eigenschaften psychologischer Studien innerhalb des Zeitraums 2010–2021 verändert haben. Die Kritikpunkte betreffen hauptsächlich die Realisierung von Stichproben, sowie die Anwendung des statistischen Nullhypothesentests als ungerichteten Nil-Nullhypothenstest, da dessen Ergebnisse im Allgemeinen die Grundlage für die inhaltlichen Schlussfolgerungen bilden. Anzumerken ist, dass diese Auflistung weitere wichtige Aspekte, wie beispielsweise die Messgenauigkeit der eingesetzten Instrumente, das Studiendesign oder die Adäquatheit der gewählten Modelle, außer Acht lässt.

3.1 Niedrige Power = kleine Stichproben = hohe Unsicherheit

Der als gering anzusehende Anteil an erfolgreichen Replikationen in der OSC-Studie macht deutlich, dass psychologische Forschungsergebnisse mit einer hohen Unsicherheit behaftet sein können. Eine allgemeine und einfache Möglichkeit, Unsicherheit in wissenschaftlichen Kontexten zu reduzieren, ist die Erhöhung, bzw. Festlegung der Genauigkeit des Schätzers $\hat{\theta}$. Weiterhin sind a-priori Überlegungen bezüglich der zu erwartenden Effektstärke für die Vermeidung von späteren Fehlschlüssen in Betracht zu ziehen. Die Teststärke oder auch Power (1-β Fehler) eines statistischen Tests bezeichnet die theoretische Wahrscheinlichkeit dafür, einen Effekt einer gewissen Größe $\delta_0 = \theta - \theta_0$, bei einem festgelegten α-Fehlerniveau (Wahrscheinlichkeit die H_0 fälschlicherweise zu verwerfen) und zugrunde gelegter Stichprobengröße, auch zu entdecken (ein signifikantes Testergebnis zu erhalten). Umgekehrt eignet sich dieser Ansatz ebenfalls zum Schätzen der benötigten Mindeststichprobengröße, um bei festgelegtem α-Niveau und gewünschter Power, Effekte (Abweichungen von der H_0) einer gewissen Mindestgröße δ_0 auch zu entdecken. Weiterhin ließe sich durch einen alleinigen Fokus auf die Messgenauigkeit die benötigte Mindeststichprobengröße schätzen, um einen zu schätzenden Parameter ausreichend genau zu schätzen (beispielsweise mit einem 95% Konfidenzintervall von ±3 Punkten).

Bereits 1962 attestierte Cohen einem Großteil der Studien des *Journals of Abnormal and Social Psychology* eine zu niedrige Teststärke (Cohen, 1962). In 50% der von Cohen untersuchten Studien lag die Wahrscheinlichkeit dafür, einen mittelgroßen Effekt zu entdecken, unter 46%. 24 Jahre später untersuchten Sedlmeier und Gigerenzer (1992) erneut die Power in psychologischen Studien und stellten fest, dass sich die Median-Power sogar weiter verringert hatte.

Dass niedrige Power ein fundamentales Problem darstellt, zeigen beispielsweise Gelman und Carlin (2014). Wenn WissenschaftlerInnen mit kleinen Stichproben und ungenauen Messinstrumenten kleine Effekte untersuchen - so wie es oft in der Psychologie geschieht - dann ist erstaunlicherweise recht häufig zu erwarten, dass ein signifikantes Ergebnis in die falsche Richtung zeigt (type S-error) und zusätzlich dazu neigt, den tatsächlichen Effekt in erheblichem Maße zu überschätzen (type M-error) (Gelman & Carlin, 2014).

Die Approximation der optimalen Stichprobengröße für einen zweiseitigen Mittelwertevergleich im Zwei-Stichproben t-Test mit der Funktion *power.t.test()* in R ergibt eine optimale Gruppengröße von $n_{opt} > 99.08$ Einheiten je Gruppe, um einen mittelgroßen Effekt von $\delta = .4$ Standardabweichungen, bei einem α-Niveau von 5% und einer Power von 80% zu identifizieren. Studien, die noch kleinere Effekte fokussieren, bzw. mit niedrigeren α-Niveaus arbeiten, bedürfen noch größerer Stichproben, um die Effekte bei unverändert hoher Power identifizieren zu können.

Blanca, Alarcón und Bono (2018) zeigen in ihrer Analyse von 287 aktuellen empirischen Forschungsarbeiten der Psychologie, dass heutzutage sehr vielfältige, fortschrittliche, statistische Techniken mit großer rechnerischer Komplexität angewendet werden. Komplexe Modelle benötigen jedoch meist große Stichproben, um zufriedenstellende Power aufzuweisen.

Ob mit der von Blanca et al. (2018) beobachteten Zunahme der Modellkomplexität ebenfalls ein Anstieg der zugrunde gelegten Stichprobengrößen zu beobachten ist und ob Power-Analysen häufiger zur Begründung der gezogenen Stichprobengröße herangezogen werden, soll hier näher untersucht werden.

3.2 Das Nil-Nullhypothesentestritual

Den meisten statistischen Methoden ist gemein, dass sie auf Annahmen über den Ziehungsprozess, die Variableneigenschaften und Modellresiduen basieren. Auf Grundlage dieser Annahmen wird die theoretische Verteilung des Schätzers $\hat{\theta}$ abgeleitet und dessen Streuung (Ungenauigkeit) $\hat{\sigma}_{\hat{\theta}}$ bestimmt, gegeben die H_0 gilt. Die erhaltenen Parameterschätzer $\hat{\theta}$ können dann mit Hilfe von Prüfgrößen und/oder Konfidenzintervallen auf überzufällige Abweichungen von dem in der H_0 festgelegten Wert θ_0 geprüft werden. In der Praxis erfolgt die Entscheidung des Hypothesentests zumeist über p-Werte, die mit einem zuvor festgelegtem α-Niveau verglichen werden. Zu betonen ist, dass die Schätzer gegen einen beliebigen Wert θ_0 getestet werden können. Ein in der Psychologie und den Sozialwissenschaften weit verbreiteter Irrtum über den Nullhypothesentest ist, dass die Nullhypothese mit einem Null-Effekt, Null-Zusammenhang, also $\theta_0 = 0$, gleichzusetzen ist (Meehl, 1967; Gigerenzer et al., 2004; Greenland, 2019). Diese Art Hypothese stellt einen Spezialfall dar und wird als Nil-Nullhypothese, Nil-Hypothese (Nickerson, 2000) oder auch Punkt-Nullhypothese bezeichnet (Berger & Sellke, 1987).

Im Idealfall folgt der Nullhypothesensignifikanztest folgendem Vorgehen:

- Hypothesen generieren: z.B. H_0: $\theta = .3$; H_1: $\theta \neq .3$ oder H_0: $\theta \geq .3$; H_1: $\theta < .3$
- Treffen von Annahmen über den datengenerierenden Prozess (z.B. unabhängige,

identisch verteilte Ziehungen) und das zugrunde gelegte Modell (z.B. linearer Zusammenhang mit normalverteilten Fehlern)

- Herleitung der Verteilung von $\hat{\theta}$ bei gegebenen Annahmen und Geltung der H_0
- α-Fehlerniveau festlegen, meist $\alpha = 0.05$
- β-Fehlerniveau/Power festlegen
- Approximation der optimalen Stichprobengröße n_{opt}
- Datenerhebung
- Bestimmung des empirischen Schätzwertes von $\hat{\theta}$
- Bestimmung des Standardfehlers/der Ungenauigkeit des Schätzers mittels $\hat{\sigma}_{\hat{\theta}}$
- Abgleich des Schätzers mit dem kritischen Wert θ_{krit}, Konstruktion eines (1-α)-Konfidenzintervalls, bzw. Bestimmung des p-Wertes (geschätzte Wahrscheinlichkeit für $\hat{\theta}$ oder extremere Ausprägungen, gegeben, die Nullhypothese gilt) und Abgleich mit α-Fehlerniveau

Fällt der Wert $\hat{\theta}$ in den Ablehnungsbereich der H_0, bzw. enthält das (1-α)-Konfidenzintervall, den in der H_0 festgelegten Wert θ_0 nicht, wird die Nullhypothese mit dem zuvor festgelegten α-Niveau verworfen. Fällt der Wert $\hat{\theta}$ nicht in den Ablehnungsbereich der H_0, bzw. enthält das (1-α)-Konfidenzintervall, den in der H_0 festgelegten Wert θ_0, wird die Nullhypothese bei einer Power von $1 - \beta$ nicht abgelehnt. Erst die Überlegungen zum β-Fehler ermöglichen es, die zur Identifikation der interessierenden Effekte nötige, bzw. optimale Stichprobengröße zu ermitteln. So können ebenfalls nicht signifikante Ergebnisse als Indiz (nicht als Beweis) für das Nichtvorhandensein eines Effekts/einer Abweichung von der H_0 gedeutet werden (bei Berücksichtigung der Power).

Da sich zwei oder mehr Gruppen von Einheiten niemals gänzlich gleichen und die meisten zur Anwendung kommenden statistischen Schätzgrößen $\hat{\theta}$ auf schwach konsistenten Schätzfunktionen basieren, was für die in der Psychologie üblichen Schätzer impliziert, dass ihre Streuung/Ungenauigkeit $\sigma_{\hat{\theta}}$ bei immer größer werdender Stichprobe gegen 0 konvergiert ($\lim\limits_{n \to \infty} \sigma_{\hat{\theta}} = 0$), weist Cohen darauf hin, dass sehr große Stichproben immer zur Ablehnung von präzisen Punkt-Nullhypothesen führen: *'The null hypothesis can only be* *true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false)*' (Cohen, 1990, p. 1308).

Der Tatsache, dass sehr große Stichproben meist zur Ablehnung von präzisen Punkt-Nullhypothesen führen, kann durch einen anderen Ansatz in der Stichprobenplanung begegnet werden. Mit Hilfe von Varianzschätzungen der zugrundeliegenden Variablen (z.B. aus vorangegangenen Studien, Normwerteangaben) wird die gewünschte, spätere (Un-)Genauigkeit des zu schätzenden Parameters θ festgelegt und die benötigte Mindeststichprobengröße bestimmt. Beispielsweise könnte der Anspruch einer Forschungsgruppe darin bestehen, den Effekt eines Phänomens mit einem 95% Konfidenzintervall auf ±2 Einheiten (Punkte, Millisekunden, o.Ä.) genau zu schätzen. Entsprechend fordert F. L. Schmidt (1996) Signifikanztestmethoden in Einzelstudien durch Punktschätzungen und Konfidenzintervalle zu ersetzen.

Meehl (1967) stellt die gegensätzlichen Anwendungsszenarien des Nullhypothesensignifikanztests, bzw. des Umgangs mit Unsicherheit, von PsychologInnen und PhysikerInnen gegenüber. Während PhysikerInnen eine konkrete Vorhersage über ihr Modell generieren und testen, ob die Messergebnisse signifikant von den Vorhersagen abweichen (was in der Regel unerwünscht ist), testen PsychologInnen in der Regel ihre Daten auf einen Null-Effekt und dies zusätzlich zumeist ungerichtet, und sehen in signifikant von Null abweichenden Ergebnissen eine Bestätigungen für ihr Modell (Meehl, 1967). Meehl führte also bereits vor über 50 Jahren an, dass nicht das bloße Vorhandensein einer Differenz zu θ_0 von theoretischem Interesse sei (d.h., dass die H_0 falsch ist, d.h., dass $\mu_g \neq \mu_b$), sondern das Vorhandensein einer Differenz in eine bestimmte Richtung (in Meehl's Fall: $\mu_g > \mu_b$). Da die allgemeine Praxis, mit angefallenen Stichproben ungerichtet Nil-Nullhypothesen zu testen, weiter allgegenwärtig war, empfahl Cohen (1990) erneut, bereits bei der Planung eines Experimentes, Gedanken über die Richtung und Stärke des zu entdeckenden/zugrunde liegenden Effekts mit einzubeziehen, um mithilfe von Power-Analysen die Größe der späteren Stichprobe zu begründen.

Spezifische Verfahren, die die Idee eines zu vernachlässigbar kleinen, bzw. grade so relevanten Effekts beinhalten, sind Äquivalenz- und Effekttests (Schuirmann, 1987;

Klemmert, 2004; Lakens, Scheel & Isager, 2018). Als Standardanwendung werden diese Verfahren bereits seit längerem in medizinischen Wirksamkeitsstudien genutzt, um die Nichtunterlegenheit bzw. Überlegenheit neuer Medikamente/Therapien zu prüfen (non-iferiority/superiority trails) (Schumi & Wittes, 2011). Als Ersatz für einen ungerichteten Nil-Nullhypothesentest empfehlen Schuirmann (1987) einen Äquivalenztest, der durch das zweimalige, einseitige Testen gegen einen redundanten, bzw. grade so relevanten Effekt erfolgt (two-one-sided testing). Diese Empfehlung wurde kürzlich für psychologische Forschungsarbeiten von Lakens et al. (2018) wiederholt.

Wie oft in psychologischen Studien mit gerichteten Tests gearbeitet wird, soll in dieser Arbeit näher analysiert werden.

3.3 α-Fehler-Kumulierung

Da mit zunehmender Anzahl statistischer Tests am gleichen Datenmaterial gleichzeitig die Wahrscheinlichkeit dafür steigt, mindestens ein falsch positives Testergebnis zu erhalten (α-Fehler Kumulierung), sollte das α-Fehlerniveau bei multiplen Tests angepasst/korrigiert werden. Ein gängiges, wenngleich regides und somit Power-reduzierendes Korrekturverfahren stellt die Korrektur nach Bonferroni (1935) dar, bei der das α-Niveau (bzw. die empirischen p-Werte) für alle Tests der Untersuchung über die Gesamtanzahl durchgeführter Tests (n) adjustiert wird ($\alpha_{adj} = \alpha/n$).

Vor allem bei kleinen Stichproben und einer großen Anzahl an Tests weist eine Adjustierung nach Bonferroni einen stark negativen Effekt auf die Power jedes einzelnen Tests auf und führt bei kleinen Stichproben meist dazu, dass tatsächlich vorhandene Effekte nach Korrektur des α-Niveaus nicht aufgedeckt werden können. Werden sehr viele statistische Tests am gleichen Datenmaterial durchgeführt, dann müsste die Stichprobengröße bei konstanten Effekten deutlich erhöht werden, um bei gleichbleibender Power die vorhandenen Effekte auch zu entdecken.

Es existieren weitere, teststärkere Verfahren zur Kompensation der α-Fehler-Kumulierung, wie die Methode von Šidák (1967), die sequentielle Korrektur nach Holm (1979), welche auch als Holm-Bonferroni Korrektur bekannt ist oder auch die

Methode von Benjamini und Hochberg (1995).

Dass, wie von Cohen (1994) aufgeführt, ungerichtete Punkt-Nullhypothesen zumeist falsch sind und gleichzeitig in der Psychologie ein inflationärer Gebrauch von Nil-Nullhypothesensignifikanztests vorherrscht, der eine Korrektur der α-Fehler erforderlich macht, erscheint zwar widersprüchlich, da der α-Fehler im ungerichteten Nil-Nullhypothesen-Setting praktisch nicht existiert, wird jedoch eindrücklich durch die Ergebnisse von Wicherts, Bakker und Molenaar (2011) untermauert. In ihrer Analyse von 49 psychologischen Forschungsarbeiten fanden sie 1.148 signifikant von Null abweichende t-, F- und χ^2-Testergebnisse. Dies entspricht, unabhängig von der zugrunde gelegten Stichprobengröße, einem Mittel von 23,4 signifikant von Null abweichenden Effekten pro Studie.

Angesichts der Omnipräsenz von signifikanten Ergebnissen und niedrigen Replikationsraten psychologischer Studien äußern einige Autoren ihre Zweifel an der Nützlichkeit eines festen α-Fehlerniveaus von 0,05 für die Entdeckung neuer Effekte und schlagen vor, es auf 0,005 bzw. 0,001 zu senken (Johnson, 2013; Ioannidis, 2018; Benjamin et al., 2018).

Untersuchungsgegenstände dieser Arbeit stellen die Anzahl extrahierter statistischer Testergebnisse je Studie, der Anteil signifikanter Ergebnisse, sowie die Anwendung von p-Wert/α-Fehler Korrekturen dar, sowie deren Entwicklung über die Zeit.

3.4 Das ‘file drawer problem’

Ein weiterer Aspekt, der in der Psychologie durch das ritualisierte Testen von Nil-Nullhypothesen und das gleichzeitige Bevorzugen von signifikanten Ergebnissen hervorgerufen wird, ist das sogenannte ‘file drawer problem’. Da HerausgeberInnen von wissenschaftlichen Forschungsartikeln eher an sensationellen, neuen Ergebnissen interessiert sind, werden Studien, die keine signifikanten Ergebnisse in Bezug zu der untersuchten Fragestellung enthalten, selten, bis gar nicht publiziert (Rosenthal, 1979; Francis, 2012), verschwinden also in der ‘Schublade‘. Zusätzlich werden Ergebnisberichte von erfolgreichen, wie nicht erfolgreichen Replikationen selten veröffentlicht (Francis, 2012). Unter dem Aspekt eines höheren Nachrichtenwerts und im Fokus stehender Absatzzahlen ist dieses Verhalten der Verlage durchaus verständlich, dient jedoch nicht einer kumulativen Wissenschaft, die auch ‘Misserfolge‘ in ihre Interpretationen und Ursachenzuschreibungen miteinbeziehen muss und kann zu Verzerrungen in Metaanalysen führen.

Obwohl sie auch andere Erklärungsansätze liefern (Power-Analyse, optional stopping), erkennen Kühberger, Fritz und Scherndl (2014) in ihrer Analyse von 1.000 psychologischen Forschungsartikeln ein Indiz für einen solchen Publikationsfehler. Sie ermittelten eine negativen Korrelation zwischen aufgeführten Effekt- und Stichprobengrößen ($r = -.45$ $[95\% KI : -.53; -.35]$). Je höher die Stichprobengröße der Studie ausfiel, desto kleiner waren die berichteten Effekte.

Die automatische Extraktion der berichteten Effektgrößen und Freiheitsgrade mit der neu entwickelten Funktion *get.stats()* kann die ‘file drawer’ Problematik für einzelne Forschungsgebiete leicht identifizierbar machen.

3.5 Der Einsatz von fragwürdigen Forschungspraktiken

Es existieren viele fragwürdige und zugleich kostengünstige Forschungspraktiken, Datenreihen dahingehend so lange zu verändern/zu ergänzen, bis die gewünschten Ergebnisse erzielt wurden (Simmons, Nelson & Simonsohn, 2011; Head, Holman, Lanfear, Kahn & Jennions, 2015). Beispielsweise ist es einfach möglich post-hoc solange Ein- bzw. Ausschlusskriterien zu definieren, bis die Ergebnisse den eigenen Wünschen entsprechen. Ebenfalls können viele ungeplante Subgruppenvergleiche durchgeführt, bzw. sich stark ähnelnde Variablen ausgetauscht werden und letztendlich nur die gewünschten Ergebnisse berichtet werden.

Ein einfaches und schwer zu identifizierendes Verfahren um signifikante Abweichungen von einer H_0 zu produzieren ist unter *optional stopping* bekannt. Nach dem Erhalt eines nicht signifikanten Ergebnisses werden solange weitere Einheiten gezogen, bis das Testergebnis signifikant geworden ist. Ein weiteres, fragwürdiges Vorgehens ist das HARKing. Unter HARKing versteht man die Darstellung einer Post-Hoc-Hypothese (d.h.

einer Hypothese, die auf den eigenen Ergebnissen basiert oder von diesen abhängig ist), als wäre sie eigentlich eine a-priori-Hypothese (Kerr, 1998).

Simmons et al. (2011) kommen nach ihrer Auflistung potenzieller Möglichkeiten, Ergebnisse in die gewünschte Richtung zu manipulieren, zu dem Schluss, dass es in vielen Fällen wahrscheinlicher ist, dass ein/e ForscherIn fälschlicherweise Beweise dafür findet, dass ein Effekt vorliegt, als dass er/sie Beweise dafür findet, dass ein Effekt nicht vorliegt.

Dass das Löschen von Ausreißern/Extremwerten bei statistischen Tests das α-Fehlerniveau drastisch erhöhen kann, zeigen Bakker und Wicherts (2014). Da das Löschen der extremsten Werte immer zu einer Verringerung der gemessenen Varianz der Wertereihe führt und somit zu kleineren Standardfehlern, ist dieser Effekt stets zu berücksichtigen und zwischen plausiblen und durch Messfehler bedingten Extremwerten zu unterscheiden. Handelt es sich um plausible Werte, können die extremsten Ausprägungen, bzw. Abweichungen eines Modells, wichtige Informationen in Bezug auf Einschränkungen, Nebenbedingungen, oder nicht lineare Trends enthalten.

John, Loewenstein und Prelec (2012) befragten mehr als 2.000 PsychologInnen danach, ob sie bereits fragwürdige Forschungspraktiken angewendet haben und ein erstaunlich hoher Anteil gab bereitwillig zu, solche Praktiken bereits verwendet zu haben, ohne dies sonderlich kritisch zu sehen.

Um die Anwendung von fragwürdigen Forschungspraktiken zu begrenzen, fordern in letzter Zeit immer mehr WissenschaftlerInnen und Verlage eine Präregistrierung von Untersuchungen. So kann zumindest klar zwischen Hypothesen prüfenden und explorativen, Hypothesen generierenden Untersuchungen unterschieden werden und weiterhin sichergestellt werden, dass lediglich die zuvor geplanten Analysen durchgeführt und berichtet werden (Nosek, Ebersole, DeHaven & Mellor, 2018).

3.6 Falsches Interpretieren von Ergebnissen

Neben der Anwendung von fragwürdigen Forschungspraktiken und der Omnipräsenz von Nil-Nullhypothesentests werden Ergebnisse teilweise auch falsch bewertet. Haller und Krauss (2002) geben einige Beispiele für falsche Interpretationsweisen von p-Werten und zeigen auf, wie weit verbreitet falsche Konzepte von, sowie Vorstellungen über diese, wohl am häufigsten berichtete, statistische Kennzahl sind. Weiter wurde gezeigt, dass diese falschen Konzepte auch bei WissenschaftlerInnen der Psychologie und sogar StatistikdozentInnen an psychologischen Fachbereichen weit verbreitet sind.

Informell ist ein p-Wert die Wahrscheinlichkeit dafür, unter einem spezifizierten statistischen Modell eine statistische Zusammenfassung der Daten zu erhalten (z. B. die mittlere Stichprobendifferenz zwischen zwei Gruppen), die gleich dem, oder noch extremer, als der beobachtete Wert ist (Wasserstein, Lazar et al., 2016). Leider werden p-Werte oft fälschlicherweise als eine unbedingte Fehlerwahrscheinlichkeit für das Ablehnen der zu testenden Hypothese oder, noch schlimmer, als die posteriori Wahrscheinlichkeit, dass die Nullhypothese wahr ist, angesehen (Sellke, Bayarri & Berger, 2001). Weiterhin besteht ein häufig begangener statistischer Fehler darin, einen signifikanten Unterschied zwischen nichtsignifikanten und signifikanten Ergebnissen zu postulieren (Gelman & Stern, 2006). Ein allgemeines Beispiel für diese Fehlinterpretation wäre: x ist ein signifikanter Prädiktor für y ($p = 0{,}04$), während z nicht signifikant mit y zusammen hängt ($p = 0{,}06$), also unterscheiden sich x und z in ihrem Effekt auf y.

Der in der Psychologie vorherrschende Wunsch, signifikante Ergebnisse zu berichten, führt teilweise zu nicht angebrachten sprachlichen Aufwertungen nicht signifikanter Ergebnisse. Dass nicht signifikante Ergebnisse dramatisch oft und sogar immer öfter als trend-signifikant deklariert werden, zeigen Pritschet, Powell und Horne (2016) in ihrer Analyse von 1.469 Forschungspapieren aus vier Dekaden (1970-2010). Trotz eines α-Niveaus von 0,05, wird in 459 der analysierten Artikel (31,2%) mindestens ein nicht signifikanter p-Wert, der zwischen 0,05 und 0,2 lag, als trend-signifikant beschrieben. Im Jahr 2010 war der Anteil mit 54% nur geringfügig niedriger, als im Jahr 2000 mit einem Maximalanteil von 59% der Artikel. Dieses Vorgehen ist angesichts der bereits nachgewiesenen hohen falsch-positiv Raten als problematisch zu bewerten.

3.7 Fehlerhaftes Berichten von Ergebnissen

Auch WissenschaftlerInnen und ReviewerInnen sind nicht unfehlbar, weshalb es durchaus passieren kann, dass die berichteten Forschungsergebnisse Fehler enthalten. Ein einfacher Test, um die Plausibilität von Mittelwerten bei Likert skalierten Tests zu überprüfen, ist der GRIM-Test (Brown & Heathers, 2017). Da die Summe Likert skalierter Items jeweils einer ganzen Zahl entspricht, kann die Plausibilität des angegebenen Mittelwerts durch Multiplikation mit der Stichprobengröße geprüft werden. Das Produkt aus Mittelwert und Stichprobengröße wird dann ab- und aufgerundet und jeweils durch die Stichprobengröße geteilt. Weichen die so erhaltenen Werte jeweils vom angegebenen Mittelwert ab, liegt eine Inkonsistenz vor. Es könnte falsch gerundet oder mit einer nicht angegebenen Substichprobe gerechnet worden sein (z.B. Ausschluss von Fällen, Missings). In 36 von 71 (50,7%) untersuchten Artikeln identifizierten Brown und Heathers (2017) mindestens eine inkonsistente Mittelwertangabe mit dem GRIM-Test. 16 dieser 36 Artikel (insgesamt 22,5%) enthielten mehr als eine Inkonsistenz.

Mit Hilfe des R Pakets *statcheck* (Epskamp & Nuijten, 2018) lassen sich einige, im APA Format berichtete, statistische Testergebnisse (Z-, t-, F-, r-, χ^2-Statistik mit dazugehörigen Freiheitsgradangaben und p-Wert) automatisiert auf Plausibilität/ Korrektheit prüfen. Nuijten, Hartgerink, van Assen, Epskamp und Wicherts (2016) geben an, dass *statcheck* in 48,8% der 30.717 analysierten Artikel Inkonsistenzen und in 11,6% der Artikel grobe Inkonsistenzen (definiert als: "Fehler führt zu einer Veränderung der statistischen Entscheidung“) identifizierte.

T. Schmidt (2017) erhebt berechtigte Kritik an der aktuellen Version von *statcheck*, da die Software vor allem bei Ergebnissen mit korrigierten p-Werten dazu tendiert, korrekte Angaben als falsch und gleichzeitig Ergebnisse, die eigentlich einer p-Wert-Korrektur bedürften, als korrekt zu markieren. Ein weiterer Kritikpunkt ist die relativ hohe Rate an nicht erkannten statistischen Ergebnissen ($\approx 40\%$). Bei eigenen Testdurchläufen zeigte sich, dass die von Nuijten et al. (2016) und T. Schmidt (2017) analysierten Artikel, die zunächst aus dem PDF Format in reinen Text umgewandelt wurden, uneinheitlich kodierte bzw. nicht extrahierbare Sonderzeichen enthielten, weshalb *statcheck* einen großen Teil der in den Studien enthaltenen Ergebnisse weder als solche erkennen noch prüfen konnte. Da eine verlässliche Extraktion von statistischen Ergebnissen für weitergehende Analysen und Auswahlmöglichkeiten unerlässlich ist und die *statcheck* Algorithmen erhebliche Mängel aufweisen, wurden für diese Arbeit eigene Extraktionsroutinen entwickelt. Vor der Systematisierung der im Text enthaltenen Ergebnisse in einer leicht weiterzuverarbeitenden Matrix, werden feinabgestimmte Textaufbereitungen und eine Vereinheitlichung der Ergebnisangaben realisiert. Die neu entwickelten Algorithmen können sowohl ForscherInnen als auch HerausgeberInnen von wissenschaftlichen Pubikationen im Qualitätsmanagement unterstützen, indem sie fehlerhafte Angaben/Übertragungen von Ergebnissen zu vermeiden helfen.

3.8 Selektive Stichproben

Zuletzt sei auf einen eher schwer zu korrigierenden Umstand hingewiesen. In Bezug zu der Grundgesamtheit, auf die die späteren Ergebnisse einer Studie übertragen werden sollen, sind die meisten der in der Psychologie gezogenen Stichproben selektiv (Arnett, 2008; Henrich, Heine & Norenzayan, 2010). Henrich et al. (2010) beziffern allein den Anteil von US-amerikanischen Stichproben, die in psychologischen Forschungsartikeln zwischen den Jahren 2003 – 2007 enthalten sind, auf 68%. Die im gleichen Zeitraum vertretenen AutorInnen stammten ebenfalls zum größten Teil aus den USA (73%), bzw. der westlichen Welt (99%). Weiterhin beziffern Henrich et al. (2010) den Anteil der in der Psychologie gezogenen Stichproben, die aus '*Western Educated Industrialized Rich and Democratic societies*' (WEIRD) stammen, auf 96%, obwohl diese Personengruppe lediglich 12% der Weltbevölkerung ausmacht. Zusätzlich werden in der Psychologie sehr viele Studierendenstichproben gezogen, wohl da diese meist sehr kostengünstig zu akquirieren sind. So wurden beispielsweise im Jahr 2007 67% der aus den USA stammenden und 80% der aus anderen Ländern stammenden Untersuchungen, die im *Journal of Personality and Social Psychology* erschienen, an Studierenden durchgeführt (Arnett, 2008).

Es erscheint sehr plausibel, dass sich die soziale, wirtschaftliche und auch politische Umgebung, in der Menschen leben und in der psychologische Wissenschaft praktiziert wird, auf einzelne, wichtige psychologische Komponenten wie Motivation, Moral und Verhalten, sowie auf mehrdimensionale, psychologische Wirkzusammenhänge auswirken kann. Die von Henrich et al. (2010) aufgeführten Beispiele dafür, dass sich WEIRD-people von anderen Sozietäten stark unterscheiden können (z.B. räumliche Wahrnehmung, soziale und moralische Entscheidungen), liefern jeweils tiefere Einblicke in den untersuchten Sachverhalt und sollten ForscherInnen darin bestärken, einen größeren finanziellen Aufwand bei der Stichprobenakquise einzuplanen. Besteht der Anspruch, eine allgemeingültige Aussage über *die Menschen* zu machen, dann ist eine stark selektive Stichprobe zumeist wenig repräsentativ und somit nicht als Grundlage wissenschaftlicher Entscheidungsprozesse geeignet. Zudem sollten WissenschaftlerInnen stets nach Nebenbedingungen bzw. Subpopulationen Ausschau halten, die die zugrunde gelegte Theorie eingrenzen und weiter präzisieren können.

4 Forschungshypothesen

Aus den in Abschnitt 3 aufgeführten Kritikpunkten abgeleitet, folgen im nächsten Abschnitt die Forschungshypothesen dieser Arbeit. Zunächst wird die Güte der neu entwickelten Extraktionsalgorithmen bewertet, die in den R-Funktionen *JATSdecoder()* und *study.character()* zusammengefasst sind. Daraufhin fokussiert die inhaltliche Analyse auf die Veränderung der in der Psychologie zur Erkenntnisgewinnung zugrunde gelegten Stichprobengrößen, der zur Anwendung kommenden statistischen Test- und Korrekturverfahren, sowie der Gesamtanzahl an statistischen Testergebnissen je Artikel innerhalb des Zeitraums 2010 – 2021.

4.1 Hypothesen bzgl. der Evaluation der Extraktionsalgorithmen

Um die allgemeine Einsatzfähigkeit von *JATSdecoder* zu untersuchen, wird die gesamte PubMed Central Datenbank mit der Funktion *JATSdecoder()* verarbeitet und die Verwendung der extrahierten Meta-Tags sowie deren Inhalt analysiert. Um die Präzision und Validität der mit *study.character()* extrahierten Studieneigenschaften kritisch zu hinterfragen, werden manuell extrahierte Studieneigenschaften anderer AutorInnen herangezogen. Der Vergleich mit der manuell extrahierten Anzahl statistisch signifikanter Ergebnisse in 49 Artikeln von Bakker und Wicherts (2011) dient dabei als erster als Indikator für die Präzision des Auslesens von statistischen Ergebnissen mit der in *JATSdecoder* enthaltenen Funktion *get.stats()*. Weiterhin werden die 49 Artikel (im PDF und HTML Format), die ebenfalls der Evaluationsstudie von *statcheck* (Nuijten, van Assen, Hartgerink, Epskamp & Wicherts, 2017) zugrunde lagen, für einen direktern Vergleich der Extraktionen von *get.stats()* und den Funktionen des R-Pakets *statcheck* (Epskamp & Nuijten, 2018) genutzt.

Die Originaldaten der manuell extrahierten Anzahl an Studien je Artikel, sowie der zur Anwendung gekommenen statistischen Softwarelösungen und Verfahren in den 288 von Blanca et al. (2018) analysierten Studien dienen der Bewertung der Extraktionsgenauigkeit dieser Studieneigenschaften. Für weitere Analysen wird der

Datensatz von Blanca et al. (2018) um manuell extrahierte Studieneigenschaften ergänzt (α-Fehler, Power, Korrekturverfahren für multiples Testen), um die Präzision der *JATSdecoder* Algorithmen zu bewerten und Einschränkungen zu diskutieren.

H_1: **_JATSdecoder()_ konvertiert NISO-JATS kodierte XML Dateien, sowie durch _CERMINE_ konvertierte PDF Dateien in eine einheitliche R-Liste mit Segmentierung der einzelnen Metadaten und Textteile.**

Alle frei zur Verfügung stehenden Dokumente der PMC-Datenbank werden mit *JATSdecoder()* verarbeitet und hinsichtlich der Verwendung und Inhalte der extrahierten Meta-Tags analysiert.

H_2: **Die Anzahl ausgelesener statistischer Ergebnisse ist merklich höher, als die des R-Pakets _statcheck_.**

Es wird angestrebt, mindestens 90% der im Text des Methoden- und Ergebnisabschnitts enthaltenen statistischen Testergebnisse zu identifizieren und auszulesen (Tabellen und Grafiken ausgeschlossen). Im Vergleich zu den in *statcheck* enthaltenen Algorithmen soll die Präzision der Extraktion mit der neu entwickelten Funktion *get.stats()* unabhängig vom Inputformat sein.

H_3 : **Die Extraktion der Anzahl Studien, verwendeter statistischer Methoden und genutzter Software ist mit einer händischen Analyse vergleichbar.**

Die händisch kodierten Rohdaten von Blanca et al. (2018) und die manuell ergänzen Studieneigenschaften werden mit den Ergebnissen von *study.character()* für 288 Studien im PDF Format direkt gegenübergestellt und bewertet.

4.2 Hypothesen bzgl. der Veränderung methodisch relevanter Qualitätskriterien wissenschaftlicher Publikationen in der Psychologie

Durch das steigende Bewusstsein für einen standardisierten Umgang mit wissenschaftlichen Forschungsprozessen, die immer leichtere und freie Verfügbarkeit statistischer Methoden, sowie Möglichkeiten große Datenmengen zu generieren und zu verwalten wird erwartet, dass sich wichtige methodische Aspekte wissenschaftlicher

Berichte der Psychologie in den Jahren 2010–2021 überwiegend zum Positiven verändert haben. Zur Analyse der Veränderung ausgewählter methodischer Studienmerkmale wird eine Auswahl aller Forschungsartikel der Jahre 2010 – 2021 aus 12 Zeitschriften zugrunde gelegt ($N = 57.909$). Die Verteilungen der aus den Studien extrahierten methodischen Eigenschaften wird sowohl global, jeweils sechs Jahre vor (2010–2015) und nach (2016–2021) Veröffentlichung des OSC Berichts betrachtet, als auch deskriptiv vergleichend zwischen den Journalen über die Zeit aufgeführt.

H_1: **Die Größe der zugrundegelegten Stichproben steigt, jedoch auf niedrigem Niveau.**

Entscheidende Kosten- und Zeitfaktoren bei psychologischen Forschungsvorhaben stellen die Akquise der Stichproben und die Durchführung der Untersuchungen dar. Heutzutage ist die Akquise großer Stichproben über Onlineerhebungen und -communities, sowie einer steigenden Anzahl an frei zugänglichen Registerdaten, leichter denn je. Da das Publizieren von vielen Studien im Wissenschaftsbetrieb meist höher honoriert wird, als die Verlässlichkeit der Ergebnisse, wird vermutet, dass die mittlere Stichprobengröße psychologischer Studien zwar steigt, jedoch immer noch niedrig ausfällt (Median $N < 200$) und dass der Anteil an Studien mit großen Stichproben ($N > 1.000$) weiterhin sehr niedrig ist ($< 5\%$), jedoch stetig wächst.

H_2: **Der Anteil Studien mit a-priori Power Analyse steigt, jedoch auf niedrigem Niveau.**

Da die Psychologie meist kleine bis mittel große Effekte untersucht und somit relativ große Stichproben benötigt werden, um eine hohe a-priori Power zu erzielen, wird vermutet, dass der Anteil an Studien mit a-priori Power Analyse noch immer gering ist ($< 20\%$), jedoch kontinuierlich steigt.

H_3: **Der Anteil Studien mit gerichteten Tests steigt, jedoch auf niedrigem Niveau.**

In der psychologischen Forschungspraxis ist der ungerichtete Nil-Nullhypothesentest omnipräsent, obwohl die Richtung eines zu erwartenden Effekts meist aus der Theorie ableitbar ist und die Power eines gerichteten Tests bei gleichem α-Fehlerniveau höher ist, als beim ungerichteten Test, sofern der wahre Effekt in die prognostizierte Richtung zeigt. Es wird erwartet, dass der Anteil an Studien mit gerichteten Tests zwar niedrig ist ($< 20\%$), jedoch über die Zeit steigt.

H_4: **Der Anteil Studien mit p-Wert/α-Fehler Korrektur ist unverändert niedrig.**

Da α-Fehler Korrekturen in Studien mit kleinen Stichproben die Power eines Tests stark verringern, wird vermutet, dass der Anteil Studien mit Korrekturverfahren für p-Werte/α-Fehler unverändert niedrig ist ($< 20\%$).

H_5: **Es besteht ein negativer Zusammenhang zwischen der Anzahl statistischer Testergebnisse und Korrekturverfahren.**

Es wird erwartet, dass der Einsatz von p-Wert/α-Fehler Korrekturverfahren negativ mit der Anzahl extrahierbarer statistischer Testergebnisse pro Studie korreliert. Je höher die Anzahl statistischer Testergebnisse in einer Studie, desto seltener kommen Korrekturverfahren zum Einsatz, da diese, bei vielen Tests und konstant gehaltener Stichprobengröße, leicht zu einem hohen Anteil nicht signifikanter Ergebnisse führen würden.

H_6: **Der Anteil nicht signifikanter Ergebnisse steigt über die Zeit.**

Da ein steigendes Bewusstsein für das *file drawer problem* vermutet wird und immer mehr Journale wie auch AutorInnen eine Präregistrierung von Studien fordern, wird erwartet, dass der Anteil berichteter, jedoch nicht signifikanter Ergebnisse über die Zeit steigt.

5 Das R Paket JATSdecoder

Text Mining bezeichnet die Entdeckung neuer, bisher unbekannter Informationen aus verschiedenen schriftlichen Quellen, durch automatische Extraktion mit Hilfe eines Computers (Gupta, Lehal et al., 2009). Das im Rahmen dieser Arbeit entwickelte Paket *JATSdecoder* (Böschen, 2022) für die Statistiksoftware R (R Core Team, 2020) dient der Überführung wissenschaftlicher Forschungsartikel, die im Format der Journal Archiving Tag Suite NISO-JATS (National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 2014) kodiert wurden, in eine semistrukturierte Datenform. Das NISO-JATS ist ein standardisiertes HTML Markup für Zeitschriftenartikel, in dem, unter anderen, alle frei zugänglichen Dokumente der PubMed Central (PMC) Datenbank (PubMed-Central, 2020) gespeichert sind. Im Januar 2023 waren bereits mehr als 5 Millionen Dokumente aus den Bereichen Biologie, Medizin und Gesundheitswissenschaften in der PMC verlinkt. Mit Hilfe von Textextraktionsalgorithmen extrahiert *JATSdecoder* Metadaten, Textteile und einige methodische Studieneigenschaften, sowie die im Text aufgeführten statistischen Ergebnisse aus dieser Art von Dokumenten. Obwohl das NISO-JATS Format nicht explizit für textanalytische Aufgaben entwickelt wurde, ermöglicht die streng einheitliche Struktur eine verlässliche Extraktion der einzelnen Elemente eines Artikels und erleichtert die Bearbeitung textanalytischer Fragestellungen, die sich auf Inhalte wissenschaftlicher Literatur beziehen. Des Weiteren eröffnen die mit *JATSdecoder* extrahierten Elemente und Studieneigenschaften neue Möglichkeiten für die Wissenschaftsforschung.

JATSdecoder besteht aus den Hauptfunktionen *JATSdecoder()* und *study.character()*, deren Unterfunktionen auch einzeln nutzbar sind, sowie einigen allgemeinen Hilfsfunktionen, die in vielen der Extraktionsprozesse Anwendung finden und auf beliebigen Textinput anwendbar sind. In Tabelle 1 sind die in *JATSdecoder* enthaltenen Hilfsfunktionen und ihre Funktionalität aufgeführt.

Mit Hilfe der Open Source Software *CERMINE* (Tkaczyk, Szostek, Fedoryszak, Dendek & Bolikowski, 2015) lassen sich PDF Dokumente in NISO-JATS kodierte XML Dateien konvertieren, wodurch eine Anwendung der *JATSdecoder* Funktionen auch auf

Tabelle 1. *Im JATSdecoder Paket enthaltene Hilfsfunktionen zur Textverarbeitung*

Funktion	Funktionalität
text2sentences()	konvertiert Fließtext in einen Satzvektor
text2num()	konvertiert ausgeschriebene Zahlen, Brüche, Potenzen und Prozentangaben in die Dezimalschreibweise
ngram()	extrahiert eine definierbare Anzahl von Wörtern, die vor und nach einem Suchbegriff stehen
strsplit2()	teilt Text vor oder nach einem Suchbegriff, ohne diesen zu entfernen
grep2()	ermöglich die Verknüpfung mehrer Suchbegriffe mit logischem UND
letter.convert()	vereinheitlicht und konvertiert hexadecimal und HTML Zeichen zu Unicode und führt *CERMINE* spezifische Fehlerkorrektur durch
which.term()	gibt die Suchbegriffe aus, die in einem Text gefunden wurden

Dokumente, die nicht in der PMC Datenbank enthalten sind, ermöglicht wird. Anzumerken ist, dass bei der Konvertierung mit *CERMINE* den NISO-JATS Tags teilweise falsche Textteile zugeordnet werden. Da *CERMINE* keine optische Buchstabenerkennung anwendet und die Kodierung von Sonderzeichen in einigen PDF Dokumenten nicht nach einer standardisierten Norm erfolgt, werden einige wichtige Sonderzeichen und Operatoren mitunter falsch (z.B. '5', bzw. '1/4' anstatt '=') oder gar nicht ausgegeben. Wie eigene Testdurchläufe zeigen, trifft dies auch auf alle anderen frei verfügbaren Softwarelösungen zu, die eine Überführung von PDF Dokumenten in Textdateien ohne optische Buchstabenerkennung durchführen. Um die im fehlerhaft kompilierten Text enthaltenen statistischen Ergebnisangaben trotzdem mit *JATSdecoder* auslesbar zu machen, wurden Textmanipulationsalgorithmen entwickelt, die falsch oder gar nicht extrahierte Sonderzeichen in statistischen Ergebnisangaben korrigieren/vereinheitlichen, bzw. ersetzen. Nicht konvertierte Operatoren werden durch '<=>' ersetzt, fehlerhaft konvertierte Operatoren und Zeichen korrigiert.

Beispiele für die Ersetzung von nicht und fehlerhaft konvertierten Operatoren und Zeichen durch die Funktion *letter.convert()*:

- 'F (2,23) 12.3, p 0.05' → 'F(2,23)<=>12.3, p<=>0.05'
- 'v2(12) 5 2.3, p 5 0.05 → 'chi2(12)=2.3, p=0.05'

5.1 Die Funktion JATSdecoder()

Das Ergebnis der Funktion *JATSdecoder()* ist eine Liste, die die Metadaten und in Abschnitte gegliederten Textteile eines NISO-JATS kodierten Dokuments enthält. Mit Hilfe dieser Liste lassen sich gezielt Textabschnitte (z.B. Abstract, Methoden-, Ergebnissektion) auswählen und hinsichtlich textanalytischer Fragestellungen weiterverarbeiten, weshalb *JATSdecoder* als allgemeine Schnittstelle für Systeme der Wissenschaftsforschung angesehen werden kann. In Tabelle 2 sind die in der Funktion *JATSdecoder()* implementierten Funktionen und die damit extrahierten Inhalte aufgelistet.

Tabelle 2. *In JATSdecoder() enthaltene Funktionen und extrahierter Inhalt*

Funktion	Extrahierter Inhalt
get.title()	Titel
get.author()	Authorennamen als Vektor
get.aff()	Involvierte Institutionen
get.journal()	Zeitschrift
get.vol()	Zeitschriftenband
get.doi()	Digitale Artikelkennung (DOI)
get.history()	Publikationshistorie
get.country()	involvierte Herkunftsländer der Autoren als Vektor
get.type()	Dokumentenart (‘research article’, ‘editorial’, etc.)
get.subject()	Stichwörter als Vektor
get.keywords()	Stichwörter als Vektor
get.abstract()	Zusammenfassung als Vektor
get.text()	Überschriften und dazugehöriger Text als Liste
get.references()	Literaturangaben als Vektor

5.2 Die Funktion study.character()

Die Funktion *study.character()* führt mit Hilfe von Wörterbuchsuchen, Textmanipulations- und Textextraktionsalgorithmen weitere Analyseschritte bezüglich methodischer Studieneigenschaften durch und gibt eine Liste mit diesen Eigenschaften aus. Neben den, im jeweiligen Methoden- und Ergebnisteil eines Artikels aufgeführten statistischen Methoden, werden die Anzahl berichteter Studien/Experimente, die verwendete/n Softwarelösung/en, α-Fehler, Korrekturverfahren für multiples Testen, sowie die im Text aufgeführten statistischen Ergebnisangaben ausgegeben. Ergebnisse aus

Tabellen und Grafiken können bisher nicht extrahiert werden. Auf Basis sprachlicher Angaben im Abstract und den identifizierten Freiheitsgradangaben in den statistischen Ergebnisangaben erfolgt zudem eine Schätzung der zugrunde gelegten Stichprobengröße. Ein bereits etabliertes Werkzeug zur Extraktion statistischer Ergebnisse aus wissenschaftlichen Berichten ist *statcheck* (siehe Abschnitt 3.7). Im Vergleich zu *statcheck* erkennt die in *JATSdecoder* enthaltene Funktion *get.stats()* jedoch auch solche Ergebnisse, die nicht dem APA-Standard entsprechen. Durch die von *letter.convert()* durchgeführten Korrekturen werden statistische Ergebnisangaben, auch in von *CERMINE* konvertierten PDF Dateien verlässlich extrahier- und prüfbar. Für ausreichend genau berichtete Ergebnisse (z.B.: t(12)=1.2, p=.25) werden p-Werte berechnet, wodurch diese einem Vollständigkeits- und Plausibilitätscheck unterzogen werden können, sofern das jeweilige Ergebnis ebenfalls einen p-Wert enthält.

Tabelle 3. *In study.character() enthaltene Funktionen und extrahierte Studieneigenschaften*

Funktion	Extrahierte Studieneigenschaft
get.stats()	statistische Ergebnisse
get.method()	statistische Methoden
get.alpha.error()	α-Fehlerniveau
get.power()	Angaben zur Teststärke
get.assumption()	statistische Annahmen
get.multiple.comparison()	Verfahren zur α-Fehler
get.software()	Name/n der verwendeten Analysesoftware
get.R.package()	verwendete R package/s
est.ss()	Schätzung der Stichprobengröße
get.n.studies()	Anzahl Studien
get.sig.adjectives()	unangemessene Adjektive, die im Kontext signifikanter und nicht signifikanter Ergebnisse genutzt werden

Die *study.character()*-Ergebnisse stellen eine neue Möglichkeit der Studienidentifikation und -analyse in großen Textdatenbanken dar. Neben der klassischen Suche nach Schlagwörtern im Titel, den Schlüsselwörtern und dem Abstract können nun diverse weitergehende Ein- und Ausschlusskriterien bezüglich der methodischen Aspekte einer Studie definiert und deren Verwendung über die Zeit analysiert werden. So lassen sich beispielsweise Studien identifizieren, die eine bestimmte statistische Methode anwenden und dabei eine große Stichprobe zugrunde legen. Über die extrahierten statistischen Ergebnisse lassen sich zusätzlich Studien mit großen/kleinen Effekten identifizieren, bzw. eine Übersicht und Analyse aller in einer Artikelauswahl enthaltenen Effekte realisieren.

6 Publikation 1: Das R-Paket JATSdecoder

Böschen, I. (2021). Software review: The JATSdecoder package – extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed central's open access database. Scientometrics 126, 9585–9601. DOI: 10.1007/s11192-021-04162-z. (Anhang A)

6.1 Einführung

JATSdecoder ist ein Softwarepaket für die freie Statistiksoftware R (R Core Team, 2020), das Textextraktionen und -analysen von NISO-JATS-kodierten XML-Dokumenten erleichtert. Die gleichnamige Funktion *JATSdecoder()* extrahiert die Metadaten, die Zusammenfassung, den in Abschnitte gegliederten und selektierbaren Text, sowie die Referenzliste eines Dokuments. Weiterhin lassen sich einige der in *JATSdecoder* enthaltenen Funktionen für textanalytische Fragestellungen mit beliebigem Textinput nutzen. Mit mehr als 3,2 Millionen Dokumenten ist PubMed-Central (PMC) eine der größten, frei zugänglichen Volltextsammlungen aus den Bereichen Biologie, Medizin und Gesundheitswissenschaften, deren Inhalte auch im NISO-JATS Format angeboten werden und hier als Datengrundlage dienen.

6.2 Methode

Dieser Bericht gibt einen Überblick über die mit *JATSdecoder()* extrahierten Metadaten aller in der PMC verlinkten Dokumente (Stand Ende 2020: 3.235.065 Artikel). Der Anteil erfolgreich extrahierter Tags (spezifischer Inhalte) wird für den gesamten Korpus über die Zeit und der Inhalt von einigen Meta-Tags im Detail dargestellt. Weiterhin werden Potentiale und Grenzen für Text-Miner, die mit wissenschaftlicher Literatur arbeiten, aufgezeigt.

6.3 Ergebnisse

Die NISO-JATS-Tags werden innerhalb der Artikel der PMC Datenbank äußerst konsistent verwendet, wodurch eine zuverlässige Extraktion der Metadaten und

Textelemente mit *JATSdecoder()* ermöglicht wird. Schlagwörter werden sowohl im subject-, als auch im keyword-Tag hinterlegt. Angaben zu EditorInnen werden am seltensten aufgeführt. Internationale Kollaborationen sind präsenter denn je. Nur etwa die Hälfte aller Artikel aus dem Jahr 2020 enthält mindestens eine/n AutorIn, die/der mit einer AutorInnen-ID (z.B. ORCiD) aufgeführt ist. Viele AutorInnen, vor allem aus dem asiatischen Raum, tragen denselben Namen und gleichzeitig stammen immer mehr Dokumente aus Asien. Es werden teilweise sehr viele verschiedene Hexadezimalcodes verwendetet, die als Synonym für die gleichen Sonderzeichen dienen. So wurden beispielsweise mehr als 20 verschiedene Kodierweisen von Leerzeichen identifiziert, die *JATSdecoder* zu einem Standardleerzeichen vereinheitlicht.

Tabelle 4. *Relative Häufigkeit der erfolgreich extrahierten/enthaltenen NISO-JATS Elemente in Artikeln der PMC-Datenbank über die Zeit*

Inhalt	Anteil	[1781; 2000]	(2000; 2005]	(2005; 2010]	(2010; 2015]	(2015; 2020]
abstract	87,9%	41,3%	71,7%	87,9%	88,9%	92,2%
affiliation	94,8%	38,4%	89,2%	96,4%	98%	98,4%
author	96,9%	56,2%	96,9%	98,9%	99,2%	99,2%
country	82,9%	15,4%	74,7%	86,9%	88,4%	86,2%
doi	100%	100%	100%	100%	100%	100%
editor	12,6%	0,1%	1,7%	10,2%	18,2%	11,5%
history	100%	100%	100%	100%	100%	100%
journal	100%	100%	100%	100%	100%	100%
keywords	59,9%	8%	42,2%	35,1%	52,9%	72%
references	90,3%	9,7%	69,6%	92,2%	94,6%	96%
sections	88,6%	8,9%	71,9%	91,3%	93,6%	93,6%
subject	99,8%	100%	99,9%	99,7%	99,8%	99,8%
text	88,5%	8,8%	71,9%	91,2%	93,5%	93,5%
title	100%	99,9%	99,8%	100%	100%	100%
type	100%	100%	100%	100%	100%	100%
volume	48,7%	93,4%	82,4%	77,1%	54,9%	36,9%
Anzahl Artikel	3.235.065	171.394	52.488	239.952	924.154	1.846.979

6.4 Diskussion

In Verbindung mit den reichhaltigen, öffentlich zugänglichen Inhalten in der PMC-Datenbank können mithilfe von *JATSdecoder* neue Monitoring-, Selektions- und Text Mining-Ansätze verfolgt werden. Bei einer Auswahl von Artikeln über Schlagwörter sollten Ein- und Ausschlusskriterien für verschiedene NISO-JATS-Tags definiert werden, da sowohl die subject- als auch die keyword-Tags recht uneinheitlich verwendet werden. Besonders bei häufig vorkommenden AutorInnen- und EditorInnennamen ist die Artikelidentifikation durch die seltene Verwendung eindeutiger Identifikatoren (ORCiD) erschwert. In einigen Dokumenten werden Ergebnisangaben nicht als Text, sondern als hyperreferenzierter bildlicher Inhalt bereitgestellt, was eine Extraktion des Ergebnisses erheblich erschwert und bisher mit *JATSdecoder* nicht möglich ist.

Das Paket *JATSdecoder*, sowie eine Dokumentation mit Anwendungsbeispielen, wird frei zugänglich bereitgestellt über:

`https://.github.com/ingmarboeschen/JATSdecoder`

und

`https://CRAN.R-project.org/package=JATSdecoder`

7 Publikation 2: Extraktion statischer Ergebnisangaben mit get.stats()

Böschen, I. (2021). Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. Scientific Reports 11, 19525. DOI: 10.1038/s41598-021-98782-3. (Anhang B)

7.1 Hintergrund

Eine automatisierte Extraktion von statistischen Ergebnissen aus wissenschaftlichen Berichten ist für eine Plausibilitätsprüfung, aber auch für die Identifikation von Studien, die bestimmte statistische Prüfgrößen anführen, bzw. bestimmte Effektmaße oder -ausprägungen berichten, von großem Vorteil. Anhand der Freiheitsgradangaben lassen sich beispielsweise Rückschlüsse auf die Anzahl an involvierten Beobachtungen tätigen. Die im R-Paket *JATSdecoder* enthaltene Funktion *get.stats()* extrahiert alle berichteten statistischen Ergebnisse, die im Haupttext eines im NISO-JATS kodierten Artikels aufgeführt sind und berechnet die p-Werte für viele statistische Standardprüfgrößen neu. Die Funktionsweise von *get.stats()* basiert auf der allgemeinen Definition, dass jede Buchstaben- oder Buchstaben-Zahlen-Kombination, die über einen Operator (<, <=, =, ...) mit einer Zahl verbunden ist, ein Ergebnis darstellt. Mit fein abgestimmten, regulären Ausdrücken werden zusammenhängende Ergebnisse identifiziert und nach einer Bereinigung von Indizes auf Angaben von vielen statistischen Standardkennwerten durchsucht (z.B.: t, F, r, β, η^2, p). Bei ausreichend genauen Angaben (z.B. '$\beta = 1.2, SE = 0.2$') werden diese Ergebnisse auch bei nicht vorhandener Angabe eines p-Wertes für die Neuberechnung der p-Werte weiterverarbeitet. Die Ausgabe kann dann auf Ergebnisse mit überprüfbaren oder berechenbaren p-Werten begrenzt werden und als Matrixausgabe leicht weiterverarbeitet werden.

7.2 Methode

In diesem Artikel wird die Fähigkeit von *get.stats()*, statistische Ergebnisse zu extrahieren, neu zu berechnen und zu überprüfen, mit einer händisch erfolgten Zählung

(Wicherts et al., 2011) und der des bereits etablierten Tools *statcheck* (Epskamp & Nuijten, 2018) verglichen. Der von Wicherts et al. (2011) manuell kodierte Datensatz, der die Anzahl statistisch signifikanter Ergebnisse aus 49 Artikeln enthält, dient als erster Indikator für die unterschiedlichen Identifikationsraten von *get.stats()* und den in *statcheck* implementierten Algorithmen. Weitere 13.531 PDF-Artikel aus 10 psychologischen Closed-Access Zeitschriften, sowie 18.744 XML-Dokumente von *Frontiers of Psychology* und 23.730 Artikel, die psychologische Forschungsarbeiten enthalten und in *PLoS One* veröffentlicht wurden (beide Open-Access), werden mit beiden Algorithmen nach statistischen Ergebnissen durchsucht und die identifizierte Anzahl an statistischen Ergebnissen gegenübergestellt.

7.3 Ergebnisse

get.stats() repliziert nahezu die manuell extrahierte Anzahl signifikanter Testergebnisse in den 49, von Wicherts et al. (2011) analysierten PDF-Artikeln. Weiterhin erhöhen die von *get.stats()* extrahierten Rohergebnisse die Erkennungsrate von Testergebnissen durch *statcheck*. Auch in den weiteren 56.005 Artikeln übertrifft *get.stats()* die *statcheck*-Funktionen bei der Identifizierung statistischer Ergebnisse, in jeder einbezogenen Zeitschrift und jedem Inputformat (siehe Publikation 2, Tabelle 2–5).

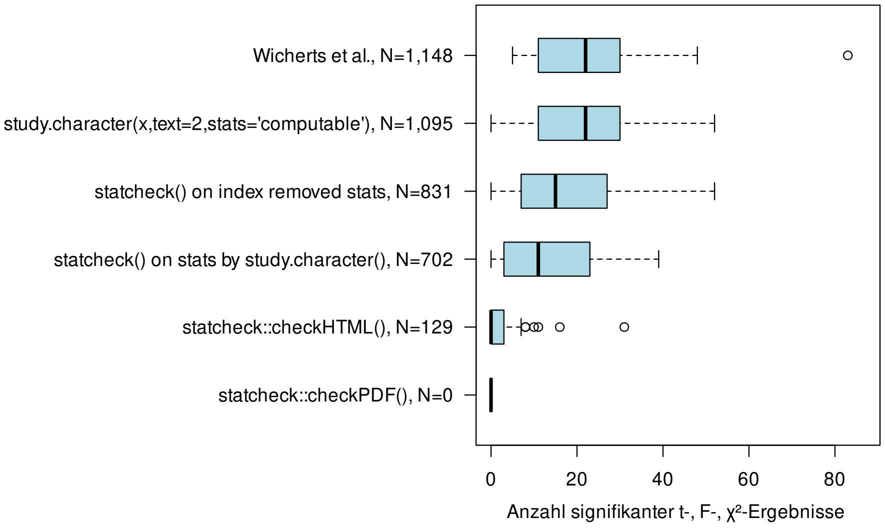


Abbildung 1. Anzahl der extrahierten, signifikanten t-, F- und χ^2-Ergebnisse aus 49 PDF-Artikeln je Methode und Inputformat (*get.stats()* wurde hier durch die Implementation in *study.character()* genutzt)

7.4 Diskussion

Die *JATSdecoder*-Funktion *get.stats()* ist ein sehr allgemeines und zuverlässiges Werkzeug zur Extraktion statistischer Ergebnisse aus reinem Textinput. *get.stats()* erleichtert somit die manuelle und automatische Überprüfung der Konsistenz und Vollständigkeit der im Text berichteten Ergebnisse und ist den in *statcheck* implementierten Algorithmen überlegen. Die Funktion erkennt ein breites Spektrum an textuellen Repräsentationen von statistischen Standardergebnissen und berechnet p-Werte für zwei- und einseitige Tests neu.

Trotzdem ist auch bei Extraktionen mit *get.stats()* mit falsch-positiven und -negativen Extraktionen, sowie Fehlern bei neu berechneten p-Werten zu rechnen. Die meisten Nichterkennungen durch *get.stats()* sind auf Ergebnisse, die in Tabellen berichtet werden, zurückzuführen. Bei Einzelfallbetrachtungen bleibt eine manuelle Überprüfung der Ergebnisse und Bewertung der Studie unerlässlich.

Eine Generalisierbarkeit der Fähigkeiten der Funktion *get.stats()* ist nur eingeschränkt möglich, da sie für Texte entwickelt wurde, die einen Punkt als Dezimaltrennzeichen in den Ergebnisangaben nutzen und in anderen Disziplinen oder Kontexten die Bezeichnungen von Kennwerten (z.B. ‘Z’), eine andere Bedeutung haben können.

Die aktuellste Version von *get.stats()* ermöglicht zusätzlich die Extraktion von Ergebnissen aus DOCX- und HTML-Dateien. Ergebnisse aus einzelnen PDF, DOCX, HTML und XML Dateien können mit *get.stats()* über eine einfache Webanwendung extrahiert und auf Kosistenz überprüft werden: `www.get-stats.app`

8 Publikation 3: Extraktion methodischer Studienmerkmale mit study.character()

Böschen, I. (2023). Evaluation of the extraction of methodological study characteristics with JATSdecoder. Scientific Reports 13, 139. DOI: 10.1038/s41598-022-27085-y. (Anhang C)

8.1 Einführung

In diesem Artikel wird das Modul *study.character()* aus dem R Paket *JATSdecoder* vorgestellt und evaluiert. Mit der Funktion *JATSdecoder()* wird zunächst der Haupttext eines, im NISO-JATS kodierten wissenschaftlichen Artikels, in Einleitung, Methoden-, Ergebnisteil und Diskussion unterteilt. Die in *study.character()* enthaltenen Funktionen extrahieren dann, mit Hilfe von fein abgestimmten, regulären Ausdrücken, methodische Studienmerkmale aus den Methoden- und Ergebnisabschnitten des Dokuments. Dazu gehören die verwendeten statistischen Verfahren, die Analysesoftware, sowie Korrekturverfahren für multiples Testen, das α-Niveau, die berichtete Power, sowie statistische Annahmen, die Anzahl berichteter Studien, die Testrichtung, eine Ausreißerdefinition in Standardabweichung, und ob Interaktionen von Kovariaten betrachtet wurden. Eine automatisiert erstellte Übersicht über und Zugänglichkeit zu diesen Studienmerkmalen auf individueller, aber auch globaler Ebene, ermöglicht neue Perspektiven für methodenbasierte Reviews und Metaanalysen sowie Monitoringansätze, stellt aber auch für den Veröffentlichungsprozess ein hilfreiches Instrument für schnelle Qualitätschecks dar.

8.2 Methode

Ein auf 287 PDF-Artikeln basierender, von Blanca et al. (2018) manuell kodierter Datensatz, dient als Indikator für die Bewertung der Genauigkeit der mit *study.character()* extrahierten Anzahl an berichteten Teilstudien pro Artikel, der angewandten statistischen Methoden und der verwendeten Softwarelösungen. Die Genauigkeit der Extraktion des berichteten α-Fehlerniveaus, der Power, der Verwendung von Korrekturverfahren für multiples Testen, der Analyse von Interaktionseffekten, der Definition von Ausreißern und der Erwähnung statistischer Annahmen wird durch einen Vergleich mit einem händisch ergänzten Datensatz derselben Artikelsammlung bewertet. Sensitivität, Spezifität und Genauigkeit werden für jede der evaluierten Funktionen angegeben.

8.3 Ergebnisse

study.character() extrahiert aus den untersuchten psychologischen Forschungsartikeln die hier fokussierten methodischen Studienmerkmale aus den Methoden- und Ergebnisabschnitten eines im NISO-JATS kodierten Dokuments mit hoher Präzision. Die entwickelten Extraktionsheuristiken weisen eine sehr niedrige Falsch-Positiv-Rate und eine hohe Genauigkeit auf (Accuracy $\geq .9$) und können eine händische Extraktion ersetzen. Die meisten Nicht-Erkennungen sind auf PDF-spezifische Konvertierungsfehler und komplexe Textstrukturen zurückzuführen, die noch nicht verlässlich verarbeitet werden können.

Tabelle 5. *Sensitivität, Spezifität und Accuracy der von study.character() durchgeführten Extraktionen in 287 Artikeln aus acht Zeitschriften*

Eigenschaft	CP	CN	FP	FN	Σ	Sensitivität	Spezifität	Accuracy
α-Niveau aus KI	105	169	0	21	295	0,83	1	0,93
Größtes α-Niveau	137	120	2	28	287	0,83	0,98	0,90
Power	61	242	2	12	317	0,84	0,99	0,96
Korrektur für multiples Testen	65	230	0	2	297	0,97	1	0,99
Entfernung von Ausreißern	24	262	0	1	287	0,92	1	0,99
Testrichtung	33	252	1	1	287	1	1	1
Interaktion/Mediator/Moderator	242	87	10	22	361	0,92	0,90	0,91
Interaktion (binärisiert)	192	87	2	6	287	0,97	0,98	0,97
Statistische Annahmen	79	215	19	12	325	0,87	0,92	0,90
Stat. Annahmen (binärisiert)	64	215	5	3	287	0,96	0,98	0,97
Software	245	105	0	8	358	0,97	1	0,98
Anzahl Studien	285	0	4	4	291	0,99	0	0,97

Hinweis: Die Gesamtanzahl der Detektionen kann Werte annehmen, die größer als $N = 287$ *sind, da Artikel mehrere Merkmale des gleichen Typs aufweisen kann.*

8.4 Diskussion

Das *JATSdecoder* Modul *study.character()* kann eine händische Extraktion der hier fokussierten methodischen Studienmerkmale ersetzen. *JATSdecoder* ermöglicht damit eine großflächige Analyse von methodischen Studieneigenschaften und Forschungspraktiken über die Zeit, für spezielle Zeitschriften und/oder Forschungsgebiete. Weiterhin lassen sich Studienrecherchen durch methodische Ein-/Ausschlusskriterien ergänzen, was für Metaanalysen und systematische Übersichtsarbeiten hilfreich sein kann. Die zur Veranschaulichung entwickelte Webanwendung `https://www.scianalyzer.com` ermöglicht eine Studienselektion und einfache Analysen der extrahierten Mermale für die Inhalte der PubMed Central Datenbank (> 5 Mio. Dokumente).

Neben den von *study.character()* extrahierten Mermalen existieren viele weitere relevante und interessante Studienmerkmale. Zum Beispiel könnten das Studiendesign (experimentell, beobachtend) oder die eingesetzten Messinstrumente (z. B.: Fragebogen, EEG, DNA-Sequenzierung) für eine Auswahl an Studien von Interesse sein. Da die meisten Studienmerkmale hochdimensional sind, wenn sie in einem breiten Spektrum wissenschaftlicher Praxis betrachtet werden, sollte jeweils die Entwicklung anspruchsvoller Werkzeuge zur Verarbeitung natürlicher Sprache in Betracht gezogen werden, die diesem Problem mit ergebnisoffenen Extraktionsprozessen begegnen.

Eine technische Einschränkung besteht darin, dass lediglich NISO-JATS kodierte XML-Dateien in englischer Sprache von *study.character()* verarbeitet werden können. Die spezifischen Funktionen zur Extraktion der Studieneigenschaften können auf beliebige Textinhalte angewendet werden und somit auch Text aus anderen Formaten verarbeitet werden.

9 Publikation 4: Analyse der Verwendung und Veränderung methodischer Studienmerkmale in psychologischen Forschungsberichten (2010 – 2021)

Böschen, I. (2023). Changes in Methodological Study Characteristics in Psychology between 2010 – 2021. PLoS One. DOI: 10.1371/journal.pone.0283353. (Anhang D)

9.1 Hintergrund

Im Jahr 2015 wiederholte die Open Science Collaboration eine Reihe von 100 psychologischen Experimenten (Open Science Collaboration, 2015). Da in einem nicht geringen Anteil dieser Untersuchungen die ursprünglichen Effekte nicht replizierten und diese teilweise in entgegengesetzter Richtung auftraten, wurde Teilen der psychologischen Forschung mangelnde Reproduzierbarkeit attestiert. Es gibt zahlreiche allgemeine Kritikpunkte, die diesen Befund erklären können, wie die inflationäre Anwendung ungerichteter Nil-Nullhypothesentests (Meehl, 1967; Nickerson, 2000), zu kleine (Cohen, 1962; Sedlmeier & Gigerenzer, 1992), selektive Stichproben (Henrich et al., 2010), fehlende Korrekturen für multiples Testen. Einige weit verbreitete, fragwürdige wissenschaftliche Forschungspraktiken wie p-hacking (Simmons et al., 2011) oder HARKing (Kerr, 1998), können die Rate falsch positiver Befunde stark erhöhen. Weiterhin bestehen viele positive Anreize, nur positive/signifikante Ergebnisse zu veröffentlichen (Rosenthal, 1979).

9.2 Methode

Um zu analysieren, ob und wie stark sich die empirische Forschungspraxis in den Jahren 2010 – 2021 im Bereich der Psychologie hinsichtlich ausgewählter Parameter verändert hat, wird eine Auswahl von 57.909 Artikeln aus 12 Zeitschriften verschiedener Teilgebiete der Psychologie, mit der Software *JATSdecoder* (Böschen, 2022) verarbeitet. Um zeitschriften- und zeitspezifische Veränderungen zu identifizieren, wird der Anteil an Artikeln mit p-Werten und die absolute Anzahl an p-Werten pro Artikel, das Berichten von Konfidenzintervallen, die Anwendung gerichteter Tests, Power-Analysen, Bayes'schen

Verfahren, α-Fehlerausprägungen, Korrekturverfahren für multiples Testen und zugrunde gelegten Stichprobengrößen im Zeitverlauf, sowie für Artikel, die vor und nach 2015 veröffentlicht wurden, analysiert. Um zu untersuchen, ob sich die Herkunft psychologischer Forschung und der zugrunde gelegten Stichproben verändert hat, wird die Verteilung der Herkunftsländer der Autorenschaften über die Zeit dargestellt.

9.3 Ergebnisse

Global betrachtet ist der Median der Anzahl im Text berichteter p-Werte, im Vergleich von Artikeln, die in den Jahren 2010 bis 2015 veröffentlicht wurden, mit Artikeln, die nach 2015 veröffentlicht wurden, von 14 auf 12 gesunken. Es wird fast ausschließlich mit dem Standard-α-Niveau von 5% gearbeitet. Gleichzeitig werden überwiegend signifikant von Null abweichende Ergebnisse berichtet. Der globale Median des Anteils signifikanter p-Werte pro Artikel liegt bei 67%. Das heißt, dass in der Hälfte aller Artikel mehr als zwei Drittel aller extrahierten Ergebnisse statistische Signifikanz aufweisen.

Tabelle 6. *Veränderung der Studieneigenschaften in Forschungsartikeln mit statistischen Ergebnissen vor und nach 2015*

Eigenschaft	$\leq$ 2015	> 2015	Gesamt
Gesamtanzahl Artikel	21.238	42.571	63.809
Anzahl empirischer Forschungsartikel	19.443	38.466	57.909
Anteil an Artikeln mit p-Wert	0,92	0,82	0,85
-> Median Anzahl p-Werte je Studie	14	12	15
-> Anteil an Artikeln mit neuberechenbaren p-Wert	0,69	0,58	0,62
-> Anteil an Artikeln mit prüfbaren p-Wert	0,67	0,55	0,60
-> Median Anteil von berichtetem p < 0,05	0,65	0,67	0,67
-> Median Anteil von neu berechnetem p < 0,05	0,73	0,75	0,74
-> Anteil an Artikeln mit α-Niveau < 0,05	0,03	0,02	0,02
-> Anteil an Artikeln mit α-Niveau < 0,01	0,01	0,01	0,01
Anteil an Artikeln mit Konfidenzintervall	0,21	0,32	0,28
Anteil an Artikeln mit Power Analyse/Wert	0,05	0,11	0,09
Anteil an Artikeln mit Bayesianische Analyse	0,02	0,05	0,04
Anteil an Artikeln mit Korrektur für multiples Testen	0,27	0,23	0,24
Anteil an Artikeln mit einseitigem Test	0,05	0,03	0,04
Anteil an Artikeln mit schätzbarer Stichprobengröße	0,83	0,77	0,79
-> Median der geschätzten Stichprobengröße	105	190	151

Der Anteil empirischer Studien, die die Anwendung von Korrekturmethoden für multiples

Testen berichten, ist etwas zurückgegangen (von 27% auf 23%), während Berichte von Konfidenzintervallen zugenommen haben (von 21% auf 32%). Gerichtete Tests und Analysen mit Bayes'schen Inferenzmethoden werden selten berichtet (jeweils 4%), wobei der Anteil an Artikeln mit gerichteten Tests zurückgegangen ist, während Bayes'sche Analysen etwas häufiger berichtet werden. Der globale Median des geschätzten Stichprobenumfangs hat sich von 105 auf 190 erhöht, wobei in den Zeitschriften *Behavioral Neuroscience* und *Psychological Neuroscience* kaum Veränderungen beobachtet wurden. In Artikeln, die nach 2015 veröffentlicht wurden, liegen der Median der geschätzten Stichprobengrößen mit 56 in *Behavioral Neuroscience* und mit 74 in *Psychophysiology*, sowie das .75-Quantil, mit 134 bzw. 143, deutlich niedriger, als in allen anderen betrachteten Zeitschriften (siehe Tabelle 10 und 11 in Publikation 4).
Weiterhin ist eine zunehmende Internationalisierung der psychologischen Forschungsliteratur zu beobachten. Während im Jahr 2010 lediglich in 12% aller Artikel mindestens eine Co-Autorenschaft aus einem nicht-westlichen Land bestand, stammen in 2021 bereits 21,4% aller Artikel aus nicht westlichen Ländern und weitere 22,4% von Kollaborationen westlicher und nicht-westlicher Co-AutorInnen.

9.4 Diskussion

Die hier betrachteten Studienmerkmale haben sich innerhalb der letzten 12 Jahre in den hier betrachteten Zeitschriften sehr unterschiedlich verändert (siehe Publikation 4 für zeitschriftenspezische Ergebnisse im jährlichen Zeitverlauf).

In empirischen Arbeiten der Psychologie werden erstaunlich viele Nil-Nullhypothesentests durchgeführt, deren Ergebnisse zumeist mit p-Werten, und immer öfter auch mit Konfidenzintervallen, aufgeführt werden. Dabei wird zuallermeist mit einem α-Niveau von 5% gearbeitet und ungerichtet getestet, was mit sehr vielen, signifikant von Null abweichenden Ergebnissen einhergeht. Es ist zu erwarten, dass sich mit weiter steigenden Stichprobengrößen der Anteil signifikanter Ergebnisse ebenfalls weiter erhöhen wird, was ein Umdenken hinsichtlich der zumeist, in der H_0 postulierten, Null-Effekte nötig macht. Würden hingegen gerichtete Nullhypothesen gegen einen marginalen oder gerade so bedeutsamen Effekt mit hoher Power getestet werden, könnten auch nicht signifikante Ergebnisse eine Theorie untermauern, bzw. signifikante Ergebnisse diese in Frage stellen. Um post-hoc Entscheidungen bezüglich der zu testenden Effekte, der Power und des α-Niveaus zu vermeiden, ist die Präregistrierung von Forschungsvorhaben weiterhin unerlässlich.

Einhergehend mit der Internationalisierung der Autorenschaften, die psychologische Forschungsergebnisse veröffentlichen, darf auch von einer Diversifizierung der involvierten Stichproben ausgegangen werden, zumindest was deren kulturellen Hintergrund angeht. Die involvierten Zeitschriften stellen eine selektive, wenngleich große Stichprobe, psychologischer Forschungsberichte aus renommierten Zeitschriften dar, weshalb eine Verallgemeinerung der Ergebnisse auf die Psychologie als Ganzes nur eingeschränkt möglich ist. Zu beachten ist, dass die meisten zeitschriftenspezifischen Baselines der Studieneigenschaften und ihre Veränderungen über die Zeit deutliche Unterschiede aufweisen (siehe: Tabellen 2–11 in Publikation 4). Weiterhin sind die globalen Effekte von den Eigenschaften der vergleichsweise vielen Open-Access Artikel überlagert, was eine zeitschriftenspezische Analyse in den Vordergrund rücken sollte.

In Tabellen enthaltene Testergebnisse sind nicht Teil der Analyse. Dadurch wird die tatsächlich berichtete Anzahl statistischer Ergebnisse unterschätzt. Weiterhin ist zu berücksichtigen, dass dies mit Verzerrungen der Ergebnisse bezüglich des hohen Anteils signifikanter p-Werte einhergehen könnte.

10 Diskussion der JATSdecoder Algorithmen

JATSdecoder ermöglicht neue Analyse- und Monitoringansätze wissenschaftlicher Veröffentlichungen und erleichtert die Auswahl von Studien anhand von methodischen Eigenschaften für Metaanalysen und systematische Reviews. Weiterhin kann *JATSdecoder* als allgemeine Schnittstelle für Projekte der Wissenschaftsforschung eingesetzt werden, in denen eine Analyse wissenschaftlicher Literatur fokussiert wird. Beispielsweise lassen sich thematische und methodische Trends mithilfe der extrahierten Metadaten und Studieneigenschaften identifizieren und analysieren.

Es konnte gezeigt werden, dass *JATSdecoder*, trotz der teils uneinheitlichen Kodierung, zuverlässig NISO-JATS kodierte XML Dateien in eine einheitliche Datenform überführt. Dies gilt auch für mit CERMINE kompilierte PDF Dateien. Durch die uneinheitliche Kodierung von Text in PDF Dateien kann es jedoch beim Kompilieren zu Fehlern kommen, was die Präzision der entwickelten Algorithmen etwas herabsetzt. Die extrahierten Metadaten und Textteile können leicht für weitergehende Analysen wissenschaftlicher Studien beliebiger Fachgebiete und Themenbereiche eingesetzt werden. Die Funktion *get.stats()* ist den bereits etablierten *statcheck* Algorithmen in Bezug zum Auslesen statistischer Testergebnisse überlegen und erkennt nahezu alle im Text berichteten Ergebnisse. Die von *JATSdecoder* durchgeführte, umfangreiche Vereinheitlichung von Sonderzeichen ist besonders für das Auslesen von statistischen Ergebnissen relevant. In jedem hier analysierten Journal und Format identifiziert die Funktion *get.stats()* mehr statistische Ergebnisse als *statcheck*. Weiterhin erhöht die Extraktion der Rohergebnisse mit *get.stats()* die Identifikationsrate von *statcheck* merklich. Für AutorInnen und EditorInnen ermöglicht *get.stats()* eine solide Überprüfung der Konsistenz berichteter Ergebnisse und kann zur Fehlervermeidung in wissenschaftlichen Veröffentlichungen beitragen. *get.stats()* ermöglicht eine Identifikation von Studien anhand der verwendeten Teststatistiken, der Größe der berichteten Effekte, sowie der Anzahl berichteter Testergebnisse. Als benutzerfreundlich Webapplikation ist *get.stats()* über die Domain: `www.get-stats.app` für einzelne Dokumente im XML-, PDF-, DOCX- und HTML-Format nutzbar.

Die automatisierte Extraktion der hier fokussierten methodischen Eigenschaften wissenschaftlicher Studien mit der Funktion *study.character()* ist mit einer händischen Analyse vergleichbar. Es wird eine schnelle Analyse und Identifikation von Studien anhand dieser Kriterien ermöglicht, was für den Veröffentlichungsprozess und weitergehende inhaltliche Fragestellungen relevant sein kann. So lassen sich beispielsweise Studien identifizieren, die bestimmte statistische Verfahren und/oder Korrekturverfahren für multiples Testen verwenden. Weiterhin ist es möglich, Studien anhand des zugrunde gelegten α-Niveaus, der berichteten Teststärke und/oder der Größe der erhobenen Stichprobe zu identifizieren.

Eine im Kontext dieser Dissertation entwickelte Webanwendung ermöglicht eine solche Auswahl, bzw. Identifikation von Studien aus der PubMed-Central Datenbank, die, bis zum Beginn des Jahres 2023, bereits mehr als 5 Mio. Artikel aus 20,902 Zeitschriften und Autoren aus 208 Ländern enthielt. Weiterhin ermöglicht die Applikation eine einfache, globale und periodische Analyse der mit *JATSdecoder* extrahierten Metadaten und methodischen Studienmerkmale einer getroffenen Auswahl an Studien. Die Rohdaten einer bis zu 20.000 Artikel umfassenden Auswahl können als CSV Datei für individuelle Analysen heruntergeladen werden:

`www.scianalyzer.com`

JATSdecoder ist eine modulare open source Software, wodurch die Einsicht in bestehende Funktionen, sowie die Implementierung neuer Funktionen, einfach möglich ist. Eine Installations- und Nutzungsanleitung, sowie die Möglichkeit der Kollaboration, werden über den *JATSdecoder* github-Account bereitgestellt:

`https://github.com/ingmarboeschen/JATSdecoder`

Die in den Veröffentlichungen genutzten Daten und Auswertungsskripte sind ebenfalls frei zugänglich und können zur Reproduktion der Analysen genutzt werden:

`https://github.com/ingmarboeschen/JATSdecoderEvaluation`.

10.1 Limitationen

Trotz der hier dargestellten, hohen Präzision der in *study.character()* implementierten Algorithmen ist, bedingt durch die Komplexität menschlicher Sprache und das Vorhandensein tatsächlicher Fehler in Dokumenten, mit falsch positiven und falsch negativen Extraktionen zu rechnen.

Die zwar zuletzt deutlich abnehmende, aber immer noch uneinheitliche Kodierung von Sonderzeichen in PDF Dokumenten stellt bei der Verarbeitung von mit CERMINE erzeugten NISO-JATS kodierten Dokumenten ein nicht vollständig gelöstes Problem dar. Die in der *JATSdecoder()* Funktion zusammengefassten Algorithmen extrahieren und restrukturieren die Inhalte eines im NISO-JATS kodierten Dokuments kontextunabhängig verlässlich. Die in *study.character()* enthaltenen Funktionen wurden für englischsprachige Berichte entwickelt, die einen Punkt als Dezimaltrennzeichen verwenden. Weiterentwicklungen von *JATSdecoder* sollten somit eine Option zum Wechsel des zu nutzenden Dezimaltrennzeichens und die Verallgemeinerung der Extraktionsalgorithmen für anderssprachigen Inhalt fokussieren.

11 Diskussion der Befunde zur Entwicklung der methodischen Studienmerkmale in psychologischen Forschungsarbeiten

11.1 Zusammenfassung der Ergebnisse im Kontext der aufgestellten Hypothesen

Einige der in dieser Arbeit untersuchten methodischen Studienmerkmale psychologischer Forschungsarbeiten haben sich im Zeitraum 2010 – 2021 deutlich verändert. Da die Veränderungen zwischen den untersuchten Zeitschriften teilweise sehr heterogen ausfallen, sollte jedoch eine zeitschriftenspezifische Betrachtung in Erwägung gezogen werden.

Positiv hervorzuheben ist, dass sowohl der globale Median, wie auch das .75-Quantil der zugrundegelegten Stichprobengrößen gestiegen ist. Unter der Annahme, dass unverändert starke Effekte betrachtet werden, ist die durchschnittliche Teststärke psychologischer Studien somit gestiegen. Verglichen mit den Artikeln, die zwischen 2010 und 2015 erschienen, haben sich in einigen der untersuchten Zeitschriften innerhalb der darauf folgenden sechs Jahre die Median- und .75-Quantil Stichprobengrößen, mehr als verdoppelt. In den Zeitschriften *Behavioral Neuroscience* und *Psychophysiology* sind der Median, sowie das .75-Quantil der extrahierten Stichprobengrößen nahezu unverändert geblieben und gleichzeitig am niedrigsten.

Insgesamt betrachtet sind in Publikationen aus dem Jahr 2021 poweranalytische Konzepte viel präsenter, als 12 Jahre zuvor. Jedoch zeigen sich auch hier deutliche Unterschiede zwischen den Zeitschriften. Während im Jahr 2021 zwei Drittel aller Studien, die in der Zeitschrift *Personality and Social Psychology Bulletin* und ein Drittel aller Studien, die in *Psychology & Aging* sowie *Psychophysiology* erschienen, die anvisierte, bzw. erreichte Teststärke berichten, ist dies auch im Jahr 2021 in lediglich 3% der Studien des *Journal of Management* der Fall.

Obwohl die Richtung der untersuchten Effekte in der Regel durch die zugrunde gelegte Theorie ableitbar ist und gerichtet durchgeführte Tests kostenfrei die Teststärke erhöhen, finden diese in den hier analysierten Artikeln sehr selten Anwendung. Der globale Anteil an Studien mit mindestens einem gerichteten Test fällt mit 4% niedriger aus als erwartet und variiert kaum zwischen den Zeitschriften. Zudem wurde beobachtet, dass der Anteil an Studien mit gerichteten Tests in Veröffentlichungen zwischen 2016 und 2021 etwas niedriger ausfällt (3%), als in Studien, die zwischen 2010 und 2015 veröffentlicht wurden (5%).

Der Forderung α-Fehlerniveaus kleiner 5% zu verwenden (Johnson, 2013; Ioannidis, 2018; Benjamin et al., 2018), um die falsch positive Rate von Ergebnissen zu reduzieren und somit höhere Robustheit im Sinne einer Replizierbarkeit zu erhalten, wurde bisher nicht nachgekommen. In 98% der analysierten Artikel wird mit α-Fehlerniveaus $\geq$ 5% gearbeitet. Da es weiterhin die allgemeine Praxis darstellt, ohne vorherige Registrierung, ungerichtet Nil-Nullhypothesen zu testen, erscheint es wenig verwunderlich, dass unverändert wenige nicht signifikante Ergebnisse berichtet werden. Insgesamt sind 68% der 453,295 identifizierten Ergebnisse, die eine Neuberechnung des p-Wertes mit *get.stats()* erlauben, bei einem α-Niveau von 0,05 signifikant. Der Median des Anteils der berichteten p-Werte $< 0,05$ pro Studie ist mit 69% unverändert hoch. Bei Beibehaltung der allgemeinen Praxis Nil-Nullhypothesen zu testen dürften sich diese Anteile, bei weiter steigenden Stichprobengrößen und unverändertem α-Niveau, noch weiter erhöhen, was die Notwendigkeit einer Abkehr von dieser Praxis verdeutlicht.

Dass sich hier, trotz der gestiegenen Stichprobengrößen, der Anteil signifikanter Ergebnisse nicht merklich erhöht hat, ist ein unerwartetes Ergebnis, über dessen Ursachen nur spekuliert werden kann. Einerseits ließe sich dieser Befund mit kleineren, im Fokus stehenden Effekten erklären, andererseits könnten die größeren Stichproben es weniger notwendig machen, fragwürdige Forschungspraktiken anzuwenden, um signifikant von null abweichende Ergebnisse zu erhalten.

Die Verwendung von Korrekturverfahren für multiples Testen wird vergleichsweise selten berichtet, wenngleich in den allermeisten Studien mehrere und teilweise sehr viele statistische Testergebnisse berichtet werden. Der globale Anteil an Studien mit Korrekturverfahren für multiples Testen liegt mit 24% etwas höher als erwartet (20%), hat sich jedoch von 27% in Studien, die zwischen 2010 und 2015 veröffentlicht wurden, auf 23% in Studien, die im Zeitraum 2015 – 2021 veröffentlicht wurden, verringert. Der

Median der Anzahl berichteter p-Werte je Studie ist von 14 auf 12 gesunken, was jedoch nicht die allgemeine Notwendigkeit von Korrekturverfahren für multiples Testen mindert. In diesem Zusammenhang ist zu betonen, dass Testergebnisse aus Tabellen nicht extrahiert wurden und somit die Gesamtanzahl an tatsächlich berichteten Testergebnissen deutlich größer sein dürfte. Anders als erwartet, steigt der Anteil an Studien, die Korrekturverfahren für multiples Testen anwenden, mit zunehmender Anzahl extrahierbarer p-Werte in fast allen hier analysierten Zeitschriften, mit großen Unterschieden zwischen den Zeitschriften. In Artikeln mit zwei extrahierbaren p-Werten liegt der globale Anteil an Artikeln mit Korrekturverfahren für multiples Testen mit 12,5% sehr niedrig und mit 35,7% in Artikeln mit 20 bis 40 extrahierbaren p-Werten immer noch niedrig. In der Zeitschrift *Behavioral Neuroscience* wird in der Hälfte aller Artikel der Einsatz von α-Fehlerkorrekturen berichtet, während im *Journal of Management*, fast unabhängig von der Anzahl der extrahierbaren p-Werte, in lediglich 3% aller Artikel. Eine händisch durchgeführte Analyse der Anwendungsszenarien von Korrekturverfahren für multiples Testen könnte der hier nicht zu beantworten Frage nachgehen, ob die Korrekturen auf alle berichteten Testergebnisse angewendet wurden, oder nur auf einzelne Vergleiche und ob es spezifische Verfahren gibt, bei denen diese Verfahren gehäuft angewendet werden. Beispielsweise wäre ein ledigliches Anwenden bei einzelnen oder post-hoc Vergleichen denkbar, da das strikte Korrigieren einen negativen Einfluss auf die Power hat und damit oft zur Folge, dass viele, der als signifikant gewerteten Ergebnisse, als nicht signifikant gewertet werden müssten.

In 50% der Studien sind mehr als 67% der berichteten und 74% der neu berechneten p-Werte kleiner als 0,05, ohne bedeutsame Veränderung dieser Anteile über die Zeit. Dieser Befund deutet darauf hin, dass nach wie vor nicht signifikante Ergebnisse eher selten berichtet, bzw. weniger Wert geschätzt werden, als statistisch signifikante Abweichungen von Nulleffekten.

11.2 Limitationen

Die Interpretation der Ergebnisse bezüglich der Veränderung der methodischen Studienmerkmale muss mit einigen Einschränkungen erfolgen.

Die 57.909 psychologischen Forschungsartikel aus 12 Jahren und 12 Zeitschriften, die hier mit *JATSdecoder* analysiert wurden, stellen eine große, wenngleich selektive Stichprobe aller psychologischen Forschungsberichte dar, weshalb eine Generalisierbarkeit der Ergebnisse auf die Psychologie als Ganzes, nur eingeschränkt möglich ist.

Gleichzeitig stellen die hier fokussierten methodischen Studieneigenschaften zwar eine große, wenngleich nicht ausschöpfende Auswahl an methodischen Qualitätsmerkmalen psychologischer Studien dar. So finden beispielsweise die Operationalisierung, die Art und Güte der eingesetzten Messinstrumente und zugrunde gelegten Stichproben für die jeweiligen Fragestellungen und Populationen, auf die generalisiert werden soll, keine Beachtung.

Die Funktion, die, basierend auf den Angaben im Abstract und den extrahierten Freiheitsgradangaben in den statistischen Ergebnissen, die Stichprobengröße schätzt, wurde noch nicht kritisch evaluiert. Dies mindert die Verlässlichkeit der Aussage gestiegener Stichprobengrößen, obwohl die beobachteten Tendenzen innerhalb und zwischen den Zeitschriften plausibel erscheinen.

Inhalte aus Tabellen und Abbildungen wurden nicht verarbeitet. Dadurch wird die tatsächlich berichtete Anzahl statistischer Ergebnisse unterschätzt, was mit Verzerrungen der Ergebnisse bezüglich des hohen Anteils signifikanter p-Werte einhergehen könnte.

Trotz der hohen Genaugkeit der Algorithmen ist davon auszugehen, dass die Eigenschaften einiger Studien fehlerhaft bzw. nicht extrahiert wurden.

11.3 Fazit und Implikationen für die Praxis

In den letzten Jahrzehnten sind einige oft wiederholte Ansätze zur Überwindung der Glaubwürdigkeitskrise der Psychologie vorgestellt worden. Bisher wurde jedoch keiner dieser Ansätze konsequent umgesetzt. Als Optimist sehe ich die positiven Veränderungen in den Eigenschaften psychologischer Forschungsarbeiten, wenngleich ich mir höhere allgemeine Standards wünsche. Als Realist erinnere ich mich an Cohen (1990), der die allgemeine Praxis der naiven Anwendung des Nill-Nullhypothesentests mit $\alpha = .05$ kritisierte und zu dem Schluss kam: *'These things take time'* (Cohen, 1990, S. 1311).

Die Psychologie hat sich innerhalb kurzer Zeit zu einer großen, weltweit vernetzten und produktiven Community entwickelt. Von den meisten ForscherInnen werden große Anstregungen unternommen, möglichst transparent und methodisch korrekt zu arbeiten. Gleichzeitig bestehen nach wie vor berechtigte Zweifel an der allgemeinen Gültigkeit, der in psychologischen Forschungspapieren getroffenen Schlussfolgerungen.

Den diversen potenziellen Ursachen für die niedrigen Replikationsraten psychologischer Forschungsergebnisse lässt sich mit dem bisherigen Forschungs- und Qualitätsstandards anscheinend nicht ausreichend begegnen. Da es weiterhin allgemeine Praxis ist, ohne Überlegungen zur Teststärke, ungerichtet und unkorrigiert gegen einen Null-Effekt mit $\alpha = .05$ zu testen, und dabei der Anteil an signifikanten Ergebnissen unverändert hoch geblieben ist, muss davon ausgegangen werden, dass viele falsch positive Befunde in der psychologischen Forschungslisteratur veröffentlicht wurden und werden. Um das Vertrauen in die Befunde der psychologischen Forschung wieder herzustellen und zu sichern, sind deshalb weitere Reformen notwendig. Ich sehe vor allem zwei vielversprechende Ansätze, die die Belastbarkeit psychologischer Forschungsergebnisse zukünftig erhöhen können, das gerichtete Non-Nil-Nullhypothesentesten und die Registrierung von Forschungsvorhaben.

Anstatt ohne begründete Vorabüberlegungen stets gegen einen Nulleffekt zu testen, sollte es sich etablieren, wie bei Äquivalenz-/Effekttests üblich, gegen gut begründete Mindesteffektgrößen zu testen. In medizinischen Wirksamkeitsuntersuchungen hat sich diese Art der Hypothesenfestlegung bereits in Nichtunterlegenheitsstudien etabliert (Schumi & Wittes, 2011; Lakens et al., 2018). Die Vorabfestlegung von Vorhersagen der zu erwartenden Effekte und geplanten Analysen ist bei diesem, aber auch dem Standardvorgehen, ein wichtiges Instrument zur Vermeidung von Bestätigungsfehlern bei der Hypothesenprüfung (Chambers & Tzavella, 2022). Auch eine Umkehr der Testlogik wäre in einem Setting, das nicht gegen Null-Effekte testet, denkbar. Würde mit hoher

Power, bzw. Messgenauigkeit gegen gut begründete a-priori definierte Mindesteffektgrößen, die eine Grenze der praktischen Relevanz markieren, gerichtet getestet werden, könnten nicht signifikant unter diesen Grenzwerten liegende Effekte zwar nicht als Beweis, dafür aber als hinreichender Beleg für eine Theorie bewertet werden. Signifikant unter den festgelegten Grenzwerten liegende Ergebnisse würden die Theorie hingegen entkräften. Dieses Vorgehen entspräche dem wissenschaftlichen Prinzip der Falsifikation nach Popper (1963) und die unternommenen Anstrengungen eine Theorie zu widerlegen würden in den Fokus der Diskussion rücken.

Ein noch relativ neuer, dafür gleich mehreren Problemen gleichzeitig begegnender Ansatz ist unter dem Bergriff 'registrierter Bericht' vorgeschlagen worden (engl. 'registered report') (Scheel, Schijen & Lakens, 2021; Chambers & Tzavella, 2022). Anders als bei prä-registrierten Forschungsvorhaben/-protokollen auf öffentlich zugänglichen Servern (z.B. ClinicalTrials.gov, OSF.io), erfolgt bei diesem Veröffentlichungsweg die Einreichung bei einer Zeitschrift, sowie das Peer-Review Verfahren bereits vor der Durchführung des Vorhabens. Da die Veröffentlichungszusage vor der Datenerhebung und -analyse erfolgt, wird die Veröffentlichung selbst unabhängig vom Ergebnis. Die Bewertung der Veröffentlichungsfähigkeit wird bei registrierten Berichten anhand der Relevanz der Forschungsfragestellung und des methodischen Vorgehens bewertet und nicht anhand des Nachrichtenwerts des Ergebnisses. Ein augenscheinlicher und entscheidender Unterschied dieser Strategie besteht darin, dass auch 'nicht erfolgreiche' Studien Wertschätzung erlangen und so dem 'file-drawer' Problem entgegengewirkt wird. Schwächen im Design, der Stichprobenplanung und dem Auswertungskonzept, können so abgemildert bzw. ganz vermieden werden, während im klassischen, post-hoc durchgeführten Peer-Review Verfahren, die identifizierten methodischen Einschränkungen, lediglich in die Limitationen mit aufgenommen werden können. Durch eine Registrierung von Forschungsvorhaben ließe sich auch am ehesten der allgemeine Standard, mit angefallenen Stichproben, ungerichtet gegen einen Nulleffekt zu testen, aufbrechen.

Ein Indiz für die Stärke des Effekts von Registrierung auf den 'Erfolg' von Studien liefert die Studie von Scheel et al. (2021). In ihrer Analyse von 71 registrierten und 152 nicht registrierten Forschungsarbeiten lag der Anteil an 'erfolgreichen' Ergebnissen in registrierten Arbeiten mit 44% deutlich unter dem Anteil von 96% bei nicht registrierten Arbeiten. Bei den 100 nicht registrierten Studien der OSC-Replikationsstudie (Open Science Collaboration, 2015) lag der Anteil 'erfolgreicher' Studienergebnisse bei 97%.

Das Center for Open Science (2023) listet bereits mehr als 300 Zeitschriften, die einen Veröffentlichungsprozess über eine Registrierung mit Peer-Review anbieten. Von den 12 in Publikation 4 involvierten Zeitschriften bieten bereits 5 diese Option an, darunter die beiden Open-Access Zeitschriften *PLoS One* und *Frontiers in Psychology*.

11.4 Ausblick

Mit zukünftig sicherlich weiter steigenden Qualitätsstandards muss diskutiert werden, wie mit den unzähligen bereits publizierten Studien umzugehen ist, die nicht diesen Standards entsprechen. Allein in den letzten 4 Jahren hat sich die Anzahl frei verfügbarer Forschungsartikel in der PubMed Central Datenbak nahezu verdoppelt. Da das Open-Access Vertriebsmodell heutzutage auch als optionaler Vertriebsweg in den meisten klassischen Zeitschriften angeboten wird, ist davon auszugehen, dass auch in den nächsten Jahren die Anzahl jährlich publizierter Forschungsergebnisse weiter steigen wird. Angesichts dieser immer größer werden Datenlage, die der Weltbevölkerung bereits in großen Teilen permanent frei zugänglich gemacht wird, stellt sich, nicht nur für das Fach der Psychologie, die immer dringlicher werdende Frage nach dem Umgang mit widerlegten bzw. fragwürdigen Studienergebnissen.

Dass Wissenschaft nicht zwingend selbstkorrigierend ist, argumentiert Ioannidis (2012). Dies zeigt sich auch bei einem Blick auf die Online-Artikel der von der Open Science Collaboration (2015) wiederholten Originalstudien. Von den 17 der 100 Originalstudien, deren Effekt in der mit hoher Power durchgeführten Replikationsstudie in die entgegengesetzte Richting zeigte, sind weiterhin alle 17 ohne einen Verweis auf diesen Befund online verfügbar. Auch die mittlerweile leicht zu identifierenden, offensichtlichen Inkonsistenzen in statistischen Ergebnisangaben, die sowohl mit *get.stats()* als auch *statcheck* identifiziert werden können, sind bisher nicht, bzw. nur teilweise behoben worden.

Angesichts der schieren Masse an Veröffentlichungen ist ein Bereinigungs- bzw. Bewertungssprozess der wissenschaflichen Forschungslandschaft durch menschliche Hand schwer vorstellbar. Die in *JATSdecoder* enthaltenen Algorithmen zur Extraktion und Systematisierung von Studienmerkmalen aus Fließtextinhalten können für Weiterentwicklungen methodischer Qualitätskontrolle nützlich sein. Mashine Learning und modellbasierte Methoden zur Verarbeitung menschlicher Sprache erleben zur Zeit einen enormen Entwicklungssprung und könnten bei einer automatisierten Bewertung von Studienergebnissen, sowie bei der Verlinkung zu aktuelleren/robusteren Ergebnissen zuküngftig eine entscheidene Rolle spielen. Denkbar wären automatisch generierte Metriken, die auf der Basis von automatisch extrahierten Studienmerkmalen und Ergebnissen die Verlässlichkeit bzw. die Güte von Publikationen und Ergebnissen bewerten. Die Begründung für die Bewertung in einer solchen Metrik sollte dabei Transparent sein, um bei offensichtlich nicht (mehr) den Standards entsprechenden Inhalten, dem Vorwurf der Zensur vorzubeugen.

12 Literatur

Arnett, J. J. (2008). The neglected 95become less american. *American Psychologist*, *63* (7), 602–614. doi: 10.1037/0003-066X.63.7.602

Bakker, M. & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior research methods*, *43* (3), 666–678. doi: 10.3758/s13428-011-0089-5

Bakker, M. & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type i error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological methods*, *19* (3), 409. doi: 10.1037/met0000014

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … others (2018). Redefine statistical significance. *Nature Human Behaviour*, *2* (1), 6–10. doi: 10.1038/s41562-017-0189-z

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57* (1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82* (397), 112–122. doi: 10.1080/01621459.1987.10478397

Blanca, M. J., Alarcón, R. & Bono, R. (2018). Current Practices in Data Analysis Procedures in Psychology: What Has Changed? *Frontiers in Psychology*, *9*. doi: 10.3389/fpsyg.2018.02558

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, 13–60.

Brown, N. J. & Heathers, J. A. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8* (4), 363–369. doi: 10.1177/1948550616673876

Böschen, I. (2021a). Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. *Scientific Reports*, *11*. doi: 10.1038/s41598-021-98782-3

Böschen, I. (2021b). Software review: The JATSdecoder package - extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed Central's open access database. *Scientometrics*. doi: 10.1007/s11192-021-04162-z

Böschen, I. (2022). JATSdecoder: A Metadata and Text Extraction and Manipulation Tool Set. Zugriff auf `https://CRAN.R-project.org/package=JATSdecoder` (R package version 1.1)

Böschen, I. (2023a). Changes in methodological study characteristics in psychology between 2010-2021. *PloS One*. doi: 10.1371/journal.pone.0283353

Böschen, I. (2023b). Evaluation of the extraction of methodological study characteristics with JATSdecoder. *Scientific Reports*, *13*. doi: 10.1038/s41598-022-27085-y

Center for Open Science. (2023). `https://www.cos.io/initiatives/registered-reports`. (Accessed: 2023-02-21)

Chambers, C. D. & Tzavella, L. (2022). The past, present and future of registered reports. *Nature human behaviour*, *6* (1), 29–42. doi: 10.1038/s41562-021-01193-7

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65* (3), 145–153. doi:

10.1037/h0045186

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45* (12), 1304-1312.

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

Epskamp, S. & Nuijten, M. B. (2018). statcheck: Extract statistics from articles and recompute p values. Zugriff auf `https://CRAN.R-project.org/package=statcheck` (R package version 1.3.0)

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19* (6), 975–991. doi: 10.3758/s13423-012-0322-y

Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9* (6), 641–651. doi: 10.1177/1745691614551642

Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465. doi: 10.1511/2014.111.460

Gelman, A. & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60* (4), 328–331. doi: 10.1198/000313006X152649

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004). The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. In *Handbook on quantitative methods in the social sciences. sage, thousand oaks, ca* (S. 391–408). Citeseer. doi: 10.4135/9781412986311.n21

Greenland, S. (2019). Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values. *The American Statistician*, *73* (sup1), 106–114. doi: 10.1080/00031305.2018.1529625

Gupta, V., Lehal, G. S. et al. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1* (1), 60–76. doi: 10.4304/jetwi.1.1.60-76

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research Online*, *7* (1), 1–20.

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. doi: 10.1371/journal.pbio.1002106

Henrich, J., Heine, S. J. & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33* (2-3), 61–83. doi: 10.1017/S0140525X0999152X

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70. Zugriff auf `http://www.jstor.org/stable/4615733`

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2* (8), e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7* (6), 645–654. doi: 10.1177/1745691612464056

Ioannidis, J. P. (2018). The proposal to lower p value thresholds to. 005. *Jama*, *319* (14), 1429–1430. doi: 10.1001/jama.2018.1536

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23* (5), 524–532. doi: 10.1177/0956797611430953

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the*

National Academy of Sciences, *110* (48), 19313–19317. doi: 10.1073/pnas.1313476110

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2* (3), 196–217. doi: 10.1207/s15327957pspr0203_4

Klemmert, H. (2004). *Äquivalenz- und effekttests in der psychologischen forschung.* Lang.

Kühberger, A., Fritz, A. & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PloS one*, *9* (9), e105825. doi: 10.1371/journal.pone.0105825

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. doi: 10.1177/2515245918770963

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34* (2), 103–115. doi: 10.1086/288135

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM). (2014). *Journal Publishing Tag Library - NISO JATS Draft Version 1.1d2.* https://jats.nlm.nih.gov/publishing/tag-library/1.1d2/index.html.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5* (2), 241–301. doi: 10.1037/1082-989x.5.2.241

Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115* (11), 2600–2606. doi: 10.1073/pnas.1708274114

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S. & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48* (4), 1205–1226. doi: 10.3758/s13428-015-0664-2

Nuijten, M. B., van Assen, M. A., Hartgerink, C. H., Epskamp, S. & Wicherts, J. (2017, 01). *The validity of the tool "statcheck" in discovering statistical reporting inconsistencies.* https://psyarxiv.com/tcxaj.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349* (6251), 943–953. doi: 10.1126/science.aac4716

Popper, K. R. (1963). Science as falsification. *Conjectures and refutations*, *1* (1963), 33–39.

Pritschet, L., Powell, D. & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27* (7), 1036–1042. doi: 10.1177/0956797616645672

PubMed-Central. (2020). *PMC Overview.* `https://www.ncbi.nlm.nih.gov/pmc/about/intro`. (Accessed: 2020-02-25)

R Core Team. (2020). R: A Language and Environment for Statistical Computing [Software-Handbuch]. Vienna, Austria. Zugriff auf `https://www.R-project.org/`

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86* (3), 638–641. doi: 10.1037/0033-2909.86.3.638

Scheel, A. M., Schijen, M. R. & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4* (2), 25152459211007467. doi: 10.1177/25152459211007467

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological methods*, *1* (2), 115-129. doi: 10.1037/14805-019

Schmidt, T. (2017). *Statcheck does not work: All the numbers. Reply to Nuijten et al.* psyarxiv.com/hr6qy.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15* (6), 657–680. doi: 10.1007/BF01068419

Schumi, J. & Wittes, J. T. (2011). Through the looking glass: understanding non-inferiority. *Trials*, *12* (1), 1–12. doi: 10.1186/1745-6215-12-106

Sedlmeier, P. & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105* (2), 309–316. doi: 10.1037/0033-2909.105.2.309

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

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62* (318), 626–633.

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22* (11), 1359–1366. doi: 10.1177/0956797611417632

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J. & Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, *18* (4), 317–335. doi: 10.1007/s10032-015-0249-8

Wasserstein, R. L., Lazar, N. A. et al. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70* (2), 129–133. doi: 10.1080/00031305.2016.1154108

Wicherts, J. M., Bakker, M. & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS one*, *6* (11). doi: 10.1371/journal.pone.0026828

Anhang A: Publikation 1

Böschen, I. (2021). Software review: The JATSdecoder package – extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed central's open access database. Scientometrics 126, 9585–9601. https://doi.org/10.1007/s11192-021-04162-z

Scientometrics
https://doi.org/10.1007/s11192-021-04162-z

Software review: The JATSdecoder package—extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed central's open access database

Ingmar Böschen[1]

Received: 19 August 2021 / Accepted: 8 September 2021

Abstract
JATSdecoder is a general toolbox which facilitates text extraction and analytical tasks on NISO-JATS coded XML documents. Its function *JATSdecoder()* outputs metadata, the abstract, the sectioned text and reference list as easy selectable elements. One of the biggest repositories for open access full texts covering biology and the medical and health sciences is PubMed Central (PMC), with more than 3.2 million files. This report provides an overview of the PMC document collection processed with *JATSdecoder()*. The development of extracted tags is displayed for the full corpus over time and in greater detail for some meta tags. Possibilities and limitations for text miners working with scientific literature are outlined. The NISO-JATS-tags are used quite consistently nowadays and allow a reliable extraction of metadata and text elements. International collaborations are more present than ever. There are obvious errors in the date stamps of some documents. Only about half of all articles from 2020 contain at least one author listed with an author identification code. Since many authors share the same name, the identification of person-related content is problematic, especially for authors with Asian names. *JATSdecoder()* reliably extracts key metadata and text elements from NISO-JATS coded XML files. When combined with the rich, publicly available content within PMCs database, new monitoring and text mining approaches can be carried out easily. Any selection of article subsets should be carefully performed with in- and exclusion criteria on several NISO-JATS tags, as both the subject and keyword tags are used quite inconsistently.

Keywords Meta-research · Text extraction · Text mining · PubMed central · Software

✉ Ingmar Böschen
ingmar.boeschen@uni-hamburg.de

[1] Psychological Methods and Statistics, Institute of Psychology, University Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany

Published online: 24 October 2021

Introduction

Scientists from all over the world cumulate knowledge to describe and better understand the causes and effects in a very wide range of research areas. Nowadays, hundreds of thousands of research-related documents are published every year in an incredibly wide range of disciplines, subjects and journals. Scientific findings are mostly published in peer-reviewed journals that are more and more willing to share their articles, or at least parts, without charging readers a fee in so-called open access journals. Combined with the increasing computational power of personal computers it is possible for anyone to purchase and deal with essentially big amounts of textual data as source of information or research interest.

The Science-Metrix report (Science-Metrix Inc. 2018) quantifies the availability of scientific research findings in the form of open access versions to be about one half of all results published. About one paper out of four is freely made available by the publishers themselves (gold open access), most of the time on their own websites but also frequently mediated by websites such as PubMedCentral, SciELO in some Romance-language countries, and JStage in Japan (Science-Metrix Inc. 2018).

Aside from the benefits of many scientific findings and information being made publicly available, selection and identification problems arise. The identification of articles topic relatedness is a key part for any researcher running a literature search. In order to make an informed judgement about a specific research topic and meta-analytical questions, the identification and preselection of relevant studies as well as the extraction of relevant information are more and more time-consuming tasks as the text corpus is constantly growing. Search engines can help to identify studies by metadata like title, author and/or keywords, but rarely by search tasks on the full textual content nor by automatically extracted methodological study features.

Knowledge discovery with text mining

In a nutshell, text mining (TM) is the process of discovering and capturing knowledge or useful patterns from a large number of unstructured textual data (Zheng and Benoit 2019). It is an interdisciplinary field that draws on data mining, machine learning, natural language processing, statistics, and more. (Zheng and Benoit 2019).

Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction or the search for and modelling of hidden patterns (Hotho et al. 2005). A wide variety of biomedical text-mining tasks are currently being pursued, such as entity recognition (e.g. finding mentions of genes, proteins, diseases) and extraction of molecular relationships (e.g. protein-protein interactions) (Gerner et al. 2010).

The increasing calculation power of computers and database capabilities combined with the focused development of neural networks and machine learning algorithms make some contemporary systems that process natural language *'artificially intelligent'*. These systems can discover hidden patterns and relationships in huge amounts of text that a person could never read or summarize during their lifetime.

For most computational text analysis, full texts must be tokenized into smaller, more specific text features, such as words or word combinations (Welbers et al. 2017). Text preparation often involves sentence and word annotations, uniformization and dimension reduction like stemming or lemmatization. Further, text may be lowerized (converting capital to lower letters) and cleaned up by removing arbitrary letters and/or words, extra spaces, numbers and/or punctuations.

Recent years have seen rapid development in algorithm based text clustering, extraction and contextual analysis methods. To name just a few applied projects that recently made use of text mining methods on scientific content:

Watanabe (2021) developed a semi-supervised text analysis technique for discovering new research domains. Westergaard et al. (2018) showed that a contextual text analysis on full text sources is substantially outperforming and more informative than an analysis on abstracts only. Anderlucci et al. (2017) described the evolution of recent scientific research topics in statistics from 1970–2015 using a semi-supervised mixture model. Nuijten et al. (2016) developed the program *statcheck* to automatically detect potential errors in documents that report statistical result in APA style. Head et al. (2015) extracted reported p-values out of the PMC article collection and found substantial support for the presence of p-hacking. (Gerner et al. 2010) developed LINNAEUS, a species name identification system for biomedical literature.

This report draws a quantitative bibliometric profile of the PMC article collection by extracting metadata and text parts with *JATSdecoder*. The unified structure and letter representations may be useful for any text mining project on NISO-JATS coded content.

PubMed central and NISO-JATS

The PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) (PubMed-Central 2020) that does not act as a publisher itself. It allows full online access to more than three million documents that consist of abstracts, full text research papers and materials. PMC provides systematic download opportunities to access its full content or only subsets (the PMC OAI service and the PMC FTP service) that have a Creative Commons BY or similar license. The full bulk download service was disabled on 18th of March 2019 and the articles are now offered as Commercial- and Non-Commercial-Use bulk packages. All documents are provided as eXtensible Markup Language (NXML) or JavaScript Object Notation (JSON) files.

PMC's NXML files follow a well documented structure called NISO-JATS (National Center for Biotechnology Information 2014). It makes most article content and meta information easily accessible if used adequately by the editorial and/or technical implementation team. The NISO-JATS tag set is the most effective and widely used archival format for journal articles (PubMed-Central 2020). NISO-JATS is a flexible HTML coding system that defines 277 elements and their use. Though Comeau et al. (2018) point out that NISO-JATS is not designed to aid text mining, as the text appears at different levels and in various structures, the great benefit of its structure is demonstrated here by successfully extracting core article metadata and content with *JATSdecoder*.

The Content ExtRactor and MINEr *CERMINE* (Tkaczyk et al. 2015) is a sophisticated PDF to text conversion tool that extracts metadata, full text and parsed references from scientific articles and outputs a NISO-JATS coded XML file that is processable with *JATSdecoder*. *JATSdecoder* contains some correction procedures to deal with *CERMINE* specific and general PDF compilation problems (e.g. coding of special characters and operators).

JATSdecoder

The R package *JATSdecoder* (Böschen 2021) is an infrastructure software that contains a set of functions to extract and unify content from scientific literature coded in NISO-JATS format. Its function *JATSdecoder()* bundles all supplied functions to extract metadata, sectioned text and reference lists. Its function *study.character()* performs multiple text extraction tasks on specific methodological characteristics like statistical methods applied, α -error, correction methods and all reported statistical test results within the text. The p-values of extracted test statistics are recomputed to facilitate a manual check on consistency and reporting style (Böschen 2021b).

JATSdecoder makes use of regular expressions in simple search and text manipulation functions like: *grep()*, *gsub()*, *substr()*, *strsplit()*, *paste()*, *tolower()* to extract the content of the targeted tag. In standard mode, *JATSdecoder*'s function *letter.convert()* transforms and unifies all hexadecimal and many HTML letter representations to Unicode letters. Distinct hyphens, spaces, operators and many other special characters are unified to a character that is part of any Western computer keyboard to facilitate further text processing tasks. For example, more than 20 codings for distinct spaces are unified to a standard space. Spacing errors are corrected, hyperlinks and HTML formatting tags are removed. In standard mode, *JATSdecoder()* unifies name and country representations to facilitate post-processing and search tasks. Therefore, the *JATSdecoder* output is not an exact representation of the original content but easy to post-process and analyze.

The NISO-JATS tags <title>, <journal>, <type>, <abstract> are simple uniformly used metadata tags. Their content can be extracted easily with simple regular expressions. Tags that are less structured and more flexible to use demand more sophisticated processes. The use of the <subject> and <kwd> (keyword) tags differs in terms of their technical implementation (with/without sub-subjects) and summarizing precision. Within the <history> tag, contributors are comparatively free to choose the precision and coding of each files publication history. *JATSdecoder*'s function *get.history()* extracts the supplied date stamps and calculates the first publishing date and year of publication if possible. Its function *get.text()* exports the contained text sectionwise. Since text sectioning is implemented inconsistently, sections and subsections are not differentiated. The sectioned text can be returned as floating text or vector with sentences for every section. *get.text()*'s argument *'sectionsplit'* can be used to split the text into main sections by defining an individual pattern vector. Section names that contain the defined pattern/s are combined with all following sections until the next matching pattern appears within the section names.

Listing 1 displays an example *JATSdecoder()* output for the article by Blanca et al. (2018). Here the text is split into four sections by setting the argument 'sectionsplit' to *'c('intro', 'method', 'result', 'discussion')'*. *JATSdecoder*'s function *text2sentences()* is used to return the abstract as sentence vector, the main text is returned as floating text per section split.

Listing 1 Example output of extracted article content with JATSdecoder()

```
$file
[1] "./JATSdecoderEvaluation/JATSdecoder/PMC6300498.nxml"
$title
[1] "Current Practices in Data Analysis Procedures in Psychology: W..."
$author
[1] "Blanca, Mar\u00eda J." "Alarc\u00f3n, Rafael"  "Bono, Roser"
$affiliation
[1] "1, Department of Psychobiology and Behavioral Sciences Methodo..."
[2] "2, Department of Social Psychology and Quantitative Psychology..."
[3] "3, Institute of Neurosciences, University of Barcelona, Barcel..."
$journal
[1] "Frontiers in Psychology"
$volume
  vol fpage lpage
  "9"    NA    NA
$editor
[1] NA
$doi
[1] "https://doi.org/10.3389/fpsyg.2018.02558"
$history
  collection      received      accepted          epub       pubDate
"2018-NA-NA"  "2018-7-19" "2018-11-29" "2018-12-13" "2018-12-13"
     pubyear
      "2018"
$country
[1] "Spain"
$type
[1] "research-article"
$subject
[1] "Psychology"         "Original Research"
$keywords
[1] "data analysis procedures" "empirical research"
[3] "quantitative research"    "methodological review"
[5] "ANOVA"                    "regression analysis"
$abstract
 [1] "This paper analyzes current practices in psychology in the us..."
 [2] "We reviewed empirical research published recently in prominen..."
 [3] "The 288 papers reviewed used 663 different DAP."
 [4] "Experimental and correlational studies were the most prevalen..."
 [5] "Two-thirds of the papers reported a single study, although th..."
 [6] "The papers mainly used parametric tests for comparison and st..."
 [7] "Regarding the former, the most frequently used procedure was ..."
 [8] "A decline in the use of non-parametric analysis was observed ..."
 [9] "Relationships among variables were most commonly examined usi..."
[10] "There was also a decline in the use of stepwise regression an..."
[11] "Overall, the results show that recent empirical studies publi..."
$sections
[1] "Introduction"
[2] "Materials and Methods;; Data Sample;; Measures;; Procedure;; R..."
[3] "Research Methods and Number of Studies;; Number and Type of Da..."
[4] "Discussion;; Author Contributions;; Conflict of Interest State..."
$text
[1] "In order to answer a specific research question, researchers h..."
[2] "The focus of analysis was scientific journals whose aim and sc..."
[3] "Table 1 shows the frequency with which different research meth..."
[4] "The aim of this paper was to analyze current practices in psyc..."
$tables
[1] "<table frame=\"hsides\" rules=\"groups\" cellspacing=\"5\" cel..."
[2] "<table frame=\"hsides\" rules=\"groups\" cellspacing=\"5\" cel..."
[3] "<table frame=\"hsides\" rules=\"groups\" cellspacing=\"5\" cel..."
...
$captions
[1] "Frequency and percentage of the research methods used in the p..."
[2] "Frequency, percentage, mean (M), and standard deviation (SD) f..."
[3] "Frequency, percentage, mean (M), and standard deviation (SD) f..."
...
$references
 [1] "Akhtar, S.; Shah, S. W. A.; Rafiq, M.; Khan, A.; (2016). Rese..."
 [2] "Asare, A. K.; Yang, J.; Brashear Alejandro, T. G.; (2013). Th..."
 [3] "Bakker, M.; Wicherts, J. M.; (2011). The (mis)reporting of st..."
...
[50] "Zientek, L. R.; Capraro, M. M.; Capraro, R. M.; (2008). Repor..."
```

Method

PubMedCentral's ftp server (ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/) was used on 01.01.2021 to bulk download all available documents in NXML format. After unzipping all folders *JATSdecoder()* is used to extract the article elements title, journal, history, author, editor, type, subject, keywords, affiliation, country, section names, full text and reference list.

The relative frequency of successfully extracted content is reported for every tag and different time intervals to display the development of the availability and consistency in use of the extracted tags over time. Each meta tag is analyzed globally to further explore the full corpus and demonstrate new monitoring techniques enabled with *JATSdecoder*. Descriptive analyses of low dimensional tags are performed with frequency tables and bar plots. Content of higher dimensional tags (subject, keywords, author and editor names) is reduced to the most frequent labels and presented in word clouds (Fellows 2018). As the representation of information in text data is mostly unstructured, some inconsistencies and general burdens for search tasks and text analytical approaches on PMC's database are outlined.

An AMD@Epyc 7452 32-core processor with 256GB ram memory running with Linux Ubuntu 20.04.1 LTS is used. All tag extractions and analytical computations are performed with the open source statistic software R R Core Team (2020) and its package *JATSdecoder* (Böschen 2021). Multi-core processes are applied with the R package *future* (Bengtsson 2020) to speed up the extraction.

The syntax and extracted data to reproduce or update this analysis can be accessed at: https://github.com/ingmarboeschen/JATSdecoderEvaluation/.

Results

Summary and NISO-JATS tag use over time

As of January 1st, 2021, the PMC database contained 3,236,331 articles published by 16,789 journals with a total file size of 260.17 GB. Table 1 displays the overall frequency of successfully extracted tags and for some subsets of publishing periods. Tags that should contain text (e.g.: title, abstract, author) are treated as not present if the tag is detected but does not contain any content. Abstracts and text parts with less than 30 characters are treated as not present.

The title, journal, history, type and subject tag contain extractable content in nearly all the documents. Material that has been published since the year 2000 tends to have much higher amounts of extractable tags in general. Within content that was published before 2000 the affiliation and country tags are rarely used, whereas nowadays, they are present and consistently used in the majority of content. Keywords are only present in about half of the articles. In 2018 about two thirds of all articles supplied keywords additionally to their subject tag. The editor tag is rarely used and shows no trend towards omnipresence in recent releases either. With 508,834 cited documents, 2020 is the most frequent year of release.

Document type

The <type> tag is consistently used to specify the global article content with one categorizing label only. 12 articles do not contribute a <type> tag. Figure 1 displays the frequency distribution of the 60 discovered <type> tag specifications on a log transformed x scale.

Table 1 Relative frequency of extracted tag use over time

Content	Total use	[1781; 2000]	(2000; 2005]	(2005; 2010]	(2010; 2015]	(2015; 2020]
Abstract	87.9%	41.3%	71.7%	87.9%	88.9%	92.2%
Affiliation	94.8%	38.4%	89.2%	96.4%	98%	98.4%
Author	96.9%	56.2%	96.9%	98.9%	99.2%	99.2%
Country	82.9%	15.4%	74.7%	86.9%	88.4%	86.2%
Doi	100%	100%	100%	100%	100%	100%
Editor	12.6%	0.1%	1.7%	10.2%	18.2%	11.5%
History	100%	100%	100%	100%	100%	100%
Journal	100%	100%	100%	100%	100%	100%
Keywords	59.9%	8%	42.2%	35.1%	52.9%	72%
References	90.3%	9.7%	69.6%	92.2%	94.6%	96%
Sections	88.6%	8.9%	71.9%	91.3%	93.6%	93.6%
Subject	99.8%	100%	99.9%	99.7%	99.8%	99.8%
Text	88.5%	8.8%	71.9%	91.2%	93.5%	93.5%
Title	100%	99.9%	99.8%	100%	100%	100%
Type	100%	100%	100%	100%	100%	100%
Volume	48.7%	93.4%	82.4%	77.1%	54.9%	36.9%
Total *n*	3,235,065	171,394	52,488	239,952	924,154	1,846,979

The vast majority of publications is labeled as ’research article’ (72%, $N = 2{,}343{,}752$), followed by reviews (8%, $N = 250{,}770$) and case-reports (5%, $N = 152{,}891$). There are 42,362 articles declared as corrections (’corrected-article’, ’correction’, ’correction-forward’) and 1902 articles tagged as retraction (’retracted-article’, ’retraction’). Some type

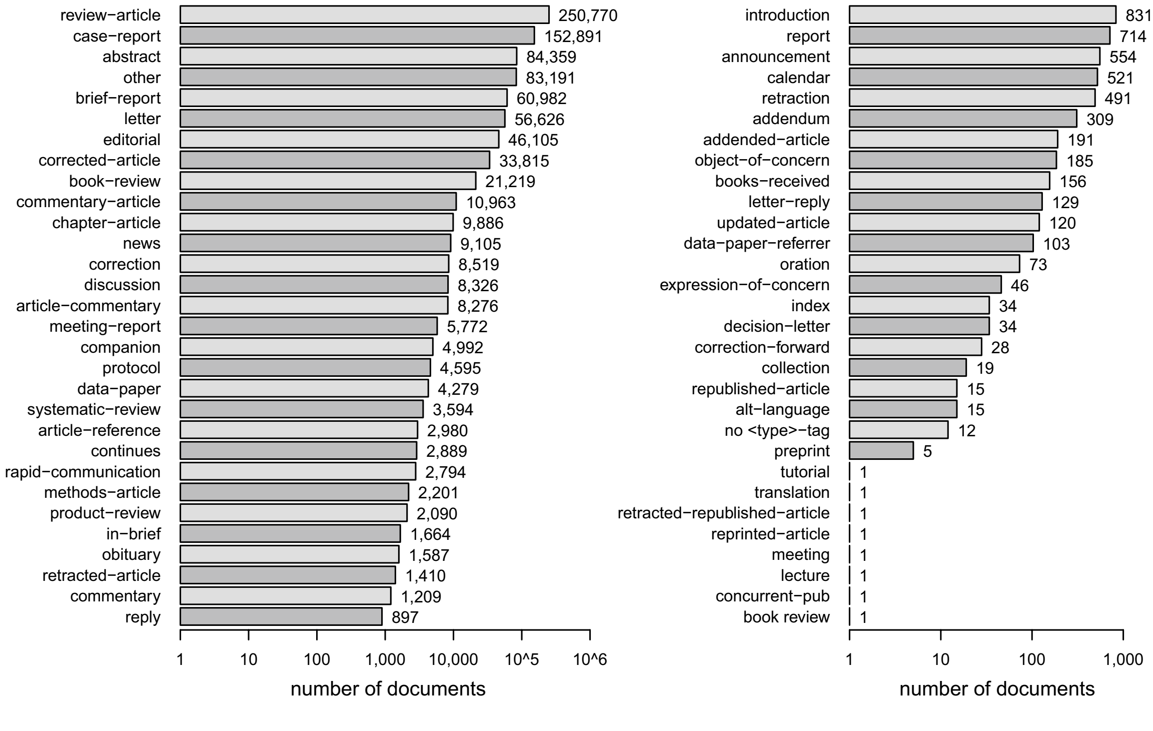


Fig. 1 Absolute frequency distribution of article types

Table 2 Absolute (h) and relative (f) frequency of journals supplying n articles

n	1	2–10	11–100	101–1001	1001–10^4	10,001–10^5	> 10^5	Sum
$h(n)$	5,078	5,178	3,553	2,461	486	31	2	16,789
$f(n)$	.302	.308	.212	.147	.029	.002	.000	1

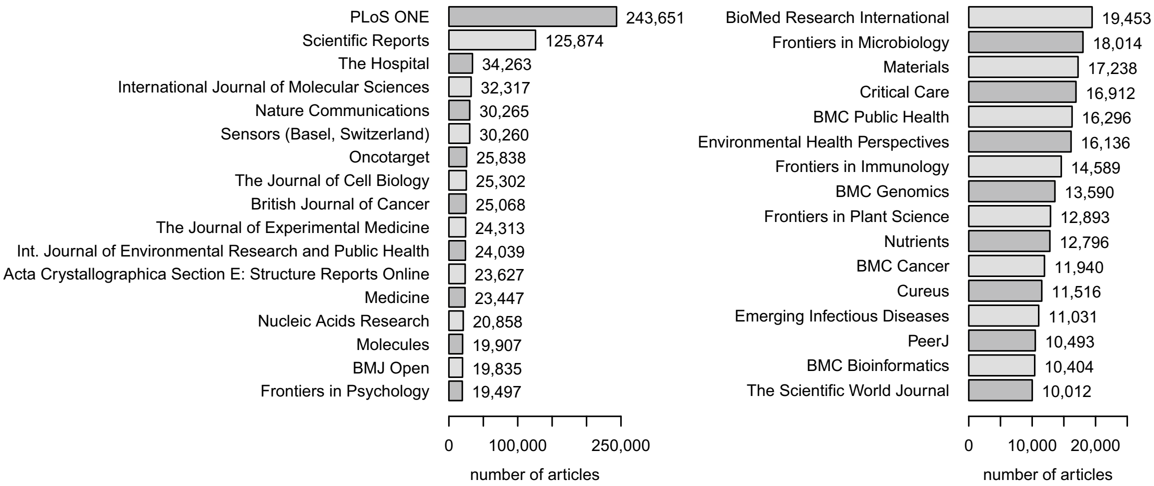


Fig. 2 Total number of published articles for journals with >10,000 articles

categories overlap. An inspection of the title tag reveals that 20,694 documents tagged as ’research-article’ have a title containing the search terms ’systematic review’ or ’literature review’. Further, 3049 documents type tagged as ’research-article’ contain the search term ’review’ inside the subject and 3987 inside the keyword tag, with 258 articles containing ’review’ in both tags.

Involved journals

The <journal> tag is used in all non-empty files and is easy to extract. Table 2 shows the absolute and relative frequency of the number of provided articles per journal. Out of 16,789 journals, 61.1% provide only 10 or fewer documents. There are 33 journals (0.2%) that provide more than 10,000 documents each. Taken together, these 33 journals contribute 971,674 articles (30%) to PMC.

Figure 2 displays the absolute number of documents by those journals that are identified more than 10,000 times. *PLoS ONE* is the most prominent publisher, followed by *Scientific Reports* and *The Hospital*.

Publication history

The publication history is usually stored within a <date> tag inside the <history> tag. A pre analysis revealed that 14 different time stamp labels are being used. *JATSdecoder()* extracts the earliest publishing date (’pubDate’) and year (’pubyear’) out of six possible publication date stamps to facilitate selection procedures.

Publishers support different and inconsistently precise information about the article's publication history. Only 1266 documents do not contain extractable date stamps. Table 3 displays the frequency of use and captured date range for each of the identified date types and standardized publishing dates, created by *JATSdecoder()*. Missing <month> and <day> tags within a history date stamp were set to 1.

The oldest publications listed are 140 books and articles published in 1781, which are distributed by the *The London Medical Journal* as scanned PDF documents.

Since then, documents from every year are available. About 82.2% of all articles were published after January 1st 2000, when the boom of the open access movement started. The 'retracted' and 'submitted' date stamps are used very rarely.

Table 3 indicates that there are obvious errors in date stamps of some documents, as they contain future dates. Further, the calculated publication date, the accept- and receive date can serve for another analysis of errors in date stamps. Calculating the time to accept (accept date–receive date) and time to publish (publishing date–accept date) reveals that there are 2,310 documents with an accept date prior to their receive date and 12,303 articles with a publishing date prior to their accept date.

Contributor

The <contrib> tag is used to store one or multiple <surname> and <name> tags of author and or editor name/s, as well as involved affiliations and contact addresses.

The representation of given names differs across journals. Some journals provide given and surnames in fully capital letters, others use one letter abbreviations of all or only given middle names. Some names with an Asian origin are provided with Latin letters, others with Asian characters. *JATSdecoder*'s functions that extract the author/s and

Table 3 Frequencies of use and time spans of extracted history date stamps

Date type	Relative use (%)	Earliest date	Latest date
Accepted	77.9	1882-3-15	2100-1-14
Collection	62.7	1973-10-NA	2021-NA-NA
Ecorrected	0.1	2010-10-04	2020-9-8
Epreprint	0.6	2005-10-10	2020-NA-NA
Epub	85.5	1957-10-01	2021-1-1
Nihms_submitted	0.9	2006-2-23	2020-9-9
pmc-release	27.5	1799-11-NA	2021-9-21
ppub	36.1	1781-1-NA	2022-NA-NA
Print	0.7	1989-NA-NA	2020-NA-NA
pub	1.1	2009-4-01	2020-NA-NA
Received	78.7	1856-1-19	2097-10-22
Retracted	0	2018-6-01	2018-8-27
Rev-recd	18.6	1800-1-16	2088-10-30
Submitted	0	2019-10-29	2020-9-7
PubDate*	90.4	1802-08-01	2021-11-01
Pubyear*	100	1781	2021

*Computed by *JATSdecoder()*

editor/s name convert fully capital letter names to a representation with an initial capital letter only and enable shortening of given names to a one letter abbreviation if desired.

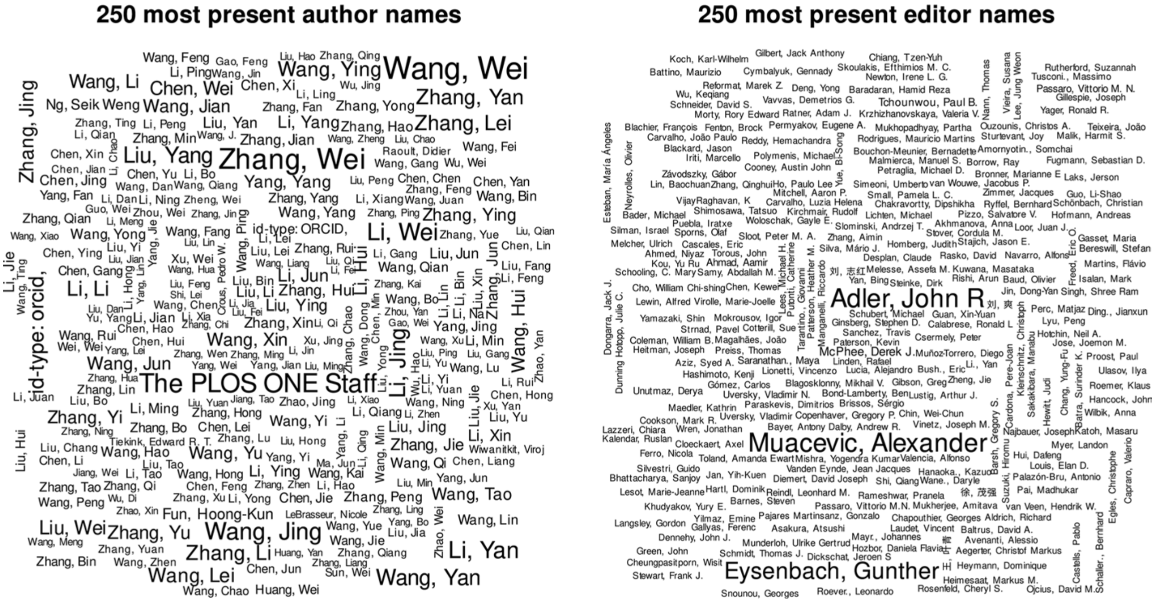
Authors

It is widely known that a considerable proportion of authors share the same last name and first initial (Falagas 2006) which is a big problem for correctly identifying individuals and networks in PMCs database. Besides that, the use of the Western-driven (surname/family name|given/first name|middle initial) system is particularly problematic for Asian biomedical researchers in general: Japan, China and especially Korea, where only a few surnames predominate and middle names often do not exist (Harrison and Harrison 2016).

This phenomenon becomes obvious in Fig. 3 which displays the 250 most frequent author names in PMC. It is obvious that most of the names have an Asian origin. First and last names are pretty short and the same names occur quite often, resulting in a lot of name siblings. All displayed author names appear at least 342 times with a maximum of 3,643 detections of 'Wang, Wei', the most present author name in PMC documents.

Several registers that supply a nomenclature for a clear author distinction exist. Harrison and Harrison (2016) have highly recommended their use. A comparably small but steadily increasing amount of articles in PMC supplies an author identification number for at least one author. 9.8% of all articles contain at least one ID coded author name, and out of these 20% have ID tagged at least half of the contributing authors. The *Open Researcher and Contributor ID* (ORCID) is the far most commonly identified identifier with 696,416 distinct IDs. 186 authors are listed with n *sciprofiles* ID, 53 with a *zoobank* ID, 10 authors are coded with their *twitter* names. In 2020 43.8% of all documents had at least one author ID tagged. Before 2009 no document contained any ID tagged author.

For a long period of time, it was quite usual to publish in very small groups or alone. But the amount and intensity of author collaborations has been growing steadily for decades. In 2020 the median of author group size grew up to 5 persons, meaning that more than 50% of content is published by researcher groups equal to or larger than 5. There are 547 articles with more than 500 authors. A manual inspection of the titles and types of


250 most present author names

250 most present editor names

Fig. 3 Wordclouds of 250 most present author and editor names

these articles reveals that there are mostly two categories of content with such big author groups. 282 (51.55%) of these articles are published by international collaborations covering topics like ’atlas’, ’detector’, ’proton’, ’collision’, ’jet’ and/or ’particle’. The remaining documents with such large author groups are mostly abstracts and poster collections from congresses or meetings.

Editors

The editor information is the least often identified meta tag. 407,359 (12.59%) of all documents contain an extractable editor name. Only 19 of all 42,455 unique editor names are represented by an author identification number. Compared to the contributing author names, most editor names have a Western-driven origin and mostly refer to the same person. The problem of the indistinguishability of editor names will become worse as soon as more journals with Asian named editors provide content to PMC.

91.9% of all documents that supply editor information were edited by one person, 6.3% list two editors. One comparably large group of involved editors was identified in the German book about infectious diseases ’Infektionserkrankungen’ (Bialek et al. 2007) with 121 listed editors.

Affiliation

The <aff> tag is usually stored within the <contrib> tag and used to supply information about the affiliations of contributing author/s. The data is stored with differing precision and technical implementation. Some documents supply the full name of a department within an institution, an abbreviation as well as full contact details, while in others, the main institution name is supplied with a country location. Some distributors use tags to locate special elements of affiliation details (name, address line, country) while others supply this data in an untagged line. Further, the naming of one and the same institution is not completely coherent (e.g.: ’University of Oxford’ and ’Oxford University’). Therefore, a global analysis of the affiliation tag is not reported here.

Since the storing of detailed affiliation data differs widely, *JATSdecoder()* removes all HTML-tags within the <aff> tag and outputs a string with comma separated specifications. Any analysis of the affiliation tag should involve a pre-targeting of specific institutions and further post-processes like uniformization of name representations and removing redundant content like postal codes.

Subjects and keywords

The <subject> tag is used within an <article-category> tag or stands alone to describe the document content or type. Without any uniformization 66,118 distinct subject labels are identified. About two-thirds of all documents contain one classifying element, which is mostly similar to the <type> tag (e.g.: type: ’research-article’, subject: ’Research Article’) but partially further specifies the document’s type. Some contributors make intense use of the subject specification. 0.5% of documents contain more than 38 subjects, with a maximum of 739 declared subjects’.

The <kwd> tag is explicitly designed to contain keywords that briefly describe the article’s content. Within the last two decades, keywords have developed from a rarely to very often used specifier of documents. As Table 1 indicates, keywords are only used in about half of all documents. In 2020 the relative frequency of keyword-tagged-documents peaks at 78.4%. An inspection of overly intense keyword uses reveals that some documents contain the abstract's text inside the <kwd> tag. Figure 4 displays the 250 most frequently extracted subject and keyword labels in separate word clouds. Bigger words indicate a higher frequency of use.

The most urgent and influential research topic in 2020 surely is the Covid-19/SARS-Cov-2 pandemia (N = 31,549). Within only one year, it has become the most common keyword specified within all PMC documents.

Country of origin

The <country> tag is used to store the involved author/s and/or affiliation/s country of origin. If no tag is used to store this information, *JATSdecoder()* performs a dictionary search at the end of each extracted affiliation address and returns a vector with uniquely identified country names. Some country namings that are different to the standard country name (e.g.: Peoples Rep. of China, UNITED STATES, UK) are unified by a manually generated list of pre-identified country namings to facilitate any post-processing on world maps. Overall, 208 out of 243 countries of origin have been detected. 82.9% of all documents contain at least one extractable country of origin. 30.8% of these documents have been released by international author groups, as they contain more than one extractable country of origin.

Figure 5 displays the development of extracted country of origin over time. Most documents that supply information of at least one country in the <contrib> tag are of US American origin. Although China has been a relatively small contributor till 1999, content distribution rapidly increased between 2000-2009. Today, China represents the second most frequently detected country of origin.

Besides the total number of countries supplying content to PMC the network structure of affiliations and countries has increased steadily. Today's scientific community is heavily connected. Besides being the most prevalent publishers, US-American affiliations most often appear in papers by international collaborations. Most frequent country cooperations

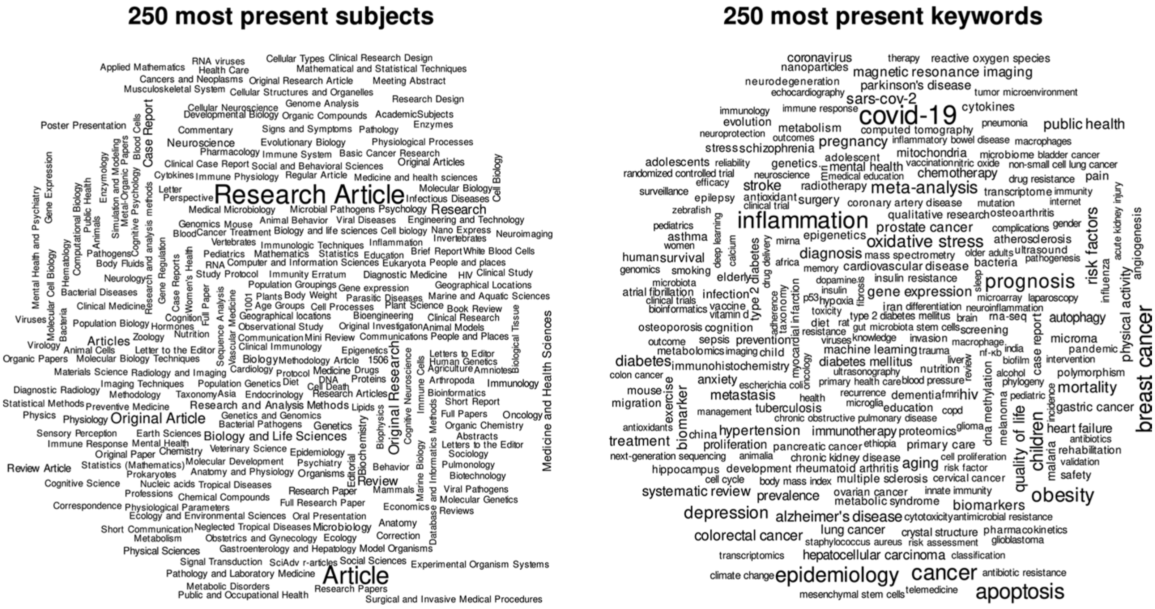


Fig. 4 Wordclouds of 250 most frequent <subject> and <kwd> tags

were found for the USA and the United Kingdom, the USA and China, as well as the USA and Germany.

Special character and character representation

A widely used PMC standard to code special characters is the hexadecimal system (e.g.: ’&#x003b1;’ for ’α’). Another common character coding system is the Unicode system (e.g.: ’\u03b1;’ for ’α’). Both coding systems contain representations of characters from nearly any alphabet, mathematical characters and symbols, as well as text formatting characters.

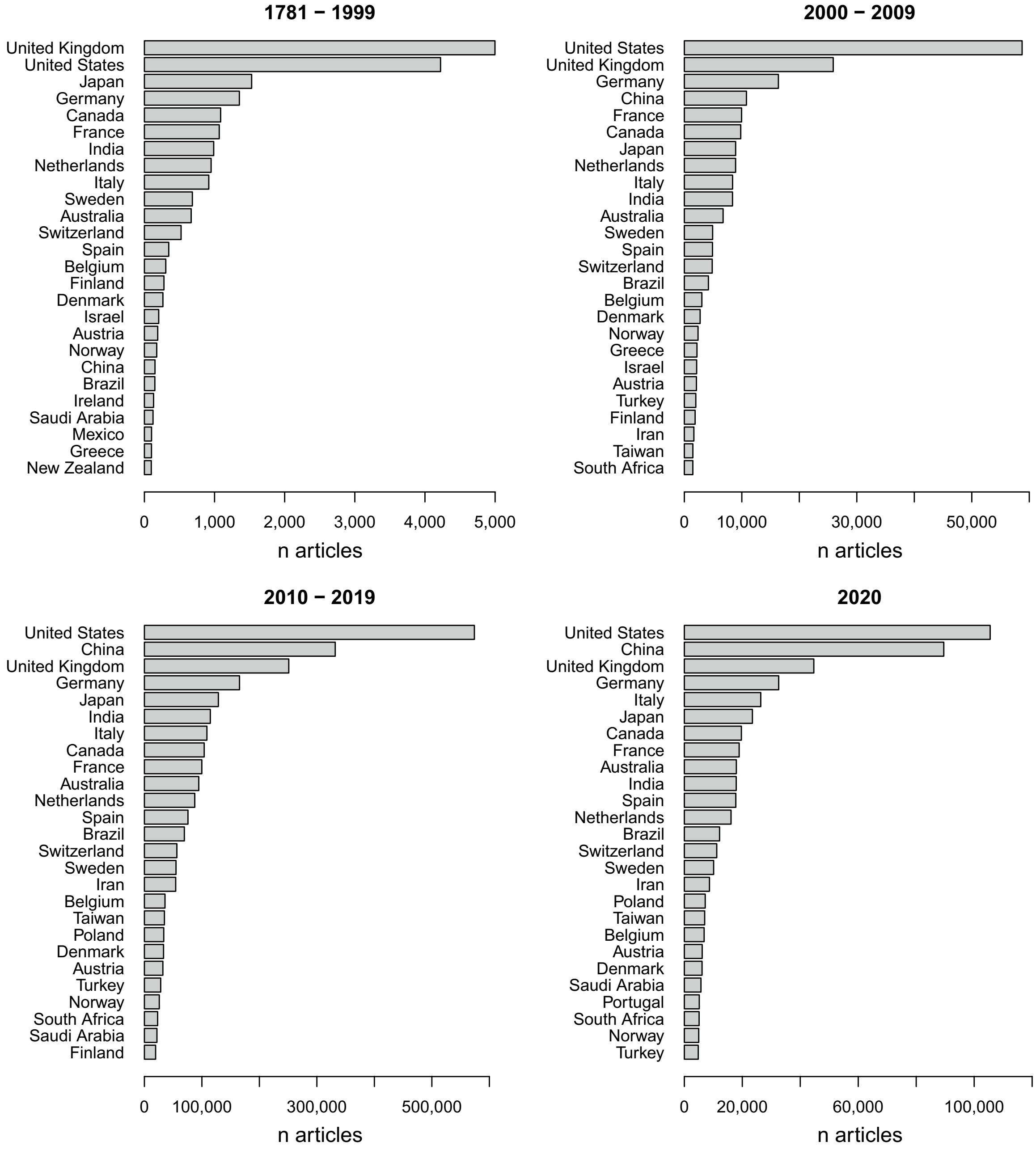


Fig. 5 Change in top 25 country involvement over time

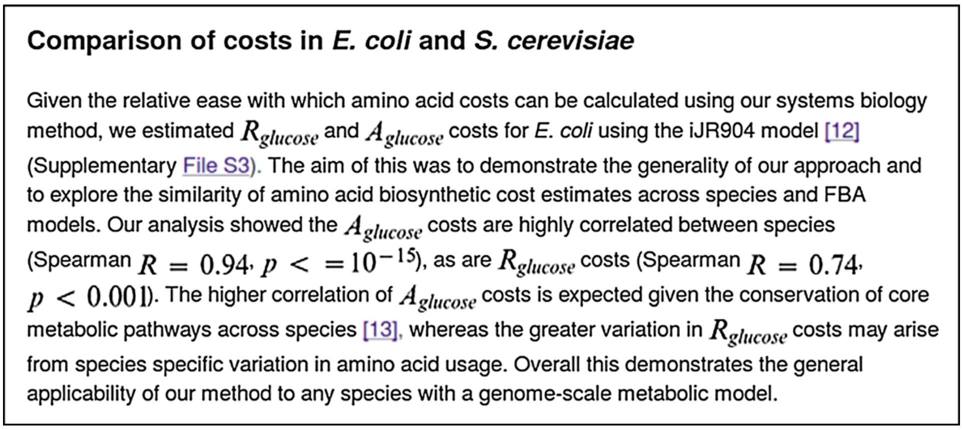
Two technically clean approaches exist to display mathematical content in nearly any browser. The Mathematical Markup Language (Ausbrooks et al. 2014) and the MathJax Tex library (The MathJax Consortium 2018) are widely used and compatible methods. MathML is made to enable mathematics to be served, received, and processed on the World Wide Web, just as HTML has enabled this functionality for text (Ausbrooks et al. 2014). MathJax is an open-source JavaScript display engine for LaTeX, MathML, and AsciiMath notation that works in all modern browsers (The MathJax Consortium 2018). Within PMC, MathML and MathJax formulas are used with hexadecimal coded special letters and LaTeX annotation.

JATSdecoder transforms MathML and MathJax formulas into a plain text representation and unifies many pre-detected synonyms of character representations (bold, cursive and other equivalents of the same character) to simple Unicode letter representations.

Some documents contain an unfavorable approach to store special characters, mathematical content like formulas and/or statistical results within text. A hyper-referenced picture is included with an <inline-formula> tag. When using a browser, this representation becomes obvious by zooming into the page. An example is given in Fig. 6. The bold pixelated text parts are represented as pictures that do not resize into line height when increasing the view size (source: Barton et al. 2010. Evolutionary systems biology of amino acid biosynthetic cost in yeast. PloS One, 5(8)).

Discussion

The PubMedCentral database is a rich and valuable resource for contextual text mining approaches on scientific content. A huge amount of scientific articles and materials from different areas of research is easily accessible in a rather consistent, pre-structured form. Although explicitly not designed to aid text mining, the NISO-JATS system enables a valid extraction of metadata and text parts with *JATSdecoder*. *JATSdecoder* copes with a wide range of possible implementations of NISO-JATS and facilitates customizable mirroring of scientific publication processes. It supports text analytical tasks on full documents or specific text parts or metadata only. Still, all results reported here rely on the assumption, that every way of tagging the targeted content was taken into consideration and correctly handled.



Comparison of costs in *E. coli* and *S. cerevisiae*

Given the relative ease with which amino acid costs can be calculated using our systems biology method, we estimated $R_{glucose}$ and $A_{glucose}$ costs for *E. coli* using the iJR904 model [12] (Supplementary File S3). The aim of this was to demonstrate the generality of our approach and to explore the similarity of amino acid biosynthetic cost estimates across species and FBA models. Our analysis showed the $A_{glucose}$ costs are highly correlated between species (Spearman $R = 0.94$, $p < =10^{-15}$), as are $R_{glucose}$ costs (Spearman $R = 0.74$, $p < 0.001$). The higher correlation of $A_{glucose}$ costs is expected given the conservation of core metabolic pathways across species [13], whereas the greater variation in $R_{glucose}$ costs may arise from species specific variation in amino acid usage. Overall this demonstrates the general applicability of our method to any species with a genome-scale metabolic model.

Fig. 6 Example of <inline-formula> use with hyper referenced pixeled pictures from Barton et al. (2010)

Although most tags are used quite consistently nowadays, any text analytical research on PMC's content has to face several challenges and limitations, depending on the topic of interest. Some tags are not being used consistently by all distributors, some change over time (e.g. retracted documents), some are set incorrectly, probably due to human error.

The full PMC text corpus is fluid and increasingly up to date, but also provides historic documents that are being made available by many distributors. Some documents received a post-processing when their status changed (retraction/correction). The document collection is highly selective for the biomedical and health sciences, as it only contains content with an open access license. It also contains some documents from other research areas (e.g. physics). Further, most of the detected journals only provide very few documents. About half of all journals that provide content to the database only link to 10 or fewer files. An analysis including the content of these journals should always consider this fact.

Most of the journals that publish with a high frequency and their editors obviously have a Western origin. The frequent absence of an author identification number is still a crucial element for analyses on publications by a specific author or network. Before the U.S. National Institutes of Health (NIH) and its National Library of Medicine (NLM) launched the modern PubMed system, the math, physics and computer science community already solved this problem with the creation of 'arXiv' in the early 1990s (Harrison and Harrison 2016). Harrison and Harrison 's claim: *"The time has come for 'DOIs for authors"* ought to be implemented for authors and editors as soon as possible. The probability that more than one author bears the same name inside the scientific community or even the same subdiscipline grows every day. A retrospective discrimination of author and editor identity will become an increasingly difficult task.

Neither keyword nor subject tag supply uniform information about the topic relatedness of a document. Researchers working with article subsets should consider various tags as in- and exclusion criteria for their selection. The selection of articles covering a certain topic should be performed by a combined search task on title, abstract, keywords and/or the subject tag.

The implementation of a general topic extracting algorithm, that accumulates information from title, subject keyword and/or abstract and text, to select content, would highly facilitate selection. It could be of future interest, if 'topic models' performed on <subject> and <kwd> tag lead to comparable results if generated from title, abstract and/or full text. Such algorithms would enable an easy visualization of the evolution of certain research topics over time.

The presentation of statistical results as images is a struggling task for automated text analysis on numerical content and should be avoided by distributors. None of the research using text analysis on scientific online literature has considered this issue. There are several open source optical character recognition (OCR) software packages available (e.g. tesseract), which may serve for a pixel to text conversion for pixeled content. The precision of such conversion could be content of future research, but should never be expected to be perfect. Simple reports on statistical parameters or results (e.g.: R = .74 or p<0.001) should be quite well extractable with OCR. Their ability to recognize complex formulas will be limited until trained adequately.

The steady increase of the amount of research findings published every year complicates the discrimination of valid and robust results from weak and questionable evidence. Text mining offers some great opportunities for selection processes and may boost a self-correcting culture in science.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was financed by a doctoral grant awarded by the Department of Psychological Methods and Statistics, Institute of Psychology, University Hamburg, Germany.

Data availability *JATSdecoder* software is freely available at: https://github.com/ingmarboeschen/JATSdecoder Skripts to reproduce this and other analyses performed with JATSdecoder are stored at: https://github.com/ingmarboeschen/JATSdecoderEvaluation.

Declarations

Conflict of interest The author declares no conflict of interest.

References

Anderlucci, L., Montanari, A., and Viroli, C. (2017). The importance of being clustered: Uncluttering the trends of statistics from 1970 to 2015. arXiv:1709.03563.

Ausbrooks, R., Buswell, S., Carlisle, D., Chavchanidze, G., Dalmas, S., Devitt, S., Diaz, A., Dooley, S., Hunter, R., Ion, P., Kohlhase, M., Lazrek, A., Libbrecht, P., Miller, B., Miner, R. d., Rowley, C., Sargent, M., Smith, B., Soiffer, N., Sutor, R., and Watt, S. (2014). Mathematical Markup Language (MathML) Version 3.0 3rd Edition. https://www.w3.org/Math/draft-spec/Overview.xml. W3C Working Draft 21 November 2014.

Barton, M. D., Delneri, D., Oliver, S. G., Rattray, M., & Bergman, C. M. (2010). Evolutionary systems biology of amino acid biosynthetic cost in yeast. *PloS One, 5*(8), e11935. https://doi.org/10.1371/journal.pone.0011935

Bengtsson, H. (2020). Future.apply: Apply function to elements in parallel using futures. R package version 1.4.0.

Bialek, R., Groll, A., Heininger, U., & Schuster, V. (2007). Infektionserkrankungen. *Therapie in der Kinder- und Jugendmedizin*. https://doi.org/10.1016/B978-343723200-8.50020-2

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.02558

Böschen, I. (2021a). JATSdecoder: A metadata and text extraction and manipulation tool set for the statistical programming language R. www.github.com/ingmarboeschen/JATSdecoder.

Böschen, I. (2021b). Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-98782-3

Comeau, D. C., Wei, C.-H., Doğan, R. I., and Lu, Z. (2018). PMC text mining subset in BioC: 2.3 million full text articles and growing. arXiv:1804.05957.

Falagas, M. E. (2006). Unique author identification number in scientific databases: a suggestion. *PLoS Medicine, 3*(5), e249. https://doi.org/10.1371/journal.pmed.0030249

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1–54. https://doi.org/10.18637/jss.v025.i05

Fellows, I. (2018). wordcloud: Word clouds. *R Package Version, 2*, 6.

Gerner, M., Nenadic, G., & Bergman, C. M. (2010). LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics, 11*, 85. https://doi.org/10.1186/1471-2105-11-85

Harrison, A. M., & Harrison, A. M. (2016). Necessary but not sufficient: Unique author identifiers. *BMJ Innovations, 2*(4), 141–143. https://doi.org/10.1136/bmjinnov-2016-000135

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. https://doi.org/10.1371/journal.pbio.1002106.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). *A brief survey of text mining*, volume 20, pages 19–62. Citeseer.

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM) (2014). Journal Publishing Tag Library—NISO JATS Draft Version 1.1d2. https://jats.nlm.nih.gov/publishing/tag-library/1.1d2/index.html.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

PubMed-Central (2020). PMC Overview. https://www.ncbi.nlm.nih.gov/pmc/about/intro. Retrieved: 2020-02-25.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Science-Metrix Inc. (2018). Analytical support for bibliometrics indicators: Open access availability of scientific publications. http://www.science-metrix.com/sites/default/files/science-metrix/publications/science-metrix_open_access_availability_scientific_publications_report.pdf.

The MathJax Consortium (2018).What is MathJax? Retrieved 25, Feb 2020 from http://docs.mathjax.org/en/latest.

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR), 18*(4), 317–335. https://doi.org/10.1007/s10032-015-0249-8

Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures, 15*(2), 81–102. https://doi.org/10.1080/19312458.2020.1832976

Welbers, K., Atteveldt, W. V., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11*(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J., & Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Computational Biology, 14*(2), e1005962. https://doi.org/10.1371/journal.pcbi.1005962

Zheng, D., & Benoit, K. (2019). Text mining for drug discovery. *Methods in Molecular Biology, 1939*, 231–252. https://doi.org/10.1007/978-1-4939-9089-4_13

Anhang B: Publikation 2

Böschen, I. (2021). Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. Scientific Reports 11, 19525. https://doi.org/10.1038/s41598-021-98782-3

scientific reports

OPEN

Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports

Ingmar Böschen

The extraction of statistical results in scientific reports is beneficial for checking studies on plausibility and reliability. The R package *JATSdecoder* supports the application of text mining approaches to scientific reports. Its function *get.stats()* extracts all reported statistical results from text and recomputes *p* values for most standard test results. The output can be reduced to results with checkable or computable *p* values only. In this article, *get.stats()*'s ability to extract, recompute and check statistical results is compared to that of *statcheck*, which is an already established tool. A manually coded data set, containing the number of statistically significant results in 49 articles, serves as an initial indicator for *get.stats()*'s and *statcheck*'s differing detection rates for statistical results. Further 13,531 PDF files by 10 mayor psychological journals, 18,744 XML documents by *Frontiers of Psychology* and 23,730 articles related to psychological research and published by *PLoS One* are scanned for statistical results with both algorithms. *get.stats()* almost replicates the manually extracted number of significant results in 49 PDF articles. *get.stats()* outperforms the *statcheck* functions in identifying statistical results in every included journal and input format. Furthermore, the raw results extracted by *get.stats()* increase *statcheck*'s detection rate. *JATSdecoder*'s function *get.stats()* is a highly general and reliable tool to extract statistical results from text. It copes with a wide range of textual representations of statistical standard results and recomputes *p* values for two- and one-sided tests. It facilitates manual and automated checks on consistency and completeness of the reported results within a manuscript.

The technical revolution goes along with a steady increase in the total number of yearly published scientific articles. Computers have become incredibly fast and enable us to deal with huge amounts of textual data, which has never been easier to preselect, access, store and process before. The PubMedCentral database[1] alone stores more than 3 million open access documents related to the biology and health sciences.

Along with the publication boom, several scientists have expressed their doubts about the robustness of many scientific results carried out[2–6]. The Open Science Collaboration[7] tried to replicate 100 experiments with a psychological background and could only replicate between 23 and 63% of the original findings, depending on the subject and definition of a successful replication. This result led to the so-called replication or reproducibility crisis in psychology.

Besides the many problems arising with overly small sample sizes[2,6,8], psychological research is often based on 'WEIRD' selective samples[9] and standardized statistical test procedures like uninformed nil-null-hypothesis testing[3,10] with an α-error probability of .05. Uninformed nil-null-hypothesis testing refers to statistical procedures that are applied on empirical data to test null-hypotheses of no correlation, zero difference or no effect with an undirected test.

A crucial, paradoxical difference between theory testing in physics and psychology is emphasized by Meehl[3]:

> 'In the physical sciences, the usual result of an improvement in experimental design, instrumentation, or numerical mass of data, is to increase the difficulty of the 'observational hurdle' which the physical theory of interest must successfully surmount; whereas, in psychology and some of the allied behavior sciences, the usual effect of such improvement in experimental precision is to provide an easier hurdle for the theory to surmount.'

Institute of Psychology, Psychological Methods and Statistics, University Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. email: ingmar.boeschen@uni-hamburg.de

Using nil-null-hypothesis tests, researchers that seek for significant results can apply several questionable research practices like optional stopping, multiple testing or postdiction[11,12] to increase the possibility of a false positive result.

Journal editors and readers usually seek for new, sensational results rather than replications reporting differing or supporting results to the original findings. Also, many researchers face a scoring system that values the pure quantity and impact of published articles more than the reliability and robustness of their findings. This surrounding may lead to a lot of flawed results and endangers scientific credibility. John et al.[12] surveyed over 2,000 psychologists about their involvement in questionable research practices and found that the percentage of respondents who committed to have engaged in at least one questionable practice was surprisingly high (up to 78% in a self report).

Within the highlighted world, an incredibly large amount of research findings is obviously contaminated with many spurious findings and errors. Still, corrections are published rather rarely and errors are preserved in the literature. Therefore, it can be quite beneficial for scientists, authors, reviewers, editors and/or search engines to summarize a scientific report in terms of its text parts and main study characteristics.

There are several tools and techniques to check specific statistical results on plausibility. The GRIM test[13] is an easily performed calculation to check a mean of a Likert scale on plausibility, if the given sample size is known. In their analysis, Brown et al.[13] found 36 out of 71 articles (51%) that reported at least one inconsistent mean. An automated tool to check p values in reports of various statistical test results is *statcheck*[14] which is described in detail later. Nuijten et al.[15] found that half of all published psychology papers that use null-hypothesis significance testing contained at least one p value that is inconsistent with its test statistic and degrees of freedom.

Besides the practical use of such checking procedures, implausible results should not be considered as a proof for a statistical error or a corrupt report and always be analyzed case by case. On the other hand, a result that passes a plausibility check does not directly imply an adequate, objective or even correct contextual decision, especially in terms of causality and generalizability. No algorithm can replace an informed expert, evaluating the whole study design, sampling methods, sample characteristics, operationalization and adequateness of the statistical procedure applied, to decide if the conclusions made are plausible, correct, or even valid. Still, an automated extraction of the reported statistical results is the key element to perform a check on completeness and plausibility quickly. An automated identification of studies that use certain statistical methods/measures or sample sizes can be very helpful for selection processes in meta-analyses and systematical reviews.

Previously to a detailed description and comparison of the algorithms by an evaluation on varying input formats, a terminology for distinct representations of statistical results is introduced.

Terminology for destinct representations of statistical results reported in text.

Generally, any letter or letter-number combination pointing to a numeric value with an operator ($<, >, =, \leq, \geq$) is here considered to be a potential statistical result. A statistical result can be a descriptive measure as well as a test result. Statistical test results mostly consist of a varying set of results (test statistic, degree/s of freedom, an effect measure, p value, confidence interval and/or a Bayes Factor). There are many widely used statistical tests. Results that contain a Z-, t-, F-, χ^2, r-, H-, Q-, G^2, U-statistic or Bayes Factor and/or a measure of effect (β, Cohen's d, η^2, OR, RR, R^2) are defined as statistical standard results here.

Although there are guidelines about how to report statistical results (e.g. APA style), they are not reported in this standardized manner consequently (e.g. p value only). It makes sense to differentiate between different reporting practices of test results in terms of the level of their completeness and post processability. A statistical test result that enables a recomputation of an also reported p value is defined as checkable here (e.g.: '$t(89) = 1.96, p = .05$'). Test results that enable a computation of a non-reported p value (e.g.: '$t(89) = 1.96$') will be called computable. A checkable result is always computable. The third set of test results is reported in a manner, in which no recomputation of a reported or unreported p value is possible (e.g.: '$r = .12, p < .05$', or '$t > 2$'). These results will be called uncomputable here.

The R Package *JATSdecoder*.

The R package *JATSdecoder*[16] supports the application of text mining approaches to scientific reports by processing XML documents that are structured with the Journal Archiving Tag System NISO-JATS[17]. The NISO-JATS is an HTML tag standard to store scientific articles without any layout parameters. Graphical content is hyper referenced.

JATSdecoder's functions make use of simple and sophisticated text extraction and manipulating algorithms that can cope with a wide range of textual and technical representations of content in NISO-JATS coded documents. The built-in function *JATSdecoder()* extracts a set of metadata (title, author, publishing dates etc.), the abstract, sectioned text and reference list. The structured output is very useful for individual searches and extraction procedures, as it facilitates these tasks on individually defined text parts (e.g. section titles, method section, reference list) and metadata.

JATSdecoder's function *study.character()* performs multiple text selection and manipulation tasks on the list created by *JATSdecoder()* and extracts key study characteristics like number of reported studies, the statistical methods, software and correction procedures for multiple testing used. Its function *get.stats()* outputs all detected statistical results including descriptive measures (mean, sd, CI, Cronbach's alpha) as a vector which is then further processed. Detected Z-, t-, F-, χ^2, r-, H-, Q-,G^2, U-statistics or Bayes Factors and corresponding effect measures R^2, β, OR, d and/or η^2 are formatted into a data frame with numerical values and operators stored in separated columns.

To increase *get.stats()*'s detection rate for computable and checkable results, users can activate its arguments '*T2t*' and/or '*R2r*'. Statistics denoted with capital letter T or R respectively will then be treated as t- or r-values, which may not be appropriate. Activating its argument '*estimateZ*' makes *get.stats()* estimate *Z*-statistics for beta- and d-values reported with standard error but no test statistic.

If possible, a recomputed *p* value for an undirected null hypothesis is added. If desired, also *p* values for directed tests can be outputted if a computation is possible (only *t*-, *Z*-, *r*-values). The resulting data frame can be reduced to computable results only (recalculation of *p* value is possible, e.g. '$r(18) = .12$'), checkable results (recomputable result with *p* value), or outputted with all detected standard results (e.g.: no *p* value check possible: '$r = .12, p = .61$' or *p* value only).

Deviations in reported and recomputed *p* values may be multicausal (directed test, rounding, typo, extraction or compilation error). Therefore, a check for completeness and plausibility of the results is not done automatically. Users checking a manuscript should always manually countercheck the extracted results and inconsistencies.

A non-computed, although expected to be computed *p* value, a non-reported but computed *p* value, or a completely missed out result in the output of detected standard results may be an indicator for an incompletely reported result within the text. Warning messages are returned if *p*-, *r*-, or R^2-values are reported that are outside their valid range.

Statistical results reported in tables are explicitly not captured by *get.stats()*, as the compilation of tables cannot be performed reliably. Here *statcheck* differs from *study.character()*, as it always analyses the whole textual content of an HTML or PDF to text converted file and captures test results from tables, if they are reported in a full textual manner and not with named columns, which is much more frequently done in practice.

To extract the statistical methods mentioned in an article, *study.character()* tries to split the NISO-JATS document into four sections (introduction, method, results, discussion). Its function *get.method()* performs a heuristic driven feature extraction process, to output the statistical methods listed in the method and results section. It finds the specification of a method, that contains the descriptive term in front of a set of search terms, which most commonly used statistical procedures have in common (e.g.: test, regression, anova, method, theorem, interval, algorithm, etc.). Users can enlarge the result space by defining additive search words in its argument '*add*'. The current heuristic enables an extraction of new, still unknown statistical procedures, if they are named with one of the already specified or user adjustable search terms at the end (e.g. '*JATSdecoder* algorithm'). Methods with a specifier behind the search term (e.g. 'test on homogeneity of variances') cannot be identified.

To identify the total number of studies reported in a document, the software and correction method used, fine-tuned dictionary searches are performed on preselected text parts and phrases. Software identification can be enhanced by adding further software search patterns.

Despite its wide extraction capabilities, the focus here is solely on *study.character()*'s function *get.stats()* and its ability to extract and post-process statistical results out of NISO-JATS formatted research articles. A simple web interface to extract and check statistical results within single articles in different formats (PDF, XML, HTML, DOCX) is hosted at: www.get-stats.app.

Several conversion tools that transform PDF documents into a post processable text object exist. One sophisticated converter is the Content ExtRactor and MINEr (CERMINE)[18] which extracts metadata, full text and parsed references from a PDF file and makes it storable in different formats (plain text, NISO-JATS XML, etc.). The implementations of most steps are based on supervised and unsupervised machine learning techniques, which simplifies the procedure of adapting the system to new document layouts and styles[18].

Language and type setting features allow very individual ways of expressing one and the same bit of information. This is especially relevant when processing text with many formulas, indices, special characters (operators, Greek letters, hyphens, separators, brackets, etc.) and synonymously used characters (Greek/Latin small letter beta: β , sharp german s: ß, HTML beta: '&*beta*;'). In electronic documents characters can be represented by different character codecs (UTF-8, ASCII, Unicode, hexadecimal, HTML, etc., or even pictures) which generally makes each extraction and compilation task on numerical results and other content more complicated.

When compiling PDFs with *CERMINE*, a wide range of compilation errors can occur (e.g.: missed operators, handling of subscripts, undetected Greek and special letters). *JATSdecoder*'s function *letter.convert()* unifies many letter representations and corrects most PDF and *CERMINE* specific conversion errors. This enables *JATSdecoder* to also reliably process PDF files that were converted to NISO-JATS coded XML files by *CERMINE*.

JATSdecoder's algorithms have been developed iteratively based on the PubMedCentral article collection and about 10,000 PDF files from different journals, that were converted with CERMINE. *get.stat()* is designed for numbers that are reported with a dot as a decimal separator.

How *get.stats()* works. A two-step process is performed to extract the reported results within a text and recalculate the reported *p* value with *get.stats()*. First, the input text is converted into sentences, squared into round brackets. Only those sentences are selected, that contain at least one letter and an operator followed by a number. To extract the reported test statistic, degrees of freedom, corresponding effect measure and *p* value, they are split at a set of words (e.g. 'and', 'or', 'were', 'of', etc.) and at words followed by a comma. If multiple test results are identified in a text snippet (e.g. more than one t- or *p* value), it is further split up, assuming a test statistic is reported in front of its *p* value. Text that appears in front and behind the results is removed with regular expressions (e.g. the text behind the last reported operator pointing to a number). The first result is a vector with unified representations of sticked results, starting with any letter or letter-number combination with, if present, degrees of freedom in round brackets, pointing to a number with an operator. Several heuristics to unify the representation of overly big and small numbers are applied. Before extracting the actual value of each standard result and the reported *p* value, regular expressions are used to remove labels of test statistics. Every targeted standard result is extracted from the sticked results with an individual heuristic that copes with a variety of reporting styles. The recognized value of the test statistic, its operator, the degrees of freedom and *p* value of each sticked result is returned as a cell in a matrix, which represents the second output. Each type of result is stored in a separated column, which greatly facilitates further post-processing and identification tasks. In standard mode, the recalculation of p values is performed based on the result matrix using basic R functions for distribution functions (‘$2 * (1 - pnorm(Z))$’, ‘$1 - pchisq(chi2, df)$’, etc.). Users can activate an additional recomputation for one-sided t- and Z-tests, as well as r-values that are reported with degrees of freedom.

The R package *statcheck*.

The R package *statcheck*[14] performs an automated detection of statistical test results reported in APA style. It is capable of extracting adequately reported Z-, t-, F-, r-, Q- and χ^2-statistics with adequately reported degrees of freedom and a p value to check the result on plausibility (see:[19]). *statcheck* recomputes the corresponding p value and flags inconsistencies to the reported p value. The built-in functions work on plain text (*statcheck()*), HTML (*checkHTML()*) and PDF files (*checkPDF()*).

Nuijten et al.[20] validated *statcheck* on the manually coded analysis of errors in all reports of statistically significant t-, F- and χ^2-test results in 48 articles, published by the Journal of Personality and Social Psychology and Journal of Experimental Psychology: Learning, Memory, and Cognition. *statcheck* extracted 648 out of 1,120 results (57.9%) in the comparative dataset (one retracted study with 28 significant results, that was part of the original analysis, was excluded by the *statcheck* authors).

Screening 39,717 articles published by eight journals with *statcheck* Nuijten et al.[15] found checkable results in 16,695 documents (42%). Here *statcheck* flagged 8,273 (49.6 %) of these 16,695 articles with at least one inconsistency. Hartgerink[21] analyzed 167,318 articles published by APA, Springer, Sage, and Taylor & Francis with *statcheck* and found 688,112 checkable statistical results in 50,845 articles (30.4%).

As noted by Schmidt[22] *statcheck*’s identification rate for statistical results is rather low. This is in part due to its inability to handle statistical results that are not reported exactly according to APA style, reported with degrees of freedom (or label) in subscript, that contain semicolons instead of commas, square brackets instead of parentheses, effect sizes in-between test statistic and p value[23].

As there is growing enforcement not to rely on the standard p value thresholds of ‘$p = .05$’ too much but rather change it to ‘$p < 0.005$’[24], report effect sizes and confidence intervals instead[25], or even turn away from frequentist methods entirely[26], *statcheck* will ever get worse in doing a good job as a detector of statistical results in text, the more these demands are implemented in practice.

As *statcheck* falsely flags inconsistencies in p values, when appropriate correction methods have been applied (p value correction for multiple testing instead of α-error adjustment) and therefore might encourage users not to use the appropriate methods, Schmidt[22] concludes that *statcheck* is an unsuitable software to detect errors in statistical results and should rather not be used.

Distinguishing features of *get.stats()* and *statcheck*.

Compared to *statcheck*, that looks out for a narrow set of exact pattern matches in a string, *get.stats()* deals with almost any result reported in text. In contrast to *statcheck* commas as well as semicolons used as separators can be handled by *get.stats()*.

Before extracting the actual value of every detected standard result, *get.stats()* selects, splits and cleans up all sentences presenting statistical results. *get.stats()* extracts and post-processes many standard results that are labeled or indexed. It performs several transformations of the textual representation of numbers in text. Fractions, as well as results reported with a ‘e^number’ or a percent sign are compiled to decimal numbers, commas in large numbers ($\geq$ 1000) are removed. The output should therefore not be treated as an exact representation of the reported results.

Whereas *statcheck*’s functions always analyze the full document or text entered, *study.character()*’s argument ‘*text.mode*’ enables an extraction with *get.stats()* on specific text parts (1: full text and abstract, 2: method and result section/s, 3: result section/s only).

statcheck treats non-significant p values reported with ‘ns’ as checkable results, whereas *get.stats()* treats such results as computable, if the reported result allows a recomputation of the p value (e.g.: ‘$t(18) = 1.1, ns$’).

Table 1 lists some potential results of a vector with identified sticked results by *get.stats(x,output = ‘stats’)*. The selected examples demonstrate how *get.stats()* and *statcheck()* differ, in terms of their ability to detect, extract and check statistical results reported in text.

In most of the listed examples, *get.stats()* extracts all contained standard results defined earlier, whereas *statcheck()* fails to detect many of the results at all and extracts some results inadequately. Any squared statistic, as well as any statistic denoted with one of 18 upper- or lowercase letters (except: B, F, N, R, T, Q, W, Z) that is reported with its degrees of freedom in brackets is interpreted as χ^2-test results by *statcheck*. ‘rp’-, ‘sr’-, ‘pr’- and ‘LR’-statistics are interpreted as correlations by *statcheck()*, which, in part, may be correct. *get.stats()* does not classify these letter combinations as standard results. Results reported as intervals may cause missing or erroneous detections by *get.stats()* as the last example in Table 1 demonstrates.

Method

To evaluate and compare the *JATSdecoder* and *statcheck* algorithms in terms of their practical precision and reliability in extracting statistical results in prespecified text parts, two analyses are performed with different input formats.

First, the total number of manually extracted statistically significant t-, F- and χ^2-statistics in the method and result section of 49 articles by Wicherts et al.[27] is compared to the number of computable, statistically significant t-, F- and χ^2-results extracted from the method and result section with *study.character(x,text.mode = 2)* and *statcheck*’s algorithms. The differences between the manually coded data and *study.character()*’s detections are described case by case.

Type	Example	*get.stats()*	*statcheck()*
APA t-test result	‘t(12) = 1.9, p > .05’	‘t(12) = 1.9, p > .05’	‘t(12) = 1.9, p > .05’
APA F-test result	‘F(2, 12) = 3.12, p < .05’	‘F(2, 12) = 3.12, p < .05’	‘F(2, 12) = 3.12, p < .05’
APA r-test result	‘r(13) = .52, p < .05’	‘r(13) = .52, p < .05’	‘r(13) = .52, p < .05’
APA Z statistic in front of line	‘Z = 1.9, p > .05’	‘Z = 1.9, p > .05’	
APA Z statistic behind white space	‘Z = 1.9, p > .05’	‘Z = 1.9, p > .05’	‘Z = 1.9, p >.05’
APA Q-test result	‘Q(13) = .52, p >.05’	‘Q(13) = .52, p >.05’	‘chi2(13) = .52, p >.05’
Non APA t-test result	‘t = 1.9, df = 12, p >.05’	‘t(12) = 1.9, p >.05’	
Non APA F-test result	‘F = 3.12, df1 = 3, df2 = 14, p <.05’	‘F(3, 14) = 3.12, p <.05’	
Semicolon as separator	‘F(1, 46) = 21; p <.05’	‘F(1, 46) = 21, p <.05’	
High df with comma	‘F(12; 1,222) = .12, p <.05’	‘F(12, 1222) = .12, p <.05’	
High F result with semicolon as separator	‘F(12; 122) = 2,123; p <.05’	‘F(12, 122) = 2123, p <.05’	
Test result with ns instead of p value	‘t(12) = 1.9, ns’	‘t(12) = 1.9’	‘t(12) = 1.9, ns’
APA t-test result with effect size	‘t(12) = 1.9, d = .2, p >.05’	‘t(12) = 1.9, d = .2, p >.05’	
Multiple completely reported results	‘all ts(27)>4.2, p <0.01’	‘t(27)>4.2, p <0.01’	
Multiple incompletely reported results	‘all rs<0.2, all ps>.01’	‘r<0.2, p >.01’	
Only p value	‘p <0.05’	‘p <0.05’	
t statistic with numbered index	‘t2(122) = 1, p >.05’	‘t(122) = 1, p > .05’	
F statistic with lettered index	‘Finteraction(1, 46) = 2.8, p <.05’	‘F(1, 46) = 2.8, p <.05’	
G^2 goodness of fit statistic	‘G2(41) = 2.3, p <.05’	‘G2(41) = 2.3, p <.05’	‘chi2(41) = 2.3, p <.05’
Result with capital T instead of t	‘T(12) = 2.33, p <.05’	‘t(12) = 2.33, p <.05’	‘t(12) = 2.33, p <.05’
Result with fraction	‘t(12) = 1/2, p > .05’	‘t(12) = .5, p > .05’	
Result with corrected p value	‘t(122) = 3, p <.05/2’	‘t(122) = 3, p <.025’	
Two reported statistics in a row	‘r(12) = .22, Z = .75, p = .45’	‘r(12) = .22, Z = .75, p = .45’	
Incomplete but p computable result	‘chi2(12) = 12.3’	‘chi2(12) = 12.3’	
Test on beta without z-/t value	‘beta = 22, SE = .77, p <0.01’	‘beta = 22, SE = .77, p <0.01’	
Test on beta without z-/t- nor p value	‘beta = 1.1, SE = .71’	‘beta = 1.1, SE = .71’	
Delta R^2 result	‘ΔR2 = 34%, p <.05’	‘R2 = .34, p <.05’	
BayesFactor result with beta and p value	‘beta = 1.2, BF(10)<1, p = .72’	‘beta = 1.2, BF(10)<1, p = .72’	
BayesFactor result withH0:H1	‘BF(01) = 2e2’	‘BF(10) = 0.005’	
Pearson correlation	‘rp(12) = .22, p = .45’	‘p = .45’	‘r(12) = .22, p = .45’
Pearson correlation	‘sr(12) = .22, p = .45’	‘p = .45’	‘r(12) = .22, p = .45’
Pearson correlation	‘pr(12) = .22, p = .45’	‘p = .45’	‘r(12) = .22, p = .45’
Other statistic: LR statistic	‘LR(12) = .1, p > .05’	‘p > .05’	‘r(12) = .1, p > .05’
Other statistic: I^2 statistic	‘I^2(22) = 1, p > .05’	‘p > .05’	‘chi2(22) = 1, p > .05’
Any hight 2 statistic	‘^{2}(22) = 1, p > .05’	‘p > .05’	‘chi2(22) = 1, p > .05’
A statistic	‘A(12) = 2.3, p <.05’	‘p <.05’	‘chi2(12) = 2.3, p <.05’
B statistic	‘B(12) = 2.3, p <.05’	‘p <.05’	
c statistic	‘c(12) = 2.3, p <.05’	‘p <.05’	‘chi2(12) = 2.3, p <.05’
d statistic	‘d(12) = 2.3, p <.05’	‘p <.05’	‘chi2(12) = 2.3, p <.05’
Interval result	‘.12<r<.22, .87 < p < .65’	‘r = .22’	

Table 1. Some examples of statistical results and the extracted standard results by get.stats() with its argument ‘T2t = TRUE’ and statcheck(). Representations are presented in an easy readable format instead of the resulting data tables extracted. Empty cells represent no detections.

The vector containing the extracted sticked results by *get.stats()*, as well as an index/label removed version are further processed with *statscheck*’s function *statcheck()* to demonstrate how the letter correction in *CERMINE* converted PDF documents increases *statcheck*’s detection rate for checkable test results.

All non- or incorrectly converted but corrected operators, that are replaced with ‘<=>’ by *letter.convert()* are converted to ‘=’ before being processed with *statcheck()*. Labels and/or indices of reported test statistics are removed with simple regular expressions. As no other α-error level was identified in the 49 studies, all results that lead to a recomputed p value $\leq$ .05 or that are reported with ‘$p \leq .05$’ are selected to compare the number of extracted significant results. Next the same article collection is analyzed by each algorithm with no limitations on p values nor type of statistics nor on the part of text. The distribution of the number of detected results is displayed in box plots for each procedure and input format.

The second analysis demonstrates that *get.stats()*’s high performance and detection rate for statistical results also holds for much bigger article collections. An unrestricted search for statistical standard results is performed

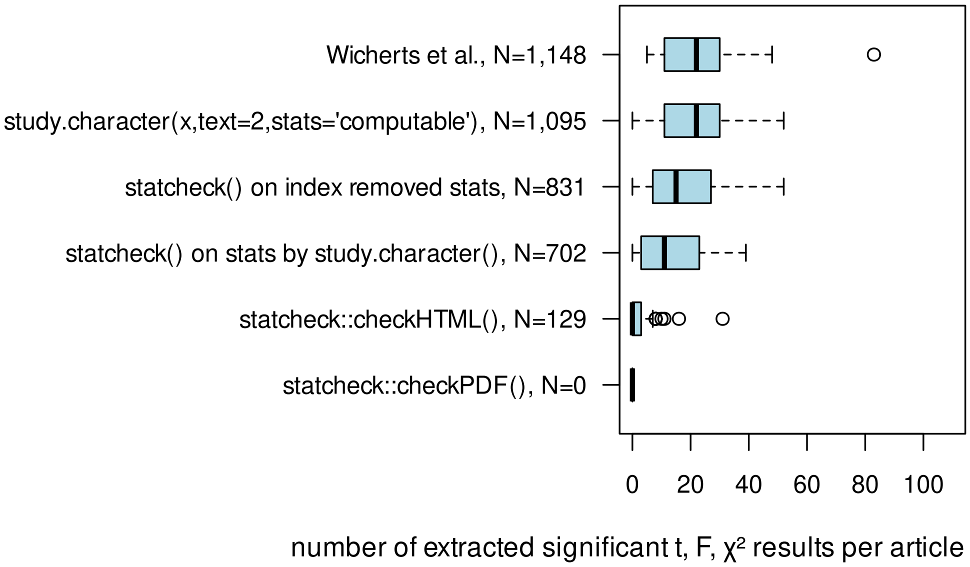


Figure 1. Total sums of extracted significant *t*-, *F*- and χ^2-test results per method and distributions of number of extracted significant *t*-, *F*-, χ^2-test results per article and method.

on 13,531 converted PDF articles, published between 2010 and 2020 in 10 mayor journals of psychology (J. of Abnormal Psychology, J. of Beh. Neuroscience, Psychophysiology, J.o. Child Psychology, Depression & Anxiety, J. of Management, Psychology & Aging, Psychological Medicine, J. of Family Psychology, Personality and Social Psychology Bulletin). A further 18,744 raw NISO-JATS coded XML documents, published by the open access journal *Frontiers in Psychology* and 23,730 'research-article' tagged documents with the pattern '*[Pp]sych*' in its *keyword*- or *subject*-tag published by *PLoS One*, serve for the analysis.

As no manually coded data exists for this big data set with varying input formats, the number of identified standard results by *get.stats()* ('all', 'computable' and 'checkable') is compared to that detected by *statcheck*'s functions with global descriptive measures. The total and relative amount of articles with detectable results and the total sum of detected results is presented for every journal and algorithm setting, as well as some descriptive measures for articles with identifiable results (mean, sd, median, IQR, .99 quantile, maximum, processing time).

All converted PDF documents are passed to *get.stats()* and *checkHTML()* as they contain HTML standard coding. The native PDF files are processed with *checkPDF()* and the preprocessed vector with sticked results extracted by *get.stats(x,output="stats")* is passed to *statcheck()*. Non-significant *p* values reported with 'ns' are excluded before counting *statcheck*'s detections to enable a comparison of the extracted number of checkable results. As the PMC bulk download contains native XML files only, no processing with *checkPDF()* is performed for these studies.

Data, input formats, PDF conversion software, hardware. Native PDF and browser (Mozilla Firefox 80) generated HTML files serve for the first analysis of 49 empirical research articles analyzed in Wicherts et al.[27]. 13,531 PDF files that were published between 2010 and 2020 in 10 mayor journals of psychology were downloaded manually with the library license owned by University of Hamburg. Letters to the editor and corrections are not part of this article collection. PMC's bulk download ftp-server (ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/) was used on 01.01.2021 to download all available native NISO-JATS coded XML documents, published by *Frontiers in Psychology* (18,744 XML files) and *PLoS One* (143,615 XML files).

The open source software *CERMINE*[18] was used to convert each PDF file into a NISO-JATS coded XML, before being processed with *JATSdecoder*'s function *study.character()* or *get.stats()*.

All extractions and analyses were performed with an AMD@Epyc 7452 32-core processor running with Linux Ubuntu 20.04.1 LTS and the open source software R 3.6[28]. To enable multicore processing, the R package *future.apply* was used[29].

Results

Evaluation of *get.stats()* detection rate with manually coded data and *statcheck*'s functions.

First the total number of significant *t*-, *F*- and χ^2-test results that was extracted manually by Wicherts et al.[27] is compared to the number of significant results extracted by *study.character()* and *statcheck*'s functions. Figure 1 displays the distribution of identified significant *t*-, *F*-, and χ^2-statistics per paper for the applied extraction method and input format.

study.character() identifies 1,095 significant results in the method and result sections compared to 1,148 results extracted by Wicherts et al.[27]. *checkHTML()* only detects 129 significant *t*-, *F*- and χ^2-test results within the full text of the browser generated HTML documents. *checkPDF()* could not extract a single statistic out of the same raw material, that was converted with *CERMINE* to become processable with *JATSdecoder*. The extracted sticked

ID	N results Wicherts	N study. character()	Δ	multiple result	Error in *p* value	p operator	Result in footnote	Other section	Fit index	Tabled result	CERM-INE	Un-clear
2	13	15	2			2						
3	33	35	2					2				
6	30	49	19	8		5						6
9	9	10	1		1							
10	21	22	1									1
13	21	22	1				1					
17	37	51	14			13						1
18	11	14	3	3								
19	39	40	1			1						
21	46	52	6				6					
23	35	39	4				4					
26	33	37	4					4				
28	21	22	1	1								
29	24	26	2				2					
32	16	20	4						3			1
33	23	27	4			2						2
34	24	26	2				1					1
36	29	30	1			1						
39	20	21	1			1						
42	28	31	3					3				
43	27	29	2			2						
44	8	20	12	1			3					8
45	7	11	4		1	1						2
46	9	12	3		2	1						
49	30	37	7				7					
5	83	33	−50							−48		−2
16	32	23	−9								−9	
22	20	7	−13							−13		
24	30	29	−1					−1				
31	6	0	−6							−6		
35	5	3	−2									−2
38	15	0	−15							−15		
41	48	33	−15							−16		1
47	36	27	−9							−11		2
48	45	8	−37							−37		
Sum			−53	13	4	29	24	8	3	−146	−9	21

Table 2. Causes for deviations in the number of extracted statistically significant *t*-, *F*-, and χ^2-test results within method and result section/s per paper by Wicherts et al.[27] and study.character().

results extracted by *get.stats()* within the method and result section/s and an additional removal of labels/indices by simple regular expressions, enhances *stacheck()*'s ability to detect and check results (702 and 831 results respectively) and supersedes *checkHTML()*'s functionality for browser generated HTML files.

Here *study.character()* extracts 53 significant results less, than were found in the manual analysis. As 146 significant results are reported in tables and not extracted by *study.character()*, 93 additive significant results are identified. Table 2 summarizes each of the 35 cases with deviations to Wicherts et al.[27]. There are several reasons for a higher detection rate by *study.character()*. 13 checkable test results that are reported for several tests (e.g. 'all ts(18)>3, ps<.05') are extracted by *study.character()*. Four results that are incorrectly reported with '$p > .05$', although they are significant, were not included by Wicherts et al.[27] but found with *study.character()*. As none of the 49 *CERMINE* converted PDF files contains readable operators, *letter.convert()* inserts "<=>" to these empty or badly captured text parts. An insignificant result reported with '$p > .05$' is therefore indistinguishable from a significant result reported with '$p < .05$'. This leads to 29 false positive inclusions in total. 24 results that are reported in footnotes and identified by *study.character()* seem not to be included in the original analysis. One result reported in the description of an experiment seems to be included in the original analysis but is not identified by *study.character()*, as only method and results sections are selected. In three articles, *study.character()* detects a total of nine results in the method sections that seem not to be included in the manual extraction. Three goodness of fit χ^2-statistics are excluded by Wicherts et al.[27] and included by *study.character()*. Nine significant result are missed by *study.character()* because some text parts or section titles got lost while PDF conversion.

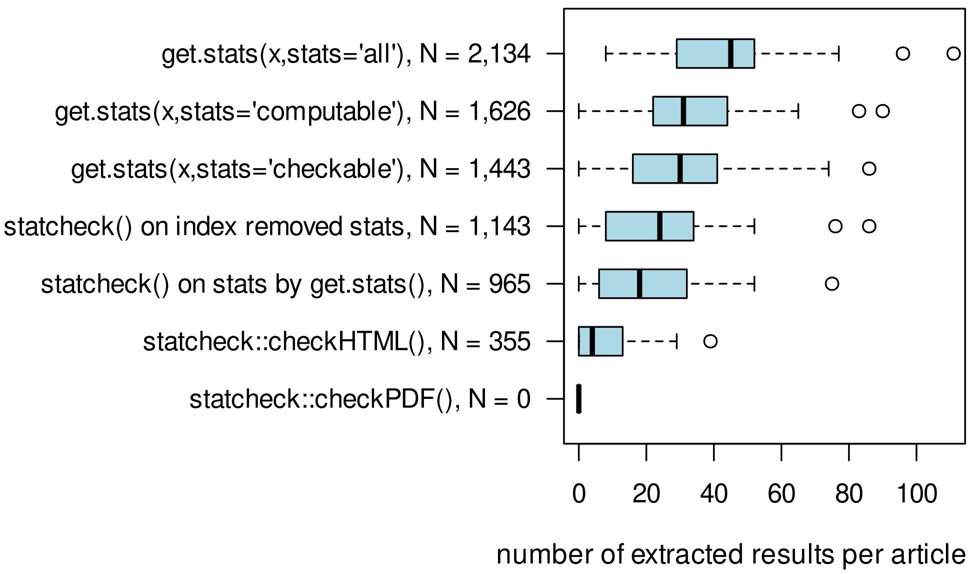


Figure 2. Total sums of extracted test results per method and distributions of number of extracted results per article and method.

Compared to the manually coded data, there are four missed results and 25 detections by *study.character()* that cannot be explained and might be due to bad captures by Wicherts et al.[27].

Figure 2 displays the distribution of all detected statistical standard results per paper for the different extraction methods and input formats, with no restrictions to significant results nor type or text parts. No manually coded data exists for this analysis. In total, *get.stats()* identifies 2,134 statistical standard results in the abstracts and full text parts. 1,626 of these results are reported in a manner, that enables a recomputation of *p* values. 1,443 results are checkable. The preprocessed and further index removed vector extracted by *get.stats(x,output="stats")* increases statchecks detection rate from 355 to 965, or even 1,143 results respectively. No false positive inclusion of a checkable result by *get.stats()* was observed.

Analysis of a large article collection with varying publishers and input formats. Next, the collection of all published PDF files by 10 mayor journals of psychology as well as all ever published XML documents by 2 open access journals is used to extend the evaluation of *get.stats()* to a bigger data set. The absolute and relative frequency of documents with extractable results per journal, different algorithm settings and input formats is listed in Table 3.

In 89% of all processed documents *get.stats()* extracted at least one statistical result (operator between letter-number combination and number). In 46% of all analyzed articles *get.stats()* detects at least one computable result and in 44% at least one checkable result (both with arguments 'T2t' and 'R2r' set to TRUE). Activating *get.stats()*'s argument 'estimateZ' has a small effect (+1%) on the total sum of identified documents with computable and checkable results.

In every journal and input format, all *statcheck* functions detect fewer documents with checkable results. In 38% of all articles *statcheck()* finds checkable results within the extracted sticked results by *get.stats()*, *checkHTML()* in 19% of all CERMXML/XML files and *checkPDF()* in 26% of all PDF files. All or most articles by four journals cannot be handled by *statcheck*'s functions *checkHTML()* and *checkPDF()*, as the compiled PDF files contain incorrectly converted operators.

The amount of articles that contain computable and/or checkable results varies greatly between journals. Overall the journal *Personality and Social Psychology Bulletin* contains checkable results in 91% of the articles, compared to 34% of all articles distributed by *Depression and Anxiety*.

The preprocessed text vector that is returned by *get.stats(x,output='stats')* enhances *statcheck()*'s ability to detect documents with checkable results in every journal. Both format specific *statcheck* functions *checkHTML()* and *checkPDF()* identify less documents in every journal.

Table 4 lists the total number of extracted results, standard results, as well as computable and checkable results in each setting and gives descriptive measures for those articles that contain extractable results. In total, *get.stats()* extracts 1,568,555 sticked results, 981,529 statistical standard results out of which 386,172 represent computable and 359,440 checkable results. Compared to the *statcheck* algorithms, the total sum of detected checkable results by *study.character()* is higher in every journal and input format. 12,249 computable results become checkable when activating *get.stats()*'s option to compute *p* values on estimated Z-values (from 347,191 to 359,440).

Within those articles that contain checkable results, the mean number of detected results is 14.2 with *get.stats()* and 13.5 with *statheck()* on the same preprocessed result vector, 10.7 with *checkHTML()* but 15.2 with *checkPDF()*. Also, the median, interquartile range (IQR), .99 quantile and maximum of checkable results detected by *get.stats()* are higher than *statcheck()*'s measures when processing the same vector and relevantly higher to

Package:		**JATSdecoder**						**statcheck**		
Function:		***get.stats(x,output=c('stats',standardStats))***						***statcheck()***	***checkHTML()***	***checkPDF()***
Input format:		**CERMXML/native XML***						**pre processed text**	**CERMXML/ XML***	**PDF**
Extra arguments: '*T2t*' and '*R2r*'				**TRUE**	**TRUE**	**TRUE**	**TRUE**			
Extra argument: '*estimateZ*'					**TRUE**		**TRUE**			
Extracted results:		**'all stats'**	**'standardStats'**	**'comp.'**	**'comp.'**	**'check.'**	**'check.'**	**'check.'**	**'check.'**	**'check.'**
Journal	**N articles**	**Absolute and relative frequency of articles with extractable results**								
Behavioral Neuroscience	783	713 (91%)	706 (90%)	643 (82%)	643 (82%)	633 (81%)	633 (81%)	616 (79%)	0 (0%)	0 (0%)
Depression & Anxiety	1261	1183 (94%)	1058 (84%)	529 (42%)	547 (43%)	413 (33%)	429 (34%)	328 (26%)	267 (21%)	277 (22%)
J. of Abnormal Psychology	966	926 (96%)	899 (93%)	629 (65%)	647 (67%)	610 (63%)	625 (65%)	589 (61%)	0 (0%)	0 (0%)
J. of Child Psych. & Psychiatry	1497	1295 (87%)	1155 (77%)	681 (45%)	708 (47%)	661 (44%)	687 (46%)	560 (37%)	543 (36%)	563 (38%)
J. of Family Psychology	1146	1131 (99%)	1,102 (96%)	797 (70%)	836 (73%)	773 (67%)	810 (71%)	733 (64%)	1 (0%)	7 (1%)
J. of Management	839	638 (76%)	559 (67%)	271 (32%)	275 (33%)	231 (28%)	236 (28%)	157 (19%)	169 (20%)	171 (20%)
Pers. and Social Psychology Bul.	1341	1332 (99%)	1330 (99%)	1219 (91%)	1237 (92%)	1204 (90%)	1218 (91%)	1169 (87%)	1179 (88%)	1182 (88%)
Psychological Medicine	2924	2741 (94%)	2542 (87%)	1162 (40%)	1208 (41%)	1086 (37%)	1129 (39%)	818 (28%)	540 (18%)	579 (20%)
Psychology & Aging	1031	1022 (99%)	1006 (98%)	789 (77%)	812 (79%)	776 (75%)	795 (77%)	729 (71%)	0 (0%)	0 (0%)
Psychophysiology	1743	1708 (98%)	1671 (96%)	1461 (84%)	1473 (85%)	1439 (83%)	1448 (83%)	1361 (78%)	592 (34%)	696 (40%)
Sum in CERMXML/native PDF	13,531	12,689 (93.8%)	12,028 (88.9%)	8181 (60.5%)	8386 (62%)	7826 (57.8%)	8010 (59.2%)	7060 (52.2%)	3291 (24.3%)	3475 (25.7%)
Frontiers in Psychology	18,744*	14,362 (77%)*	13,091 (70%)*	9222 (49%)*	9408 (50%)*	8902 (47%)*	9070 (48%)*	7734 (41%)*	4465 (24%)*	
PLoS One	23,730*	22,675 (96%)*	20,211 (85%)*	8432 (36%)*	8558 (36%)*	8043 (34%)*	8153 (34%)*	6573 (28%)*	2815 (12%)*	
Sum in native XML files	42,474	37,037 (87.2%)	33,302 (78.4%)	17,654 (41.6%)	17,966 (42.3%)	16,945 (39.9%)	17,223 (40.5%)	14,307 (33.7%)	7280 (17.1%)	
Total Sum	56,005	49,726 (89%)	45,330 (81%)	25,835 (46%)	26,352 (47%)	24,771 (44%)	25,233 (45%)	21,367 (38%)	10,571 (19%)	3475 (26%)

Table 3. Absolute and relative frequency of articles with extractable, computable or checkable statistic by journal, input format, additive settings and algorithm.

checkHTML() and *checkPDF()*. *get.stats()* detects the highest number of checkable results in one study with 199 results, whereas *statcheck()* identifies 149 results as maximum.

No unexpected processing times occurred. As many preprocessing operations are performed, the extraction of the sticked results with *get.stats(x,output='stats')* takes 1.3 seconds on average per paper and processor. The mean processing time of this vector differs slightly between *statcheck()* (.6 sec.) and *get.stats()* (.5 sec. per document and processor). In total, both file specific *statcheck* functions work a lot faster, as no case specific letter conversion nor uniformization is performed before extracting the results.

Table 5 displays the increase factors with which *get.stats()* identifies more checkable results per journal. As no PDF files are analyzed for *Frontiers in Psychology* and *PLoS One*, these fields are left blank for *checkPDF()*. *get.stats()* outperforms *statcheck()* in detecting checkable results by a varying factor of 1.07 for *Behavioral Neuroscience* to 1.73 for *Journal of Management* when processing the same preprocessed vector of sticked results extracted with *get.stats(x,output="stats")*. This pattern holds for *checkHTML()* when processing *CERMINE* converted PDFs (1.13 to 2.84) and *checkPDF()* processing the original PDF files. Three PDF article sets mostly contain non-standard coded operators and cannot be processed in their native version by *checkPDF()* nor in their *CERMINE* compiled version by *checkHTML()*. Compared to *checkHTML() get.stats()* extracts 3.33 (*Frontiers in Psychology*) to 4.1 (*PLoS One*) times more checkable standard results within the native XML files with most results coded in HTML style.

Conclusion

get.stats()'s high precision and flexibility in extracting statistical results from research papers in NISO-JATS formatted XML files has been demonstrated. It facilitates plausibility checks on many standard results reported in text, and can help scientists as well as editors to summarize and check a study regarding reporting style and

Package:		**JATSdecoder**						**statcheck**		
Function:		**get.stats(x,output=c('stats',standardStats))**						***statcheck()***	***checkHTML()***	***checkPDF()***
Input format:		**CERMXML**						**processed text**	**CERMXML**	**PDF**
Extra arguments: '*T2t*' and '*R2r*'				TRUE	TRUE	TRUE	TRUE			
Extra argument: '*estimateZ*'					TRUE		TRUE			
Extracted statistics		**'all stats'**	**'standardStats'**	**'comp.'**	**'comp.'**	**'check.'**	**'check.'**	**'check.'**	**'check.'**	**'check.'**
Journal	**N articles**	**total number of extracted results**								
Behavioral Neuroscience	783	26,239	22,274	14,365	14,370	13,517	13,522	12,658	0	0
Depression & Anxiety	1261	29,930	15,091	4376	4615	3319	3512	2518	2095	2359
J. of Abnormal Psychology	966	33,093	22,893	8902	9255	8372	8669	7922	0	0
J. of Child Psych. & Psychiatry	1497	38,093	20,244	6686	7243	6195	6667	5093	4969	5390
J. of Family Psychology	1146	32,064	20,343	5642	6627	5137	5944	4707	1	11
J. of Management	839	16,247	10,210	2028	2246	1544	1726	998	1013	1063
Pers. and Social Psychology Bul.	1341	89,066	53,733	28,377	31,410	27,229	29,588	25,159	26,221	26,261
Psychological Medicine	2924	69,799	41,633	10,415	10,858	9452	9821	7373	3922	4579
Psychology & Aging	1031	44,318	30,653	16,314	17,071	14,993	15,506	12,751	0	0
Psychophysiology	1743	68,170	49,853	30,415	30,941	28,799	29,115	25,436	10,245	13,143
Sum in CERMXML/native PDF	13,531	447,019	286,927	127,520	134,636	118,557	124,070	104,615	48,466	52,806
Frontiers in Psychology	18,744	458,136	287,485	127,036	132,771	120,675	125,032	98,842	37,555	
PLoS One	23,730	663,400	407,117	115,873	118,765	107,959	110,338	85,463	26,915	
Sum in native XML files	42,474	1,121,536	694,602	242,909	251,536	228,634	235,370	184,305	64,470	
Total sum	56,005	1,568,555	981,529	370,429	386,172	347,191	359,440	288,920	112,936	52,806
Mean		31.5	21.7	14.3	14.7	14	14.2	13.5	10.7	15.2
SD		26.1	19.9	14.8	15	14.4	14.6	13.8	12	15.5
median		25	16	10	10	9	10	9	7	10
IQR		[14; 42]	[8; 29]	[4; 20]	[4; 20]	[4; 19]	[4; 19]	[4; 18]	[3; 14]	[4; 22]
Quantile99		125	93	69	70	68	68	65	56.3	71
Max		406	329	184	199	184	199	149	126	126
Total time in seconds		1172	463	415	416	419	424	563	153	70
Seconds per paper per processor		1.256	0.496	0.445	0.446	0.449	0.455	0.603	0.68	0.311

Table 4. Total sum of extractable, computable and checkable statistics by journal, input format, additive settings and algorithm. Non checkable results reported with 'ns' are removed from statchecks output, descriptive measures are calculated on those articles with detected results per setting.

Journal	**get.stats() versus *statcheck(get.stats(x, output="stats"))***	**get.stats() versus *checkHTML()***	**get.stats() versus *checkPDF()***
Behavioral Neuroscience	1.07	Inf	Inf
Depression & Anxiety	1.39	1.68	1.49
J. of Abnormal Psychology	1.09	Inf	Inf
J. of Child Psychology & Psychiatry	1.31	1.34	1.24
J. of Family Psychology	1.26	5944.00	540.36
J. of Management	1.73	1.70	1.62
Personality and Social Psychology Bul.	1.18	1.13	1.13
Psychological Medicine	1.33	2.50	2.14
Psychology & Aging	1.22	Inf	Inf
Psychophysiology	1.14	2.84	2.22
Frontiers in Psychology	1.26	3.33	
PLoS One	1.29	4.10	

Table 5. Increase factor of detection rate for checkable results by get.stats() compared to statcheck's functions.

checkability of reported results. However, fully reported and plausible results do not tell us anything about the methodological quality of a study.

get.stats() is heavily outperforming all three *statcheck*'s algorithms in extracting statistics from floating text. The vague definition of a statistical result being any letter-number combination pointing to a number with an operator makes *get.stats()* a very general and valid tool to detect statistical results within text. Incomputable, computable and checkable results become clearly distinguishable. If possible, *p* values are recomputed and become checkable if also reported. Incompletely or inconsistently reported results can be detected by a manual check of reported and computed *p* values.

JATSdecoder's functions can handle most PDF and *CERMINE* specific conversion errors in statistical results, except in cases with non compiled text parts (e.g. footnotes, listings, section titles). Incorrectly converted operators and some Greek letters are corrected, while completely missing operators are replaced with '<=>' for many statistical results. The extracted vector of sticked results by *get.stats()*, converts CERMINE converted PDF files, that are unprocessable for *checkPDF()*, into a format that is post-processable with *statcheck()*.

The results of Nuijten et al.[20] could not be replicated with neither input format. Compared to the original paper, *checkPDF()* does not detect a single checkable result in the PDF files, while *checkHTML()* just detects a small proportion in the browser generated HTML files. Finally, *statcheck()* identifies more checkable results within the preprocessed output of *get.stats()* than were found by Nuijten et al.[20]. Therefore, *get.stats()* preprocessed output enhances any automated plausibility check with *statcheck()*, especially for those PDF files that compile with errors, which applies to full article collections of some journals.

In all cases, *get.stats()* outperforms all *statcheck* algorithms. Even compared to a manual extraction, its precision on extracting statistical results from text can be considered very high. In some rare cases, the compilation by *CERMINE* failed to cover all text parts, leading to some undetected results. However, this problem only needs to be considered when PDF conversion was applied.

Most deviations observed to the manually coded data by Wicherts et al.[27] are caused by their representation in tables, differing inclusion criteria and/or differing definitions of a checkable result. No false positive detections of checkable results by *get.stats()* were observed.

A non-negligible part of all reported results in the surveyed articles is presented in tables and cannot be extracted nor checked by neither *get.stats()* nor *statcheck*. Converting tables in PDF files to text mostly produces spurious artifacts in the resulting output, as they allow very individual layout and coding styles. *statcheck* detects results reported in tables if they are reported in a full textual manner in one cell of the table, which is a rather rare event. Up to now, as only a very small portion of tabulated results can be extracted with *statcheck*, it is sensible to restrict checking procedures to results reported within the main text only. Descriptive measures of the total number of reported results in text therefore tend to be mostly negatively biased estimates for the actual number of reported results. Correlation matrices and regression tables often contain a high amount of test results. For test results reported with asterisks instead of *p* values, a precise plausibility check is generally not possible.

As no algorithm can be perfect, false positive and negative detections may occur when *get.stats()* tags a reported result as a standard result. Many PDFs lose their special characters during conversion to NISO-JATS coded XML files which may lead to false positives and negatives, when a missing Greek letter other than χ is used but χ is imputed by *letter.convert()*. Results that are labeled equally to the above defined standard results but represent other measures, will be treated as a standard result. Especially wrongly interpreted Z-values (e.g. in a coordinate: 'x = 1, y = 2, z = 3') will automatically lead to the computation of a *p* value and suggest that the result is computable. Special or anomalous labels of results and special letter uses that are not captured by *get.stats()* may lead to a non-detection as checkable standard result.

JATSdecoder enables a wide range of possibilities for meta-analytical research and mirroring techniques. The reported degrees of freedom in some test statistics allow an estimation of the sample size which a study is based on. Another option is to analyze all ever reported statistics by an author, affiliation, subject and/or other subsets of metadata. A p-curve analysis of the reliably extracted results from one or many article/s may help to identify questionable research practices performed by individuals or groups. Its ability to split an article into selectable sections and phrases enables sentence detection in specific text parts of a study (e.g. discussion/conclusion only). With a little additive text extraction effort, it is possible to detect all investigated variables or effects within a research topic.

Data availability

JATSdecoder software is freely available at: https://github.com/ingmarboeschen/JATSdecoder A simple web interface enables the use of *JATSdecoder*'s function *get.stats()* on single files of different formats: www.get-stats.app Scripts to reproduce this and other analyses performed with *JATSdecoder*, as well the extracted results from *Frontiers in Psychology* and the selection of *PLoS One* articles are stored at:https://github.com/ingmarboeschen/JATSdecoderEvaluation.

Received: 30 June 2021; Accepted: 14 September 2021
Published online: 30 September 2021

References

1. PubMed-Central. PMC Overview https://www.ncbi.nlm.nih.gov/pmc/about/intro (2019).
2. Cohen, J. The statistical power of abnormal-social psychological research: A review. *J. Abnorm. Soc. Psychol.* **65**, 145–153. https://doi.org/10.1037/h0045186 (1962).
3. Meehl, P. E. Theory-testing in psychology and physics: A methodological paradox. *Philos. Sci.* **34**, 103–115. https://doi.org/10.1086/288135 (1967).
4. Gigerenzer, G. Mindless statistics. *J. Socio-Econ.* **33**, 587–606. https://doi.org/10.1016/j.socec.2004.09.033 (2004).

5. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2**, e124. https://doi.org/10.1371/journal.pmed.0020124 (2005).
6. Gelman, A. & Carlin, J. Beyond power calculations: Assessing type S (Sign) and type M (Magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641–651. https://doi.org/10.1177/1745691614551642 (2014).
7. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, 6251. https://doi.org/10.1126/science.aac4716 (2015).
8. Sedlmeier, P. & Gigerenzer, G. Do studies of statistical power have an effect on the power of studies?. *Psychol. Bull.* **105**, 309–316. https://doi.org/10.1037/0033-2909.105.2.309 (1992).
9. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world?. *Behav. Brain Sci.* **33**, 61–83. https://doi.org/10.1017/S0140525X0999152X (2010).
10. Gigerenzer, G., Krauss, S. & Vitouch, O. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In *Handbook on quantitative methods in the social sciences* Vol. 21 (ed. Kaplan, D.) 389–406 (Sage, Thousand Oaks, 2004). https://doi.org/10.4135/9781412986311.n21.
11. Simmons, J. P., Nelson, L. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psych. Sci.* **22**, 1359–1366. https://doi.org/10.1177/0956797611417632 (2011).
12. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psych. Sci.* **23**, 524–532. https://doi.org/10.1177/0956797611430953 (2012).
13. Brown, N. J., Heathers, J. A. & Test, The GRIM. A simple technique detects numerous anomalies in the reporting of results in psychology. *Soc. Psychol. Pers. Sci.* **8**, 363–369. https://doi.org/10.1177/1948550616673876 (2017).
14. Epskamp, S., Nuijten, M. B. statcheck: Extract Statistics from Articles and Recompute p Values. R package version 1.3.0 https://CRAN.R-project.org/package=statcheck (2018).
15. Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S. & Wicherts, J. M. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* **48**, 1205–1226. https://doi.org/10.3758/s13428-015-0664-2 (2016).
16. Böschen, I. JATSdecoder: A meta data and text extraction and manipulation tool set for the statistical programming language R. https://www.github.com/ingmarboeschen/JATSdecoder (2021).
17. National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM). Journal Publishing Tag Library—NISO JATS Draft Version 1.1d2. https://jats.nlm.nih.gov/publishing/tag-library/1.1d2/index.html (2014).
18. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J. & Bolikowski, Ł. CERMINE: automatic extraction of structured metadata from scientific literature. *Int. J. Doc. Anal. Recognit.* **18**, 317–335. https://doi.org/10.1007/s10032-015-0249-8 (2015).
19. Nuijten, M. B. & Polanin, J. R. "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Res. Synth. Methods* **11**, 574–579. https://doi.org/10.1002/jrsm.1408 (2020).
20. Nuijten, M. B., van Assen, M. A., Hartgerink, C. H., Epskamp, S., Wicherts, J. The Validity of the Tool ”statcheck” in Discovering Statistical Reporting Inconsistencies. Preprint at https://psyarxiv.com/tcxaj (2017).
21. Hartgerink, C. H. 688,112 Statistical results: Content mining psychology articles for statistical test results. *Data* **1**, 14. https://doi.org/10.3390/data1030014 (2016).
22. Schmidt, T. Statcheck does not work: All the numbers. Reply to Nuijten et al. (2017). Preprint at https://psyarxiv.com/hr6qy.
23. Journal of Experimental Social Psychology. JESP piloting the use of statcheck. www.journals.elsevier.com/journal-of-experimental-social-psychology/news/jesp-piloting-the-use-of-statcheck (2017).
24. Benjamin, D. J. *et al.* Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6. https://doi.org/10.1038/s41562-017-0189-z (2018).
25. Cumming, G. The new statistics: Why and how. *Psychol. Sci.* **25**, 7–29. https://doi.org/10.1177/0956797613504966 (2014).
26. Wagenmakers, E. J., Wetzels, R., Borsboom, D. & Van Der Maas, H. L. Why psychologists must change the way they analyze their data: The case of psi. *Comment on Bem. J. Pers. Soc. Psychol.* **100**, 426–432. https://doi.org/10.1037/a0022790 (2011).
27. Wicherts, J. M., Bakker, M. & Molenaar, D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE* **6**, e26828. https://doi.org/10.1371/journal.pone.0026828 (2011).
28. R Core Team. R. A Language and Environment for Statistical Computing. https://www.R-project.org/ (2021).
29. Bengtsson, H. future.apply: Apply Function to Elements in Parallel Using Futures. R package version 1.4.0. https://CRAN.R-project.org/package=future.apply (2020).

Author contributions

The manuscript and all analytics were created by the author.

Funding

Open Access funding enabled and organized by Projekt DEAL. This research was financed by a doctoral Grant awarded by the Department of Psychological Methods and Statistics, Institute of Psychology, University Hamburg, Germany.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to I.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Anhang C: Publikation 3

Böschen, I. (2023). Evaluation of the extraction of methodological study characteristics with JATSdecoder. Scientific Reports 13, 139.

https://doi.org/10.1038/s41598-022-27085-y

scientific reports

OPEN

Evaluation of the extraction of methodological study characteristics with JATSdecoder

Ingmar Böschen

This paper introduces and evaluates the *study.character* module from the *JATSdecoder* package which extracts several key methodological study characteristics from NISO-JATS coded scientific articles. *study.character* splits the text into sections and applies its heuristic-driven extraction procedures to the text of the method and result section/s. When used individually, *study.character*'s functions can also be applied to any textual input. An externally coded data set of 288 PDF articles serves as an indicator of *study.character*'s capabilities in extracting the number of sub-studies reported per article, the statistical methods applied and software solutions used. Its precision of extraction of the reported α-level, power, correction procedures for multiple testing, use of interactions, definition of outlier, and mentions of statistical assumptions are evaluated by a comparison to a manually curated data set of the same collection of articles. Sensitivity, specificity, and accuracy measures are reported for each of the evaluated functions. *study.character* reliably extracts the methodological study characteristics targeted here from psychological research articles. Most extractions have very low false positive rates and high accuracy ($\geq$ 0.9). Most non-detections are due to PDF-specific conversion errors and complex text structures, that are not yet manageable. *study.character* can be applied to large text resources in order to examine methodological trends over time, by journal and/or by topic. It also enables a new way of identifying study sets for meta-analyzes and systematic reviews.

In scientific research practice, many individual decisions can be made that affect the scientific quality of a study. There are also changing standards set by journal editors and the community. This applies not only to the study design, but also to the choice of statistical methods and their settings. With new methods and standards, the way research is planned, conducted and presented changes over time and represents an interesting field of research. One aspect to consider is the ever-increasing number of scientific publications coming out each year. Numerous studies have investigated the use and development of statistical techniques in scientific research practice[1–7]. Most of these studies used manually coded data of a limited number of articles, journals, topics or time interval. The selectivity of these samples therefore severely limits the generalizability of the findings to a wider scope. For example, Blanca et al.[7] analyzed the use of statistical methods and analysis software solutions in 288 articles (36 articles each from 8 journals), all from a publication period of about one year.

A technology that is suitable for analyzing large amounts of text and helps to overcome the problem of small samples in the analysis of scientific research practice is text mining. Text mining is the process of discovering and capturing knowledge or useful patterns from a large amount of unstructured textual data[8]. It is an interdisciplinary field that draws on data mining, machine learning, natural language processing, statistics, and more[8]. It facilitates extraction and unification tasks that cannot be done by hand when the analyzed text corpus becomes large. In addition to rudimentary computer commands on textual input (regular expressions), there are also many software programs and toolkits that provide model-based methods of natural language processing (NLP).

Well-known NLP libraries such as *NLTK*[9] or *spaCy*[10] provide users with a variety of programs for linguistic evaluation of natural language. This often involves the use of statistical models and machine learning. In contrast, the *JATSdecoder* package[11] focuses on metadata and study feature extraction (in the context of the NISO-JATS format). This extraction is implemented using expert-driven heuristics. Thus, unlike in the aforementioned large multipurpose NLP libraries, no further programming effort is required to perform specific extraction.

Research on scientific practice can benefit greatly from NLP techniques. Compared to manual coding, an automated identification of study characteristics is very time and cost-efficient. It enables large-scale and trend analyzes, mirroring of scientific research practices and identification of studies that meet certain methodological

Institute of Psychology, Research Methods and Statistics, University Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. email: ingmar.boeschen@uni-hamburg.de

Function	Functionality
letter.convert()	Hexadecimal and HTML to UTF-8 conversion, CERMINE specific error correction
text2sentences()	Converts text string to vector of sentences
text2num()	Converts different representations of numbers to a digit representation
which.term()	Performs search task for multiple patterns and returns hit vector
ngram()	Extracts ±n-gram bag of words aroud a search term hit
strsplit2()	Splits text before, after, or at search pattern
grep2()	Enables search tasks on multiple search patterns connected with a logical AND

Table 1. General text processing functions implemented in JATSdecoder and their functionality.

requirements for meta-analyses and systematic reviews. In addition, automated plausibility checks and global summaries can support quality management.

In general, most methodological study characteristics (e.g., statistical results, α-level, power, etc.) are reported in a fairly standard way. Here, the module *study.character* from the R package *JATSdecoder*[11] is presented and evaluated as a tool for extracting key methodological features from scientific reports. The evaluation of the built-in extraction functions is performed on a medium-sized collection of articles (N = 287) but highlights the possibilities in mirroring and identifying methodological trends in rather big article collections. Although the use of method-based NLP methods might be appropriate for the study features focused here, all functions run fine-tuned expert-driven extraction heuristics to achieve a robust extraction and traceability of errors. While many NLP libraries can be thought of as a toolbox for a variety of problems, *JATSdecoder* represents a precision tool for a specific problem.

The JATSdecoder package

Scientific research is mostly published in two ways. In addition to a printable version which is distributed as PDF file, machine-readable versions are accessible in various formats (HTML, XML, JSON). The PubMed Central database[12] currently stores almost five million open access documents from the biology and health sciences, distributed as XML files and structured using the Journal Archiving Tag System NISO-JATS[13]. The NISO-JATS is an HTML tag standard to store scientific article content without any graphical parameters (website layout, text arrangement, etc.), graphical content is hyper referenced.

JATSdecoder[11] is a software package for the statistical programming language R[15]. Its function *JATSdecoder* converts NISO-JATS encoded XML documents into a list with metadata, user-adjustable sectioned text and reference list[16]. The structured list is very useful for costum search and extraction procedures, as it facilitates these tasks on selectively defined text parts (e.g., section headings, method or results section, reference list).

The algorithms of *JATSdecoder* were iteratively developed based on the PubMed Central article collection (at that time $\approx$ 3 million native NISO-JATS XML) and more than 10,000 PDF files from different journals that were converted to XML files with the Content ExtRactor and MINEr[14](*CERMINE*).

CERMINE is a sophisticated PDF conversion tool which extracts metadata, full text and parsed references from scientific literature in PDF format. The output can be returned as plain text or NISO-JATS encoded content. Compared to a pure text extraction, the transfer into the NISO-JATS format with *CERMINE* is a great advantage for post-processing. Article metadata can be accessed directly and the text of multi-column blocks is extracted correctly, which is often not the case with the output of other conversion software. Supervised and unsupervised machine learning algorithms enable *CERMINE* to adapt to the different document layouts and styles of the scientific literature. Large file collections can be converted using batch processing. Thus, with the help of CERMINE, the publications of another large group of publishers can be processed with *JATSdecoder*.

In addition to the extraction of metadata and study features, *JATSdecoder* provides some convenient, purely heuristic-driven functions that can be useful for any text-analytic approach. An overview of these functions and their functionality is given in Table 1. All functions are based on the basic R environment and make intense use of regular expressions. *letter.convert()* unifies hexadecimal and many HTML letters into a Unicode representation and corrects most PDF and *CERMINE* specific conversion errors. For example, more than 20 different hexadecimal characters that encode a space are converted to a standard space, invisible spaces (e.g.: 'u200b') are removed. When extracting text from PDF documents, special characters can often not be read correctly, as they can be stored in a wide variety of formats. Badly compiled Greek letters (e.g., 'v2' not 'χ^2') and operators (e.g., '5' not '=') are corrected, a '<=>' is inserted for missing operators (e.g., 't<=>1.2, p<=>0.05' for 't 1.2, p 0.05'). These unifications are important for further processing and facilitate text search tasks and extractions. *text2sentences()* converts floating text into a vector of sentences. Many not purely digit-based representations of numbers (words, fractions, percentages, very small/high numbers denoted by 10^x or $e + x$) can be converted to decimals with *text2num()* (e.g., 'five percent' $->$ '0.05', '0.05/5' $->$ '0.01'). *ngram()* extracts a definable number of words occurring before and/or after a word within a list of sentences (±n-gram bag of words). The presence of multiple search patterns can be checked with *which.term()*. The output is either a binary hit vector for each search pattern or a vector of detected search patterns. The functions *grep2()* and *strsplit2()* are useful extensions of the basic R functions *grep()* and *strsplit()*. *grep2()* enables the identification and extraction of text using multiple search patterns linked with a logical AND. Compared to *strsplit()*, which deletes the search pattern when splitting text into pieces, *strsplit2()* allows to preserve the search pattern in the output by supporting splits before or after the recognized pattern.

The *study.character* module bundles multiple text selection and manipulation tasks for specific contents of the list created by *JATSdecoder*. It extracts important study features such as the number of studies reported, the statistical methods applied, reported α-level and power, correction procedures for multiple testing, assumptions mentioned, the statistical results reported, analytical software solution used, and whether the results include an analysis of interacting covariates, mediation and/or moderation effects. All functions use sophisticated, expert-guided heuristics for text extraction and manipulation, developed with great effort and domain expertise. One advantage of the time-intensive development of efficient rules is the robust recognition of a wide range of linguistic and technical representations of the targeted features, as well as a clear assignment of the causes of incorrect extractions. A functional limitation of most *study.character* functions is that they can only handle English content.

In general, *study.character* attempts to split a document into four sections (Introduction, Methods, Results, Discussion). The text of the introduction, which explains the theory and describes other work and results, and the discussion section, which contains implications, limitations, and suggestions for future procedures, can easily lead to false-positive extractions of actually realized study features. This also applies to the information in the bibliography. Therefore, mostly only the methods and results sections and captions are processed to extract the study characteristics from an article.

It has been demonstrated that *study.character*'s function *get.stats()* outperforms the program *statcheck*[17] in extracting and recalculating p-values of statistical results reported within an article in both PDF and XML format[18]. Here, *study.character*'s functions to extract the statistical methods applied, statistical software used, number of studies per article, reported α-level and power, test direction, correction method for multiple testing, and mentioned assumptions are evaluated using manually coded data of the study characteristics.

Description of the extraction heuristics. A brief description of the targeted study feature and the implemented extraction heuristic of each function is given in the following section. Minor uniformization tasks are not listed, but can be traced using the source code of each function. The text processing and feature extraction are implemented with basic R functions (e.g., *grep()*, *gsub()*, *strsplit()*) and *JATSdecoder*'s text processing solutions, which are also based on these basic functions. A main feature of these functions is that they can be used with regular expressions, which makes them very powerful if used wisely. The *grep()* function performs search queries, *gsub()* finds and replaces text. Using *strsplit()*, text input can be split into a vector at text locations that match a search pattern. The search pattern itself is removed.

Statistical method. To draw contextual conclusions, researchers use various statistical methods and procedures to process and summarize their data. Although any descriptive as well as inferential method can be considered a statistical method, the focus here is on inferential methods. Inferential methods are based on either simple or more complex models, which also allow differing depths of data analysis and inference. Some of these methods are widespread in the literature (e.g., t-test, correlation, ANOVA, multiple regression), while other techniques are rarely used.

The function *get.method()* extracts the statistical methods mentioned in the input text. It detects sentences containing a statistical method with a list of search terms that most commonly used procedures share as an identifier (e.g., test, correlation, regression, ANOVA, method, theorem, interval, algorithm, etc.). After lowerization, up to seven preceding words with the identified search term at the end are extracted with *ngram()* and further cleaned up with an iteratively generated list of redundant words (e.g., prepositions, verbs). Users can expand the possible result space by passing additive search words to the '*add*' argument of *get.method()*. The current heuristic enables the extraction of new, still unknown procedures (e.g., '*JATSdecoder* algorithm'), if their name ends with one of the prespecified or user-adjusted search terms. Simple descriptive measures (e.g., mean, standard deviation, proportion) are not extracted, because they are overly common and therefore do not differentiate well. Methods with a specifying term after the search term (e.g., 'test for homogeneity of variances') cannot be identified by *get.method()* yet.

α-level. Theoretically, any frequentist decision process requires an a-priori set significance criterion, the α-level or type-1 error probability. The type-1 or α-error is the probability of rejecting a correct null hypothesis. Because it has become a widespread standard to work with an α-level of 0.05, it is often not explicitly stated in practice. Among many synonyms (e.g., 'alpha level', 'level of significance', 'significance threshold', 'significance criterion') and made up terms (e.g., 'level of confidence', 'level of probability'), it may be reported as critical p-value (e.g., 'p-values $<$ 0.05 are considered significant') and/or with a verbal operator (e.g., 'the α-error was set to 0.05'), making it difficult to detect and extract reliably. In addition, the α-level may be reported with a value, that has been corrected for multiple testing, which does not lower the nominal α-level. Another indirect but clearly identifiable report of an α-error probability is the use of 1-α confidence intervals.

The text of the method and result sections/s, as well as the figure and table captions, are passed to *get.alpha.error()*. Prior to the numerical extraction of the reported α-level/s, several unification tasks are performed on synonymously used terms for α-errors and reporting styles. Levels of different p-values that are coded with asterisks are not considered α-levels. When a corrected α is reported by a fraction that also contains the nominal value (e.g., '$\alpha = 0.05/4$') both values are returned (0.05 and 0.0125). The argument '*p2alpha*' is activated by default to increase the detection rate. This option allows extraction of p-values expressing α-levels (e.g., 'Results with p-values < 0.05 are considered significant.'). The final output is a list distinguishing between detected nominal, corrected α-level/s and extractions from 1-α confidence intervals. Since some articles report multiple α-levels, all detected values are max- and minimized to facilitate further processing.

Correction for multiple testing. The nominal α-level refers to a single test situation. When multiple tests are performed with the same α-level, the probability of obtaining at least one significant result increases with each test and always exceeds α. There are several correction procedures to control the inflation of the α-error or false discovery rate, when running multiple tests on the same data.

A two-step search task is performed for the text of the methods and results section/s, as well as figure and table captions by *get.multiple.comparison()*. Sentences containing any of the search terms 'adjust', 'correct', 'post-hoc' or 'multiple' are further inspected for twelve author names (e.g., 'Benjamini', 'Bonferroni') that refer to correction procedures, as well as four specific procedures (e.g., 'family-wise error rate', 'false discovery rate') that correct for multiple testing (see Online Appendix A for the full list of specified search terms). The output is a vector with all identified authors of correction methods. Common spelling errors (e.g., 'Bonfferoni' instead of 'Bonferroni') are also detected, but returned with the correct name.

Test power. The concept of power describes the probability of correctly rejecting a false null hypothesis given a theoretical (a-priori) or empirical (post-hoc) effect. It can be used to estimate an optimal sample size (a-priori) or as a descriptive post-hoc measure.

get.power() extracts the reported aimed and achieved power value/s that are reported in the full text of the document. Since the term power is used in different contexts, sentences containing certain terms are omitted (e.g., volts, amps, hz). To reduce the likelihood of false positives, detected values that fall outside the valid power range ([0; 1]) are omitted. *get.power()* unifies some synonyms of power (e.g., $1-\beta$) and extracts the corresponding value/s if they fall within the range of 0–1. When β-errors are reported instead of power values, they are converted to power values by replacing β with $1-\beta$.

Interaction effects. Analyses with more than one independent variable can be conducted with or without an interaction effect of the covariates. The term interaction effect refers to any type of interplay of two or more covariates that have dynamic effects on an outcome. In most research settings, the analysis of interactions is of great interest, as it may represent the central research hypothesis or lead to restrictions and/or reinforcement for the hypothesis/theory being tested. In addition to statistical models that explicitly include an interaction effect, mediation- and moderation analyses focus on dynamic effects of covariates on an outcome.

has.interaction() searches the lowerized text of the methods and results section/s for specific search patterns that relate to an interaction effect. To avoid false positive hits when analyzing articles dealing with interactions of organisms instead of variables, sentences containing specific search terms (e.g., social, child, mother, baby, cell) are removed. The output distinguishes between an identified interaction, mediator and/or moderator effect.

Test direction. Most research is based on theories that allow a prediction about the direction of the effect under study. Besides several procedures, that do not allow a direct conclusion about the direction of an observed effect (e.g., χ^2-Test, ANOVA), others can be applied to test directed hypotheses (e.g., t-test). Adjusting an undirected test to a directed test increases its power, if the sample and effect size are held constant, and the effect is present in the predicted direction.

Sentences containing a statistical result or one of several search terms (e.g., 'test', 'hypothesis') are searched by *get.test.direction()* for synonyms of one- and two-sided testing and hypothesis (e.g., directed test, undirected hypothesis). To avoid false positives for one-sidedness, sentences containing certain reference words (e.g., paper, page, pathway) are excluded and detected values less than one are omitted.

Outlier definition. Since many popular statistical measures are sensitive to extreme values (e.g., mean, variance, regression coefficients), their empirical values may not be appropriate to describe a sample. In practice, there are two popular techniques to deal with extreme values and still compute the desired statistic. Simple exclusion of outliers reduces the sample size and test power, while adjustments towards the mean preserve the original sample size. Both procedures can, of course, distort the conclusions drawn from the data because the uncertainty (variance) is artificially reduced. It is difficult to justify why valid extreme values are manipulated or removed to calculate a particular measure rather than choosing an appropriate measure (e.g., median, interquartile range). On the other hand, outliers may indicate measurement errors, that warrant special treatment. A popular measure for detecting outliers is the distance from the empirical mean, expressed in standard deviations.

get.outlier.def() identifies sentences containing a reference word of a removal process or an outlier value (e.g., outlier, extreme, remove, delete), and a number (numeric or word) followed by the term 'standard deviation' or 'sd'. Verbal representations of numbers are converted to numeric values. Since very large deviations from the mean are more likely to indicate a measurement error than an outlier definition, and to minimize erroneous extractions of overly small values, the default result space of the output is limited to values between 1 and 10.

Statistical assumptions. Any statistical procedure/model is based on mathematical assumptions about the sampling mechanism, scaling, the one and/or multidimensional distribution of covariates and the residual noise (errors). The underlying assumptions justify the statistical properties of an estimator and a test statistic (e.g., best linear unbiased estimator, distributional properties, α-error/p-value). There may be serious consequences for the validity of the conclusions drawn from these statistics, if the underlying assumptions are violated.

To extract the mentioned assumptions within an article, *get.assumption()* performs a dictionary search in the text of the methods and results sections. A total of 20 common assumptions related to the model adequacy, covariate structure, missing and sampling mechanisms can be identified (see Online Appendix C for the full list of specified search terms).

Analysis software. Statistical software solutions are a key element in modern data analysis. Some programs are specifically designed to perform certain procedures, while others focus on universality, performance, or usability.

To identify the analytic software solution mentioned in the methods and results sections, *get.software()* is used to perform a manually curated, fine-grained dictionary search of software names and their empirical representation in text. Tools for data acquisition or other data management purposes are not part of the list. However, they can be tracked down with a vector of user-defined search terms, passed to the '*add*' argument. A total of 55 different software solutions can be detected in standard mode (see Online Appendix B for the complete list of specified search terms).

Number of reported studies. Research reports may contain single or multiple study reports. To determine the total number of studies reported in an article, the section titles and abstract text are passed to *get.n.studies()*. Enumerated studies or experiments are identified, and the highest value is returned. The function returns '1' if no numbering of the studies is identified.

Methods

To evaluate the extraction capabilities of *study.character*, a manually coded dataset serves as reference data. The statistical methods used, the number of studies reported, and the software solutions used were coded by Blanca et al.[7] and provided to the author. All articles were manually rescanned for those study characteristics that are extracted by *study.character* but were not part of the original dataset.

The collection of articles by Blanca et al.[7] consists of 288 empirical studies published in eight psychological journals (British Journal of Clinical Psychology, British Journal of Educational Psychology, Developmental Psychology, European Journal of Social Psychology, Health Psychology, Journal of Experimental Psychology-Applied, Psicothema, Psychological Research) between 2016 and 2017.

The absolute frequencies of the identified statistical procedures used in the main analysis by Blanca et al.[7] are contrasted with those of *study.character*. The manually created categories of the statistical methods from Blanca et al.[7] are compared to the uncategorized statistical methods extracted using *study.character*. The search tasks for counting the frequency of articles using a specific category of procedures are implemented with regular expressions. An exploratory view of the entire result space of *get.method()* is displayed in a word cloud.

To explore the correct/false positive/negative detections by *study.character* all other extracted features are compared to the manually recoded data. A correct positive (CP) detection refers to an exact match to a manually coded feature within an article. A false positive (FP) refers to an extraction that is not part of the manually coded data. Articles that do not contain a feature and for which no feature has been detected are referred to as a correct negative (CN). Finally, a false negative (FN) refers to a feature that was not detected but was manually identified.

If a target feature is identified multiple times in an article, *study.character* will output this feature once. Therefore, the evaluation of the detection rates is carried out at the article level. Since most of the features focused on here can potentially have multiple values per article, the extractions may be fully or partially correct. This can be illustrated by the example of the extraction of the α-level. If the manual coding revealed the use of a 5% and a 10% α-level and *study.character* identifies the 5% and an unreported 1%, this is counted to be 1 correct positive, 1 false negative and 1 false positive for this article. It follows, that the number of correct (CP+CN) and total decisions (CP+FN+CN+FP) may be larger than the total number of articles analyzed.

Global descriptive quality measures (sensitivity, specificity, accuracy) are reported for every extracted feature.

Sensitivity refers to the proportion of correctly detected features within all features present (CP+FN).

$$sensitivity = CP/(CP + FN) \tag{1}$$

Specificity refers to the proportion of correct non-detections within all articles that do not contain the searched pattern (CN+FP).

$$specificity = CN/(CN + FP) \tag{2}$$

Finally, accuracy is the proportion of correct detections (CP+CN) within all existing features and non-existing features (CP+FN+CN+FP).

$$accuracy = (CP + CN)/(CP + FN + CN + FP) \tag{3}$$

Absolute frequency tables of manual and automatic detections are presented for each characteristic, and a causal association of the deviations that occurred is provided.

Data, input fomats, PDF conversion software, hardware

The 288 articles in the raw data provided by Blanca et al.[7] were manually downloaded as PDF files. The PDF files were converted to NISO-JATS encoded XML using the open-source software *CERMINE*[14], before being processed with *study.character*. Since the compilation with *CERMINE* can lead to various errors (text sectioning/structuring, non-conversion of special characters), this can be considered as a rough test condition for the evaluated functions. All processes are performed with a Dell 4-core processor running with Linux Ubuntu 20.04.1 LTS and the open-source software R 4.0[15]. To enable multicore processing, the R package *future.apply*[19] is used. The word cloud of the identified methods is drawn using the *wordcloud*[20] package.

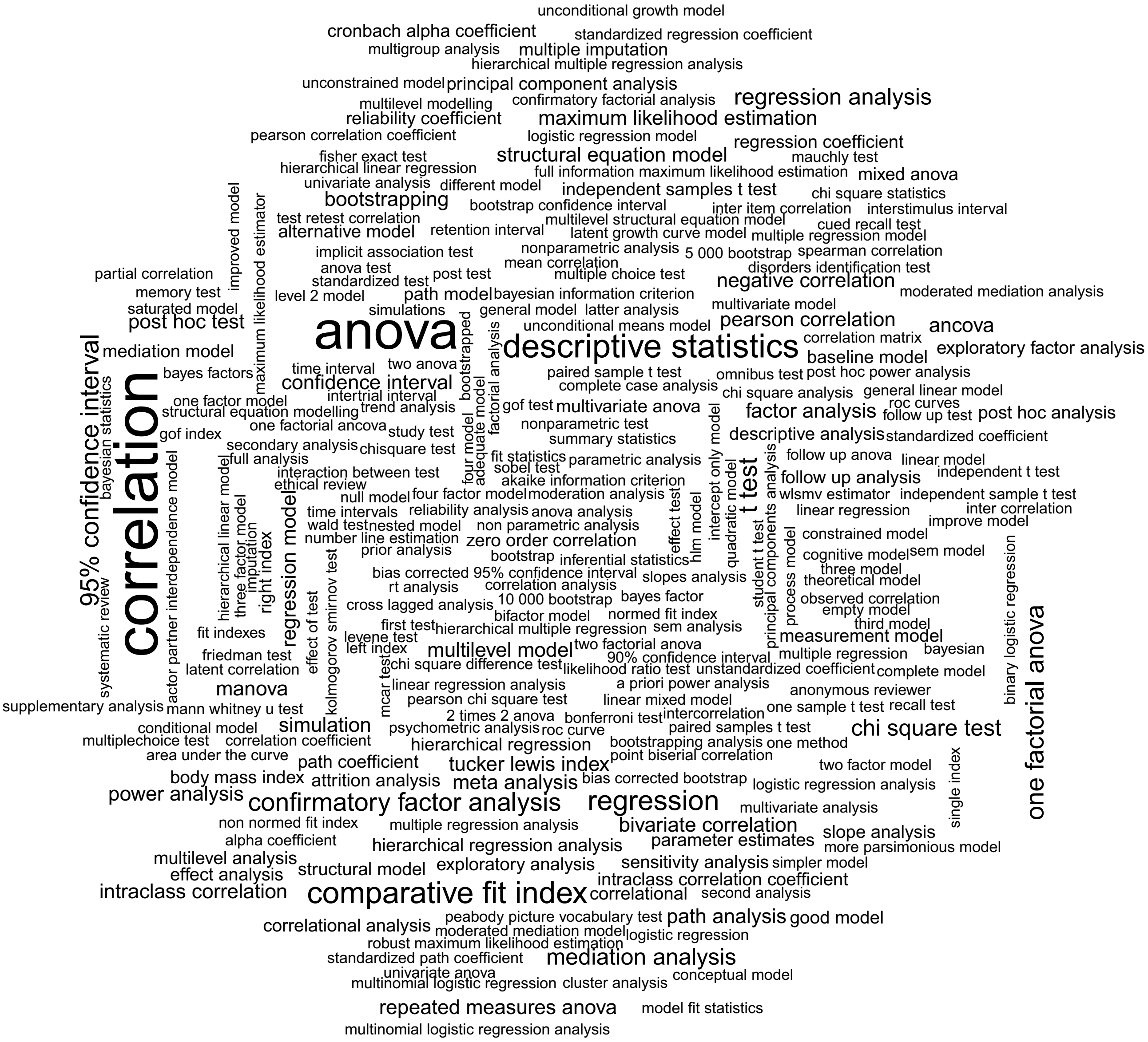


Figure 1. Word cloud of the extracted statistical methods by *study.character*.

Results

The extraction properties and causes of deviations from the manually coded study features are given in the following section for each function. A total of 287 articles are included in the analyses, as the Blanca et al.[7] data contain one article twice.

It should be noted that the extractions of statistical methods and software solutions from Blanca et al.[7] are not directly comparable to the output of *study.character* as they coded the statistical methods used in the main analyses (rather than each mention) and that are explicitly reported to be used for these main analyses.

Statistical methods.

An insight into the overall result space of the statistical methods extracted by *study.character* is given in Fig. 1, where the frequency table of the extractions is shown as a word cloud. Bigger words indicate higher frequencies. It is obvious, that correlation analysis and ANOVA are the most frequently mentioned methods in this article selection.

In order to compare the extractions of *get.method()* with the extractions of the main analysis procedure of Blanca et al.[7] the absolute frequencies of the detected studies using a specific class of methods are listed in Table 2. The regular expressions listed are used as search terms to count the hits of *get.method()* per categorized method.

Because Blanca et al.[7] coded the statistical method used in the main analysis (all methods reported in preliminary analyses or manipulation checks, footnotes, or in the participants or measures section, were not coded), most methods are more commonly identified by *get.method()*. Two rare categories cannot be identified at all with the search terms used ('correlation comparison test', 'multilevel logistic regression' ^).

The large differences in most of the identified methods (e.g., descriptive statistics, correlation, χ^2-statistics) are due to the different inclusion criteria (each mentioned method vs. method of main analysis). In addition, using

<table>
<tr><th>Statistical method</th><th>Search term</th><th>Blanca et al.</th><th>get.method()</th><th>Δ</th></tr>
<tr><td>Descriptive statistics (M; SD; OR; RR; percentages; etc.)</td><td>‘Descriptive statistics|descriptive analysis|descriptives’</td><td>15</td><td>63</td><td>48</td></tr>
<tr><td>Distribution fitting</td><td>‘Distributional analysis|distribution analysis|distribution fitting’</td><td>3</td><td>1</td><td>-2</td></tr>
<tr><td>Inter-rater agreement (Kappa; Intraclass correlation coefficient)</td><td>‘Inter rater reliability|kappa|intraclass correlation|intra class correlation’</td><td>3</td><td>30</td><td>27</td></tr>
<tr><td>Pearson’s correlation coefficient</td><td>‘Pearson correlation|pearson product|product moment|zero order correlation| ^correlation$’</td><td>54</td><td>138</td><td>84</td></tr>
<tr><td>Pearson’s correlation/regression coefficient comparison test</td><td>‘Correlation comparison|coefficient comparison’</td><td>3</td><td>0</td><td>-3</td></tr>
<tr><td>Spearman’s correlation coefficient</td><td>‘Spearman corr|spearman brown|spearmancoef |spearman rank|spearman rho’</td><td>8</td><td>12</td><td>4</td></tr>
<tr><td>Other nominal/ordinal correlation measures (gamma; Cramer’s V; Somers’ d; contingency coefficient)</td><td>‘Gamma|cramer v|somer d|contingency table analysis|contingency coef’</td><td>4</td><td>3</td><td>-1</td></tr>
<tr><td>Pearson’s Chi-square</td><td>‘^chi square|[^d] chi square’</td><td>27</td><td>52</td><td>25</td></tr>
<tr><td>Other test of contingency tables and proportion comparison (Fisher; McNemar tests; Cochran’s Q; z statistic)</td><td>‘Fisher exact|fisher z|mcnemar|cochran q|z statistic’</td><td>5</td><td>9</td><td>4</td></tr>
<tr><td>Mann-Whitney U test</td><td>‘Mann whitney|mannwhitney|u test’</td><td>5</td><td>7</td><td>2</td></tr>
<tr><td>Wilcoxon signed-rank test</td><td>‘Wilcoxon|signed rank’</td><td>4</td><td>4</td><td>0</td></tr>
<tr><td>Kruskal-Wallis test</td><td>‘Kruskal |wallis ’</td><td>5</td><td>5</td><td>0</td></tr>
<tr><td>Friedman test</td><td>‘Friedman test’</td><td>5</td><td>7</td><td>2</td></tr>
<tr><td>One sample t-test</td><td>‘One sample t test|single sample t test’</td><td>11</td><td>7</td><td>-4</td></tr>
<tr><td>Independent t-test</td><td>‘^t test$ |[^n][^e] t test|independent t test|two sample t test’</td><td>45</td><td>62</td><td>17</td></tr>
<tr><td>Paired t-test</td><td>‘Paired samples t test|paired sample t test|paired t test’</td><td>20</td><td>18</td><td>-2</td></tr>
<tr><td>ANOVA</td><td>‘Anova| ^anova’</td><td>138</td><td>156</td><td>18</td></tr>
<tr><td>ANCOVA</td><td>‘Ancova| ^ancova’</td><td>15</td><td>15</td><td>0</td></tr>
<tr><td>MANOVA/MANCOVA</td><td>‘Manova|mancova’</td><td>13</td><td>19</td><td>6</td></tr>
<tr><td>Regression analysis</td><td>‘Regression’</td><td>82</td><td>127</td><td>45</td></tr>
<tr><td>Multilevel regression</td><td>‘Multilevel.*?regression|hierarchic.*?regression|mixed.*?regression’</td><td>25</td><td>31</td><td>6</td></tr>
<tr><td>Multivariate regression</td><td>‘Multivariate.*?regres|multiple*?regres’</td><td>1</td><td>3</td><td>2</td></tr>
<tr><td>Poisson regression</td><td>‘Poisson regression’</td><td>1</td><td>1</td><td>0</td></tr>
<tr><td>Logistic regression</td><td>‘Log.*?regression’</td><td>15</td><td>28</td><td>13</td></tr>
<tr><td>Multilevel logistic regression</td><td>‘Multilevel logistic|logistic multilevel’</td><td>5</td><td>0</td><td>-5</td></tr>
<tr><td>Multinomial/ordinal regression</td><td>‘Multinom.*?regres|ordin.*?regres’</td><td>9</td><td>10</td><td>1</td></tr>
<tr><td>Generalized estimation equation for ordinal data</td><td>‘Gee|generalized estimation equation’</td><td>1</td><td>1</td><td>0</td></tr>
<tr><td>Path analysis</td><td>‘Path analysis|path model|path coefficient|path estimate|structural equation’</td><td>46</td><td>53</td><td>7</td></tr>
<tr><td>Multilevel SEM</td><td>‘Multilevel structural equation’</td><td>6</td><td>4</td><td>-2</td></tr>
<tr><td>Growth curve modeling</td><td>‘Growth curve|growth model’</td><td>10</td><td>11</td><td>1</td></tr>
<tr><td>Multilevel growth curve modeling</td><td>‘Multilevel growth|multigroup.*?growth’</td><td>2</td><td>1</td><td>-1</td></tr>
<tr><td>Confirmatory factor analysis (CFA; SEM)</td><td>‘Confirmatory factor’</td><td>21</td><td>35</td><td>14</td></tr>
<tr><td>Exploratory factor analysis (EFA)</td><td>‘Exploratory factor’</td><td>13</td><td>14</td><td>1</td></tr>
<tr><td>Cronbach’s α</td><td>‘Cronbach alpha|reliability coeff|cronbach coeff’</td><td>17</td><td>25</td><td>8</td></tr>
<tr><td>McDonald’s omega</td><td>‘Mcdonald|mc donald|omega coef|omega estimate’</td><td>4</td><td>3</td><td>-1</td></tr>
<tr><td>Test’retest reliability</td><td>‘Test retest reliability|test retest corr’</td><td>3</td><td>4</td><td>1</td></tr>
<tr><td>Convergent/Discriminant validity indexes (AVE; MSV)</td><td>‘Convergent validity|discriminant validity|[^a-z]ave |^ave |msv’</td><td>2</td><td>2</td><td>0</td></tr>
<tr><td>Sensitivity and specificity measures</td><td>‘Sensitivity|specificity|specifity’</td><td>2</td><td>14</td><td>12</td></tr>
<tr><td>Item analysis: classical test theory</td><td>‘Item analysis|items analysis’</td><td>2</td><td>4</td><td>2</td></tr>
<tr><td>Item analysis: item response theory (item calibration; DIF)</td><td>‘Item response analysis|dif analysis’</td><td>2</td><td>2</td><td>0</td></tr>
<tr><td>Cluster analysis</td><td>‘Cluster analysis’</td><td>3</td><td>4</td><td>1</td></tr>
<tr><td>ROC curve analysis</td><td>‘Roc curve|receiver operating’</td><td>7</td><td>7</td><td>0</td></tr>
<tr><td>Markov models</td><td>‘Markov|marcov’</td><td>1</td><td>1</td><td>0</td></tr>
</table>

Table 2. Number of articles with mentions of specific statistical methods extracted with *study.character* and frequency of main analytical methods reported in Blanca et al.’s Table 4.

of ‘regression’ as a search term in the output of *get.method()* also results in hits when more complex regression models were found (e.g., multilevel or multivariate regression), whereas Blanca et al.[7] consider simple regression models and more specific regression models to be disjoint.

Feature	CP	CN	FP	FN	Σ	Sensitivity	Specifity	Accuracy
α-level from CI	105	169	0	21	295	0.83	1	0.93
Max of α-level and CI (p2alpha=FALSE)	125	122	1	40	288	0.76	0.99	0.86
Max of α-level and CI (p2alpha=TRUE)*	137	120	2	28	287	0.83	0.98	0.90
Power	61	242	2	12	317	0.84	0.99	0.96
Correction for multiple testing	65	230	0	2	297	0.97	1	0.99
Outlier removal	24	262	0	1	287	0.96	1	1
Test direction	33	252	1	1	287	0.97	1	0.99
Interaction/mediator/moderator	242	87	10	22	361	0.92	0.90	0.91
Interaction (binary)	192	87	2	6	287	0.97	0.98	0.97
Assumptions	79	215	19	12	325	0.87	0.92	0.90
Assumptions (binary)	64	215	5	3	287	0.96	0.98	0.97
Software	245	105	0	8	358	0.97	1	0.98
n studies	283	0	4	4	291	0.99	0	0.97

Table 3. Sensitivity, specificity and accuracy of *study.character*'s extractions. Detections may sum up to values $> N = 287$ as results may be multidimensional and false positives are included. *Default setting of study. character.

α-level	study.character	Manual coding	False positive	False negative
0.01	1	1	0	0
0.02	1	1	0	0
0.05	89	109	0	20
0.1	14	15	0	1
Sum	105	126	0	21

Table 4. Absolute frequencies of detected α-level from 1-α confidence intervals.

Sensitivity, specificity, accuracy. Table 3 shows the sensitivity, specificity, and accuracy measures for *study.character*'s extractions based on the manually coded data. Most of the extractions work very accurately and can replace a manual coding.

Except for the α-level detection with '*p2alpha*' activated, all extractions have low false positive rates. In default mode, the empirical sensitivity of all extractions is above 0.8, the specificity above 0.9. Since there are usually very few false positive extractions, five specificity measures reach 1.

Accuracy is lowest for α-level detection (0.86 with 'p2alpha' deactivated, 0.9 in default mode) and statistical assumption extraction (0.9). The accuracy of all other extractions is above 0.9. The binarized outputs for the extracted interaction and the stated assumptions have higher accuracy than the raw extractions.

α-level. Although most of the studies examined make use of inferential statistics, only 78 (27%) explicitly report an α-level. In all cases, where no α-level is reported, the standard of $\alpha = 5\%$ is applied, but not considered an extractable feature. Since some studies report the use of multiple α-levels, the total number of detected and undetected α-levels exceeds the number of articles. Eight articles report the use of a 90% confidence interval and a 95% confidence interval.

The absolute frequency of α-levels extracted from 1-α confidence intervals by *study.character* and the manual analysis are shown in Table 4. *study.character* correctly extracts the α-value in 105 out of 126 (83%) confidence interval reports in 97 out of 118 (82%) articles. No false positive extraction is observed. Seven non-detections by *study.character* are due to *CERMINE* specific conversion errors of figure captions, 11 to the non-processing of column names and content of tables. Two reports of confidence intervals cannot be recognized due to unusual wording ('95% confidence area', 'confidence intervals set to 0.95'), one due to a report in the unprocessed discussion section.

The corrected α-level cannot be well distinguished from an uncorrected α-value. Only one out of eight corrected α-levels is correctly labeled and extracted by *study.character*, one is a false positive detection of a nominal α. The extracted nominal α-level contains three of the manually extracted corrected α-values.

For simplicity, the extracted nominal and corrected α-levels are merged with the extraction from the confidence intervals and reduced to their maximum value, which corresponds to the nominal α-level. Table 5 shows the frequency distribution of the extracted maximum α-level with the deactivated conversion of p- to α-values and the default setting.

The conversion procedure of p-values increases the accuracy of α-level extraction, but brings one additional false positive extraction, which is caused by a statistical test result reported with a standard threshold of p-values

α-level	study.character	Manual coding	False positive	False negative
0.003	0 (0)	1	0 (0)	1 (1)
0.01	2 (3)	2	1 (2)	1 (1)
0.0125	1 (1)	1	0 (0)	0 (0)
0.017	1 (1)	1	0 (0)	0 (0)
0.02	1 (1)	1	0 (0)	0 (0)
0.05	107 (119)	144	0 (0)	37 (25)
0.1	14 (14)	15	0 (0)	1 (1)
Sum	126 (139)	165	1 (2)	40 (28)

Table 5. Distribution of extracted maximum α-level with option ‘p2alpha’ deactivated and in default mode (numbers in brackets).

Power	study.character	Manual coding	False positive	False negative
(0,0.2]	5	5	0	0
(0.2,0.5]	3	3	1	1
(0.5,0.79]	12	12	0	0
(0.79,0.8]	25	24	1	0
(0.8,0.9]	6	8	0	2
(0.9,1]	12	20	0	5
> 1	0	1	0	0
Sum	63	73	2	8

Table 6. Absolute frequencies of extracted categorized test power.

($p < 0.01$). Thus, enabling the conversion of p- to α-values slightly increases the false positive rate of explicitly reported α-levels, especially for rather rarely applied levels (0.1, 0.01 and 0.001).

Test power. Since test power can be reported as both a-priori and a-posteriori results, some articles contain multiple power values. The absolute distribution of categorized power values found by *study.character* and manual coding is shown in Table 6. The evaluation of the categorized power values differs from the results in Table 3 because here, four unrecognized values in articles with several power values of the same category are evaluated as fully correct. There are two false-positive extractions caused by a poorly compiled table and a citation of Cohen's recommendation to plan studies with at least 80% power[21]. Both errors occur in documents that contain other correctly extracted power values. Overall, 61 of 73 (84%) manually coded and categorized power values are correctly extracted in 42 of 45 (93%) articles. Nine of the 12 unrecognized reports of power follow a text structure that is still unmanageable (e.g., ‘The final sample size ensured sufficient power (i.e., 0.99)’, ‘The statistical power was very high (0.99)’). This also applies to the specification of a power interval (‘with a power ranging between 0.80 and 0.90’). Here, only the first value (0.8) was extracted and considered a correct positive, while the second limit of the interval is missing and considered a false negative. One non-detection is caused by an uncompiled and unimputated Greek letter β. In addition, one erroneous report of a power value of 80 is not extracted by *study.character*, because it falls outside the defined result space [0; 1]. Further, one power value reported in an unprocessed figure caption is not detected.

Correction for multiple testing. Table 7 shows the absolute frequency of detected correction method for multiple testing by *study.character* and the manual coding. Within the collection of articles analyzed, ten of 15 detectable authors/correction methods for multiple testing are identified by *study.character* without a false positive. There are two non-identifications. One article reports a p-value correction, but not the specific method. In another article, the reported use of a ‘Bonferroni Test’ is not detected as a correction procedure, because it is not mentioned that something is corrected/adjusted with it.

Interaction effects. The distinction between moderation, mediation, and interaction effects works in 242 out of 264 mentions (92%) (see Table 3). Table 8 shows the frequency of extracted type of interaction effect by *study.character* and the manual coding.

Overall, 22 specific mentions are not recognized in 20 articles, and 10 false positive hits occurred in 10 articles. Unrecognized mentions are mostly due to reports within the non-scanned abstract, introduction, discussion, or section headings as well as simple but badly handled sentences (e.g., ‘The model also included the interactions between A and each predictor variable.’ or ‘We tested the effect of A and B, along with their interaction.’). The false positive extractions are mainly observed in studies that infer moderating/mediating effects of covariates but

Author	study.character	Manual coding	False positive	False negative
Bonferroni	38	39	0	1
Tukey	9	9	0	0
Holm	6	6	0	0
Fisher LSD	3	3	0	0
Hochberg	3	3	0	0
Scheffé	2	2	0	0
Benjamini	1	1	0	0
Duncan	1	1	0	0
Keuls	1	1	0	0
Newman	1	1	0	0
not specified	0	1	0	1
Sum	65	67	0	2

Table 7. Absolute frequencies of authors of multiple test correction procedures.

	study.character	Manual coding	False positive	False negative
Interaction	151	158	2	9
Mediator	68	70	3	5
Moderator	33	36	5	8
Sum	252	264	10	22

Table 8. Absolute frequencies of specific interaction effects.

SD	study.character	Manual coding	False positive	False negative
2	4	5	0	1
2.5	8	8	0	0
3	12	12	0	0
Sum	24	25	0	1

Table 9. Absolute frequencies of extracted outlier definition expressed in standard deviations (SD).

do not perform an explicit moderator/mediator analysis. In two articles examining mother-infant interaction and quality of interactions among peers, exclusion of target sentences fails for the term 'interaction' and causes false positive extractions.

The presence of an interaction effect can be localized very well with the binarized output (no detection vs. any detection of interaction). The presence of at least one type of interaction analysis is correctly detected in 192 of 198 (97%) articles. In total, six articles analyzing an interaction effect of variables are not identified, two detections are false positives.

Outlier definition.

The distribution of detected outlier definitions is shown in Table 9. Twenty-four out of 25 (96%) outlier definitions, expressed by standard deviations, are correctly extracted. One report of an outlier removal is not detected due to an non-compiled special character ('±'). This error does not occur when parsing the original sentence ('Twenty-nine additional infants were excluded for following reasons: ..., extreme looking times (±2 SD) ...') Because only removal processes reported with standard deviations are targeted, one outlier removal based on an interquartile range of 1.5 is not part of this analysis.

Test direction.

The absolute distribution of detected test sidedness by *study.character* and the manual coding is shown in Table 10. Thirty-three of 34 (97%) reports of a reported test direction are correctly extracted. One false positive hit is observed in an article dealing with 'one-sided aggression' One report of a two tailed test setting is not detected in a sentence about power considerations ('...a difference of $d = 0.4$ ($\eta^2 = 0.04$; two tails, $\alpha = 0.05$)...'), because 'two tails' is not defined as inclusion pattern.

Statistical assumptions.

Seventy-nine of 91 manually coded assumptions (87%) are extracted correctly (see Table 3). Extraction of specific assumptions results in more false positive (19) than false negative (12) detections. Both, the false positive and negative detections mainly concern the very general assumptions of normally distributed variables, independence of measurements and linearity of relationships. More specific assumptions

Test direction/s	study.character	Manual coding	False positive	False negative
One and two sided	2	2	0	0
One sided	6	5	1	0
Two sided	26	27	0	1
Sum	34	34	1	1

Table 10. Absolute detections of test direction/s.

Assumption	study.character	Manual coding	False positive	False negative
Normal distribution	23	19	6	2
Sphericity	18	18	1	1
Missing at random	14	15	0	1
Missing completely at random	8	8	0	0
Independency	7	5	3	1
Linearity	7	5	4	2
Multivariate normal	5	5	0	0
Homogeneity of variances	4	2	2	0
Multicollinearity	4	5	0	1
Equal variances	3	2	1	0
Homoscedasticity	2	2	0	0
Autocorrelation	1	1	0	0
Proportional hazards	1	0	1	0
Proportional odds	1	0	1	0
Homogeneity	0	1	0	1
Homogeneity of covariance matrices	0	1	0	1
Nonproportionality	0	1	0	1
Sampling adequacy	0	1	0	1
Sum	98	91	19	12

Table 11. Absolute frequencies of detected statistical assumptions.

are extracted very accurately. Four manually extracted assumptions are missed because they are not part of the result space of *get.assumption()* (homogeneity of covariance matrices, sampling adequacy, non proportionality) or are too unspecific (homogeneity). For one article containing the manually coded assumption of nonproportionality, *study.character* outputs the proportional hazards and proportional odds assumptions, which is appropriate. The absolute frequency distribution of the extracted assumptions by *study.character* and the manual coding is shown in Table 11.

Analysis software. Compared to the manually recoded software mentions, the dictionary search tasks work very accurately. Eight software mentions are missed, no false positive extraction is observed. In total, *get.software()* identifies 245 usages of 23 different software solutions in 181 articles. This is significantly more than reported by Blanca et al.[7] (180 uses of 13 software programs in 155 articles). The absolute frequencies of studies explicitly reporting the use of a statistical software to perform the main analysis coded by Blanca et al.[7] and all extracted software mentions by the *get.software()* function of *study.character* are listed in Table 12.

A total of six discrepancies to the recoded data of software mentions are due to extraction errors of *study.character* involving non-captures of G*Power (2), the tool PROCESS MACRO (2), MPlus (1) and Amos (1). Compared to the analysis of Blanca et al.[7], with the exception of AMOS and SPSS PROCESS MACRO, all other software solutions are identified more frequently or as frequently by *study.character*. A comparison with the manual coding shows that most of the discrepancies (47) are due to different inclusion criteria (mentioned vs. explicitly stated software for the main analysis). Since G*Power is never used to perform the main calculations of an analysis, it is not extracted by Blanca et al.[7]. All mentions of Matlab (13) extracted with *get.software()* are not included in the data of Blanca et al.[7], since it was only indicated that it was used to design the computerized experiment, with no explicit indication of its use as analysis software. In eight cases, neither Blanca et al.[7] nor *study.character* extracted the specifications of comparatively rarely used software solutions (e.g., ConQuest 2.0, ASY-DIF), because they were outside their area of interest, or results space. Nevertheless, these software solutions could be easily detected by *get.software()* by adding these names to the 'add' argument. Some tools such as Omega, ChiSquareDif or Excel extracted by *get.software()* were outside the domain of interest of Blanca et al.[7].

Software	study. character	Manual coding	Blanca et al.
SPSS	71	71	52
Mplus	45	45	44
PROCESS MACRO	25	25	26
AMOS	18	18	19
G*Power	14	16	0
MATLAB	13	13	0
Stata	11	11	8
SAS	10	10	6
R	9	9	7
HLM	5	5	0
EQS	3	3	3
JASP	3	3	0
Excel	3	3	0
LISREL	2	2	2
FACTOR	2	2	2
Python	2	2	0
Sum	238	236	169
Psychtoolbox	2	2	0
AcqKnowledge	2	2	0
MLwiN	1	1	1
Warp PLS	1	1	0
Statistica	1	1	0
Praat	1	1	0
ImageJ	1	1	0
TimeStudio	1	0	0
Rmediation Webapp	1	0	0
Omega	1	0	0
ConQuest 2.0	1	0	0
ChiSquareDif	1	0	0
ASY DIF	1	0	0
HTLM	0	0	6
other	0	0	4
Sum	15	9	11

Table 12. Absolute frequencies of detected software solutions by *study.character*, the manually coded data, and the explicitly stated software solution used for the main analyis reported in Blanca et al.'s Table 7.

N studies:	1	2	3	4	5	6
study.character	205	36	24	15	6	1
Manual coding	202	37	26	15	6	1
Blanca et al.	200	39	29	15	4	0

Table 13. Absolute frequencies of extracted number of studies per paper by *study.character*, the manually coded data and Blanca et al.'s Table 2.

Number of studies reported per paper. In 283 articles (98.6%), the number of contained studies is correctly extracted by *study.character* (see Table 3). Since a numeric value is returned in each case, the function has no specificity. The output can only be right or wrong, and never false negative. Compared to the manually recoded data, there are four false detections by *study.character*. All are due to missing section names in the compiled XML file. In one case, only some relevant section names were not compiled by *CERMINE*, resulting in an output of '2' instead of '3' studies. In three cases, the missing section names result in the default output of '1' study. The full distribution of the number of reported studies per paper extracted by *study.character*, the manual recoding and Blanca's analysis are shown in Table 13.

In contrast to the manually coded data of Blanca et al.[7] (Table 2 in the original article), there are seven deviations from the recoded number of studies per article. Some differences can be explained by different coding approaches. Two control tasks in one study were treated as two individual studies by Blanca et al.[7], while *study.*

character does not consider them as a study. One study with two sub-studies in a single study was coded as two studies by Blanca et al.[7]. No error assignment can be made for the remaining five discrepancies. These include an article with six studies coded as an article with three studies by Blanca et al.[7].

Discussion

The accuracy analysis of the study feature extraction presented here contributes to the establishment of *JATSdecoder* as a useful tool for science research. In addition to the extraction algorithms for statistical results that have already been evaluated[18], key methodological features of studies can now be extracted in a rubust way, opening up a wide range of new possibilities for research on research projects.

JATSdecoder's module *study.character* can facilitate meta-research and identification tasks on methodological study features of scientific articles, that are written in English. The primary focus on NISO-JATS encoded content and the purely heuristics-based approach distinguish *JATSdecoder* from most other text processing packages. Nevertheless, the implemented extraction heuristics exhibit high accuracy and can replace overly time-consuming human coding. Moreover, *study.character* facilitates the monitoring of methodological research practices in large text databases. Such analysis can be performed for one or many journals, authors, subjects, and/or disciplines when the output of *study.character* is combined with the metadata extracted by *JATSdecoder*.

Although *study.character* cannot distinguish between a statistical method and other text phrases that are not statistical methods (e.g., 'IQ test'), its output facilitates research on research that is focused on the application and distinction of statistical methods in practice. A reduction of dimensionality by search terms enables rapid method-specific identification of studies and mapping of developments in scientific research practice.

However, some aspects should be considered in an individual or overall analysis of studies with *study.character*. Compared to Blanca et al.[7], who extracted the main statistical method used and the explicitly stated software used to perform the main calculations, *study.character* outputs all mentioned methods and software solutions within the methods and results sections.

A key feature of *study.character* is the selection of section-specific text parts, which only works for NISO-JATS encoded XML files. The very low false positive rates indicate that the text selection and exception handling work properly. Nevertheless, applying the *study.character* functions to any unstructured plain text is possible, but may lead to more false positive extractions, if discussions and reference lists are also processed.

The evaluation was done on psychological studies only. Whether the high level of precision can be achieved in other scientific disciplines, thus allowing a general analysis of scientific procedures in other disciplines, remains to be answered by future research.

There are other interesting study features that are not yet extracted by *study.character*. For example, an identification by study design (e.g., experimental, observational), study type (e.g., randomized treatment control, placebo waitlist control), or measurement tools (e.g., questionnaire, EEG, DNA-sequencing) might be of interest in a meta-analysis. Since all of these features are high-dimensional when considered across a broad range of scientific practice, consideration should be given to developing more sophisticated natural language processing tools that can address this issue. In any case, it should be carefully investigated whether model-based text extraction tools (e.g., Named Entity Recognition) can outperform both methods. One aspect that will be complicated when using these methods is assigning the cause of incorrect extractions.

A web application enabling the analysis and selection of the extracted metadata and study characteristics of content within the PMC database is provided at: https://www.scianalyzer.com. The handling is simple and allows even inexperienced users to make use of the *JATSdecoder* extractions and perform individual analysis and search tasks. The raw data of searches with less than 20,000 results can be downloaded for further processing.

Since *JATSdecoder* is a modular software, externally developed extraction functions can be easily implemented. Collaboration on *JATSdecoder* is very welcome to further improve and extend the functionality. It can be initiated via the GitHub account: https://github.com/ingmarboeschen/JATSdecoder.

Data availability

An interactive web application for analyzing study characteristics and identifying articles linked in the PubMed Central database is accessible at: https://www.scianalyzer.com.

Code availability

JATSdecoder software is freely available at: https://cran.r-project.org/package=JATSdecoder; https://github.com/ingmarboeschen/JATSdecoder. Scripts to reproduce this and other analyses performed with *JATSdecoder* are stored at: https://github.com/ingmarboeschen/JATSdecoderEvaluation.

Received: 20 December 2021; Accepted: 26 December 2022
Published online: 04 January 2023

References

1. Cohen, J. The statistical power of abnormal-social psychological research: A review. *Psychol. Sci. Public Interest* **65**, 145–153. https://doi.org/10.1037/h0045186 (1962).
2. Reis, H. T. & Stiller, J. Publication Trends in JPSP: A three-decade review. *Pers. Soc. Psychol. Bull.* **18**, 465–472. https://doi.org/10.1177/0146167292184011 (1992).
3. Schinka, J. A., LaLone, L. & Broeckel, J. A. Statistical methods in personality assessment research. *J. Pers. Assess.* **68**, 487–496. https://doi.org/10.1207/s15327752jpa6803_2 (1997).
4. Bangert, A. W. & Baumberger, J. P. Research and statistical techniques used in the Journal of Counseling & Development: 1990–2001. *J. Counsel. Dev.* **83**, 480–487. https://doi.org/10.1002/j.1556-6678.2005.tb00369.x (2005).
5. Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M. & Depaoli, S. A systematic review of Bayesian articles in psychology: The last 25 years. *Psychol. Methods* **22**, 217–239. https://doi.org/10.1037/met0000100 (2017).

6. Anderlucci, L., Montanari, A. & Viroli, C. The Importance of Being Clustered: Uncluttering the Trends of Statistics from 1970 to 2015. arXiv preprint arXiv:1709.03563 (2017).
7. Blanca, M. J., Alarcón, R. & Bono, R. Current practices in data analysis procedures in psychology: What has changed? *Front. Psychol.* **9**. https://doi.org/10.3389/fpsyg.2018.02558 (2018).
8. Zheng, S., Dharssi, S., Wu, M., Li, J. & Lu, Z. Text mining for drug discovery. *Methods Mol. Biol. (Clifton, NJ)* **1939**, 231–252. https://doi.org/10.1007/978-1-4939-9089-4_13 (2019).
9. Bird, S., Klein, E. & Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit* (O'Reilly Media, Inc., 2009).
10. Honnibal, M. & Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017) (to appear).
11. Böschen, I. *JATSdecoder: A Metadata and Text Extraction and Manipulation Tool Set* (2022). https://CRAN.R-project.org/package=JATSdecoder. R package version 1.1.
12. PubMed-Central. PMC Overview. Accessed: 2021-12-20. https://www.ncbi.nlm.nih.gov/pmc/about/intro (2020).
13. National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM). Journal Publishing Tag Library - NISO JATS Draft Version 1.1d2. https://jats.nlm.nih.gov/publishing/tag-library/1.1d2/index.html (2014).
14. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J. & Bolikowski, Ł. CERMINE: automatic extraction of structured metadata from scientific literature. *Int. J. Doc. Anal. Recognit. (IJDAR)* **18**, 317–335. https://doi.org/10.1007/s10032-015-0249-8 (2015).
15. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. (2020).
16. Böschen, I. Software review: The JATSdecoder package - extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed Central's open access database. *Scientometrics* **126**, 9585–9601. https://doi.org/10.1007/s11192-021-04162-z (2021).
17. Epskamp, S. & Nuijten, M. B. statcheck: Extract statistics from articles and recompute p values. R package version 1.3.0. https://CRAN.R-project.org/package=statcheck (2018).
18. Böschen, I. Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. *Sci. Rep.* **11**. https://doi.org/10.1038/s41598-021-98782-3 (2021).
19. Bengtsson, H. future.apply: Apply function to elements in parallel using futures. R package version 1.4.0. https://CRAN.R-project.org/package=future.apply (2020).
20. Fellows, I. wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud (2018).
21. Cohen, J. *Statistical power analysis for the behavioral sciences* (Erlbaum, Hillsdale, NJ, 1988). https://doi.org/10.4324/9780203771587.

Acknowledgements

I would like to thank María J. Blanca for her cooperation in providing the data of her original research. I also thank Marcella Dudley very much for the linguistic revision of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This research was financed by a doctoral grant awarded by the Department of Psychological Methods and Statistics, Institute of Psychology, University Hamburg, Germany. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The author declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-27085-y.

Correspondence and requests for materials should be addressed to I.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Anhang D: Publikation 4

Böschen, I. (accepted in March 2023). Changes in methodological study characteristics in psychology between 2010 – 2021. PLoS One. https://doi.org/10.1371/journal.pone.0283353.

PLOS ONE

RESEARCH ARTICLE

Changes in methodological study characteristics in psychology between 2010-2021

Ingmar Böschen *

Department of Research Methods and Statistics, Institute of Psychology, University Hamburg, Hamburg, Germany

* ingmar.boeschen@uni-hamburg.de

Abstract

In 2015, the Open Science Collaboration repeated a series of 100 psychological experiments. Since a considerable part of these replications could not confirm the original effects and some of them pointed in the opposite direction, psychological research is said to lack reproducibility. Several general criticisms can explain this finding, such as the standardized use of undirected nil-null hypothesis tests, samples being too small and selective, lack of corrections for multiple testing, but also some widespread questionable research practices and incentives to publish positive results only. A selection of 57,909 articles from 12 renowned journals is processed with the *JATSdecoder* software to analyze the extent to which several empirical research practices in psychology have changed over the past 12 years. To identify journal- and time-specific changes, the relative use of statistics based on p-values, the number of reported p-values per paper, the relative use of confidence intervals, directed tests, power analysis, Bayesian procedures, non-standard *α* levels, correction procedures for multiple testing, and median sample sizes are analyzed for articles published between 2010 and 2015 and after 2015, and in more detail for every included journal and year of publication. In addition, the origin of authorships is analyzed over time. Compared to articles that were published in and before 2015, the median number of reported p-values per article has decreased from 14 to 12, whereas the median proportion of significant p-values per article remained constant at 69%. While reports of effect sizes and confidence intervals have increased, the *α* level is usually set to the default value of .05. The use of corrections for multiple testing has decreased. Although uncommon in each case (4% in total), directed testing is used less frequently, while Bayesian inference has become more common after 2015. The overall median estimated sample size has increased from 105 to 190.

OPEN ACCESS

Citation: Böschen I (2023) Changes in methodological study characteristics in psychology between 2010-2021. PLoS ONE 18(5): e0283353. https://doi.org/10.1371/journal.pone.0283353

Editor: Enrico Toffalini, University of Padova, ITALY

Received: May 19, 2022

Accepted: March 7, 2023

Published: May 10, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0283353

Data Availability Statement: The data and script to reproduce the analysis are available at: dx.doi.org/10.17504/protocols.io.j8nlkk775l5r/v1.

Funding: This research was financed by a doctoral grant awarded by the Department of Research Methods and Statistics, Institute of Psychology, University Hamburg, Germany. The funders had no

Introduction

For decades, many scientists have been voicing well-founded criticism about the general application of statistical methods, specifically for the social sciences. Their critique points to several widespread problems such as the ritualized application of significance tests on nil-null

role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The author has declared that no competing interests exist.

hypothesis [1–5], the arbitrary selection of mostly small [6, 7] or selective samples [8, 9], the standardized or even wrong interpretation of results [10, 11] or questionable research practices to find the desired (significant) results [12].

An indication of the practical relevance of these concerns in the field of psychology is provided by the large-scaled replication study of the Open Science Collaboration (OSC) [13], which gives empirical evidence for the so-called *Replication Crisis in Psychology*. The OSC conducted 100 high-powered replications of psychological experiments (average power of .92), but only a disappointingly small part lead to the same results as the original studies. The overall replication rates ranged between 36% and 68%, depending on the definition of a replicated result and varied greatly between journals.

The highest replication rates were observed for cognitive studies published by *Psychological Science*. Fifty-three per cent of the replicated experiments from this journal lead to effects of the original direction and p-values $< .05$. Also, 53% were subjectively considered a successful replication and 92% reached meta analytical p-values $< .05$. One qualification about this result is the possibility that the original studies have inflated effect sizes due to publication, selection, reporting, or other biases [13].

The widespread problematic or even flawed application of statistical inference and conclusions made therewith, led Ioannidis to the general, pessimistic conclusion that *"most published research findings are false"* [14]. In the light of the OSC study, however, this statement appears to be exaggerated for psychology, but it points to severe credibility problems.

Journals have developed some general and individual quality standards intended to increase the reliability of their publications. For quite some time, peer review has been an important standard in quality assurance and is undoubtedly irreplaceable for a credible science. However, the replication crisis was not prevented by the peer-review process. Other journal requirements designed to ensure research quality vary from journal to journal and have and will continue to evolve over time. These requirements mostly relate to justification of sample size, precision of description of data preparation, reporting of statistical test results, reporting of effect sizes and confidence intervals, publication of evaluation protocols and data, preregistration of research projects. Taken together, the standards set by journals have a strong influence on how scientists plan and conduct their studies and how they report and interpret their findings. However, in many cases, even these so-called 'best practices' have apparently not been sufficient to ensure the replicability of a study result in the past.

What is going wrong? Possible causes of the replication crisis

Despite many great scientific achievements, something seems to be going wrong in some areas of scientific reasoning. The here listed critique is not new, but still of major relevance for psychological research practice, as the low replication rates in the OSC study suggest. One of the most important issues is certainly the overly ritualized use of undirected nil-null hypothesis significance testing.

Most frequentist methods have in common that they are based on assumptions about the sampling process and the distributional properties of involved variables. Based on these assumptions and a given null hypotheses (H_0), the theoretical distribution of an estimator $\hat{\theta}$ and its variance (inaccuracy) $\hat{\sigma}^2_{\hat{\theta}}$ is determined. The realized estimator $\hat{\theta}$ can then be checked for random deviation from the theoretical value θ_0, which was specified in H_0. It should be emphasized that, unlike in practice, the empirical effect estimator $\hat{\theta}$ can be tested against any value of θ_0. θ_0 can be the difference in means of two groups, a correlation, model fit index, or any other statistical measure of interest.

Predefined critical values of the determined test statistic serve as a threshold for a binary decision about θ_0. If the test statistic falls within the rejection region, the H_0 is rejected. Otherwise H_0 is not rejected, which is not to be confused with proved or accepted.

Low power = small samples = high uncertainty

Undoubtedly, the replication crisis is a symptom of a situation of great uncertainty that can be caused by low statistical power, noisy measurements and/or questionable research practices. An obvious way to reduce uncertainty in scientific contexts is to increase the measurement accuracy of the estimated parameters θ. The test strength or power ($1 - \beta$ error) of a statistical test is the theoretical probability of correctly rejecting the null hypothesis—getting a significant result—with an a priori set α error and planned or realized sample size N, assuming an effect of a certain size $\delta_{true} = \theta_{true} - \theta_0$. Conversely, this idea is suitable for estimating the required minimum sample size in order to discover effects (deviations from the H_0) of a certain minimum size δ_0, or to estimate the optimal sample size, to achieve a power or length of a $1 - \alpha$ confidence interval, with an adequate α level.

A power analysis for a two-sided comparison test of means (T-test) with the R function *power.t.test()* yields an optimal size of $n_{opt} > 99.08$ units per group to identify a medium-sized effect of $\delta = .4$ standard deviations, with an α level of .05 and a power of .8. Studies that focus on smaller effects or work with lower α levels require much larger samples in order to identify the effects (get a significant result) with the same power.

In 1962, Cohen found that most of the studies published in the *Journal of Abnormal and Social Psychology* (46%) had low power ($< .5$). When one posits medium effects in the population (Cohen used $\delta = .5$ sd) the studies have, on average, a slightly less than 50-50 chance of successfully rejecting their major null hypotheses [6]. None of the investigated studies had a power $> .5$, when considering small effects ($\delta = .25$ sd).

Again, 24 years later Sedlmeier & Gigerenzer [7] examined the power of psychological studies of several journals and found that the median power had decreased even further. Among other critics, Gelman & Carlin [15] show that low-powered studies have fundamental problems and introduce the terms type M (magnitude) and type S (sign) error. *"When researchers use small samples and noisy measurements to study small effects—as they often do in psychology as well as other disciplines—a significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an effect"* [15]. The practical relevance of these error types is supported by the OSC report, in which the replication effects were half the size of the original effects and some effects pointed in the opposite direction of the original effects.

One approach to sample size planning is to estimate the desired (in-) accuracy of the parameter to be estimated and use variance estimates, for example from previous studies or standard values. A researcher might want to estimate the effect of a phenomenon with a 95% confidence interval that has an accuracy of ± 3.5 units (points, milliseconds, etc.).

Further, focusing on the β error (not rejecting H_0, although it is false) or test power enables the determination of a necessary or even optimal sample size to identify the effect sizes of interest. This approach also permits an interpretation of insignificant results as an indication (not as evidence) for the absence of an effect of a certain size, if the power of the test is high.

The undirected nil-null test scenario

A common misconception about the null hypothesis test in psychology and the social sciences is that the null hypothesis is to be equated with a zero effect ($\theta_0 = 0$), or zero correlation [1, 10, 16]. This type of hypothesis represents a special case and is also known as the nil-null hypothesis, the nil-hypothesis [3] or the point null hypothesis [2]. The term nil-null refers to a test setting that hypothesizes a true null effect or zero correlation.

Since two or more groups of units are never completely alike and most statistical estimates $\hat{\theta}$ are weakly consistent estimates, meaning that their variance/inaccuracy $\sigma^2_{\hat{\theta}}$ converges to 0 when the number of observations (N) grows to infinity ($\lim\limits_{n \to \infty} \sigma_{\hat{\theta}} = 0$), Cohen points out that very large samples always lead to the rejection of any precise point-zero hypotheses [17]. Twenty-eight years after his first study on the power of psychological studies [6], Cohen reiterated his recommendation to psychologists, that they should include thoughts about the direction and strength of the effect to be discovered, when planning a study [17]. *"The null hypothesis can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false)"* [17].

This position coincides with Meehl's paradox [1] of the opposing application scenarios of the null hypothesis significance test and the handling of uncertainty, by psychologists and physicists. *"While physicists generate a specific prediction about their model and test whether the measurement results deviate significantly from the predictions, psychologists usually test their data for a zero effect and this in addition mostly in an undirected manner in order to underpin their model with significant results. Further, it had been more explicitly recognized that what is of theoretical interest is not the mere presence of a difference (i.e., that H_0 is false) but rather the presence of a difference in a certain direction"* [1].

Psychologists seem to be sceptical about generating a priori set theoretical point estimates or borders of interest apart from zero. This surely limits them to use uninformed nil-null hypothesis tests and seek for a significant result, which is then assessed as support for their theory.

It is common practice to report the p-value of a statistical test result. The p-value is the probability of obtaining a value $\hat{\theta}$ or more extreme deviations from θ_0, when H_0 is in fact true. The p-value is compared with a previously defined α level (the probability of incorrectly rejecting H_0) to make a binary decision. If a p-value is smaller than the α level, the test result is reported as significant, if $p > \alpha$ the H_0 is not rejected. In practice the α level is mostly set to the undefined standard of .05.

Another frequentist approach was developed by Neyman [18] three decades before Meehl's criticism. Neyman introduced the $1 - \alpha$ confidence interval for a parameter θ. Cox & Hinkley [19] describe confidence intervals in terms of a random sample. *"Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward $1 - \alpha\%$"*. Besides the additional information about uncertainty, it includes a test on a statistical null hypothesis. If the $1 - \alpha$ confidence interval contains the value of θ_0, the test statistic leads to a p-value $> \alpha$. If θ_0 is not included, the resulting p-value is below α, what is called a significant deviation from H_0. In fact this simple check does not add any information to the p-value based procedure. Rather, the empirical effect and the width of the interval should be in focus.

In 1996, Schmidt [20] calls for the replacement of significance tests by point estimates and confidence intervals, as this approach leads to much more informative results. Although confidence intervals have been around for a long time, in 2014 Cumming [21] still encourages an increased use of effect estimates combined with confidence intervals and a focus on their width, which he jovially called 'The New Statistics'.

A rather seldom proclaimed procedure are equivalence tests. Equivalence tests enable to falsify predictions about the presence, and declare the absence, of meaningful effects [22]. This type of testing differs from the standard method in that interval hypotheses are tested and support for the null hypothesis can be generated. In principle, equivalence tests demand an a priori set definition of trivially small, meaningless effects. An undirected hypothesis is then tested with two one-sided tests against the bounds of the predefined interval of meaningless effects. Non-inferiority and superiority trials are another type of interval null hypotheses primarily used in medical treatment evaluation. The null hypothesis in non-inferiority trials states that a new treatment is worse than an old by more than $-\Delta$, where $-\Delta$ is the 'non-inferiority margin' [23].

Although it is the most commonly applied threshold, several authors have expressed their doubts about the usefulness of a fixed alpha error of .05 for the discovery of new effects and have proposed to lower it to .005 [24–26]. These and several other authors [21, 27–29] have even proposed to move away from frequentist methods entirely and to use Bayesian statistics instead (e.g. Bayes factor, posterior distribution). Compared to standard null hypothesis testing, Bayesian methods update prior assumptions about an effect or the probability of a hypothesis with the incoming data. This approach allows more intuitive statements about the probability of a hypothesis given the data, or the ratio of the probabilities of competing hypotheses given the data. For a detailed comparison of the frequentist and Bayesian approach, the reader is referred to the literature.

As a consequence of the ongoing debate about the utility of p-values, the editors of the journal *Basic and Applied Social Psychology* have banned p-values and hypothesis tests in 2015 [30]. However, according to Lakens [31], this change has not been accompanied by a shift toward Bayesian methods nor confidence intervals and has tended to degrade the quality of studies published in the journal as measurement inaccuracy is vanished.

Questionable research practices

Questionable research practices (QRPs) that manipulate the data or the analytical process, until the desired effects are found, are widespread [12, 32]. After listing several potential ways to manipulate the data collection and analytical processes, Simmons et al. conclude that in many cases it is more likely that a researcher will find false evidence for the presence of an effect than that he/she will find evidence that there is no effect [12].

For example, optional stopping refers to a practice that is applied if a result is not significant. Additional cases are drawn until the desired result has been achieved, without reporting or correcting for this sampling procedure (e.g. sequential testing). Further, unplanned subgroup comparisons can easily be carried out, especially in large data sets with many potential dependent variables, and only the significant effects are reported later. HARKing (Hypothesizing After Results are Known) is a special form of this approach. HARKing refers to the report of a post hoc hypothesis, as if it were actually an a priori hypothesis [33]. Further, post hoc exclusion and inclusion criteria can be defined until the results meet one or many specific requirements. The file drawer problem, which was introduced by Rosenthal [34], refers to the fact that mostly studies that 'worked' are brought to publication and that there is therefore a bias towards 'positive' results in the literature.

In 2012, John et al. surveyed over 2,000 American psychologists and found that the percentage of respondents who reported to have engaged in QRPs was surprisingly high [35]. For certain practices, their inferred actual estimates approached 100%, which suggests that these practices may constitute the de facto scientific norm [35]. It should be noted, that the survey did not ask how often these practices were used, only whether they were ever used at all. Ten years later, Fox et al. estimated that 18% of American psychologists have used at least one QRP in the past 12 month and that QRP users are a stigmatized sub-population of psychologists [36].

An effective solution to minimize the use of HARKing and optional stopping is to determine the research questions and analysis plan before observing the research results—a process called preregistration [37]. However, it remains questionable whether preregistration also counteracts the file drawer problem, since authors and editors can continue to make the publication dependent on the result. In registered reports a study protocol containing the hypotheses, planned methods, and analysis pipeline undergoes peer review before the data collection and the results are published regardless of whether the hypotheses are supported or not [38]. By analyzing the first hypothesis of standard and registered reports, Scheel et al. show that the latter publication process leads to significantly lower success rates (96% positive results vs. 44% positive results in registered reports) and therefore reduces publication bias and/or α error inflation [38].

Steegen [39] proposed multiverse analyzes as an approach to coping with the researcher's degrees of freedom within an analytical process. A multiverse analysis displays the stability or robustness of a finding, not only across different options for exclusion criteria, but across different options for all steps in data processing [39].

Multiple testing

Since the probability of receiving at least one false positive (significant) test result increases with each statistical test performed on the same data (α error accumulation), the α level or p-values should usually be adjusted/corrected for multiple tests of the same family (family-wise error rate). Thus, in the absence of strong a priori expectations about the tests that are relevant, this alpha inflation can be substantial and be a cause for concern [40]. It is rarely emphasized that this also applies to the reporting of multiple 1-α confidence intervals. A simple, rigid/conservative method for an α level/p-value adjustment is the Bonferroni correction, in which the underlying α level/empirical p-values are corrected with the total number of tests carried out. There are various other, more powerful methods for adjusting the α level/p-values for multiple testing, such as the sequential method by Holm [41] or Benjamini & Hochberg [42]. If a large number of statistical tests is carried out on the same data, with constant effects and corrected α levels, then the sample sizes must be increased in each case in order to discover the effects with the same power. Another technique to control the α error in multi group comparisons is partial pooling, which is applied in multilevel models with undiminished power [43].

Selective samples

Although sample characteristics are rather unrelated to the low replication rates of the OSC report, sampling may be a critical issue in drawing valid conclusions about humanity. Henrich et al. point out that psychological research is mostly based on samples from and written by authors from western educated industrial rich and democratic (WEIRD) societies [9]. Their analysis of articles by the top journals in six sub-disciplines of psychology from 2003 to 2007 shows that although WEIRD residents represent only a small proportion of the world population (12%), 73% of first authors were at American universities, and 99% were at universities in Western countries [9]. Annett examined studies published in the *Journal of Personality and Social Psychology* in 2007 and found that in 67% of U.S. and 80% of non-U.S. studies, the samples consisted of undergraduate psychology students [8], which is certainly an even more specific kind of humanity.

Method

The extent to which some of the methodological features highlighted here have changed over the past 12 years is examined using a collection of 57,909 psychological research articles from 12 journals.

The R [44] package *JATSdecoder* [45] is used to extract the methodological study features in focus here. *JATSdecoder* is a general toolbox which facilitates text extraction and analytical tasks on NISO-JATS coded XML documents [46]. *JATSdecoder* consists of two main modules. The *JATSdecoder()* function extracts metadata such as the title, abstract, author, keywords, country of origin, reference list and the main text, which is stored as a vector of sections. The *study.character()* function mainly processes the text of the methods and results sections and extracts important study characteristics using expert-guided heuristics. It has been demonstrated that *study.character()* reliably extracts statistical results reported in text [47] and methodological study features such as the statistical methods applied, the α level and correction procedures for multiple testing, a priori and a posteriori power and the test sidedness [48]. For example, the statistical methods extraction algorithm uses an n-gram bag of words approach for a set of pre-specified inclusion terms (e.g., test, interval, bayes, analysis). This allows for accurate identification of most statistical procedures while keeping the result space open. The underlying sample size is estimated based on reports within the abstract (textual and numerical representations) and by a user-definable quantile of all extracted degrees of freedom, that are detected within the reported statistical test results within the text. The .9-quantile is used here to reduce overestimation by reports that mostly contain subgroup analyses or mishandling of repeated measures as independent units. It should be emphasized that the latter function has not yet been evaluated and is still in the developmental stage.

All research articles published between 2010 and 2022 by ten highly renown journals (*Behavioral Neuroscience*, *Depression & Anxiety*, *J. Abnormal Psychology*, *J. Child Psychology*, *J. Family Psychology*, *J. Management*, *Personality and Social Psychology Bulletin*, *Psychological Medicine*, *Psychology & Aging*, *Psychophysiology*) of five psychological subdisciplines (Biological, Clinical, Developmental, Social, Work and Organizational Psychology) were downloaded manually with the license of the University of Hamburg as a PDF version. The Content ExtRactor and MINEr *CERMINE* [49] was then used to convert the original PDF files into XML files, structured in the *Journal Article Tag Suite* standard markup (NISO-JATS) [50]. In order to further include articles published in open access journals in the analysis, research articles from *Frontiers in Psychology* and *Plos One* were selected. The full PubMed Central (PMC) database was downloaded in bulk from https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/ in XML format in January 2022 (4,048,608 files), processed with *JATSdecoder* [45], and reduced to articles that were published by *PLoS One* and *Frontiers in Psychology* between 2010 and 2021 and type tagged as 'research-article'. Articles on psychology from *Plos One* were selected using a search task with the pattern 'psycholog' on the lowerized title, abstract, keyword, subject and affiliation tags.

Articles that contain any of the search terms 'meta-analysis', 'review', 'letter to the editor', 'corrigendum' or 'commentary' in their title and that do not contain any statistical result in the main text are removed in order to select only empirical research articles for the analysis.

Global and journal-specific trends and changes in study characteristics are analyzed over time with conditional frequency tables and bar plots. The distribution of the number of reported p-values per journal is shown in a box plot.

In terms of reporting style, a distinction is made between the results as follows. While extractable results with p-value contain a p-value pointing to a number with an operator (e.g.: $p < .05$), computable p-values can also be extracted from incompletely reported results (e.g.: t

(32) = 1.96). A checkable result must be computable and also contain a p-value (e.g.: *t*(32) = 1.96, *p* = .05). It must be emphasized that only results reported in the text can be extracted. Results in tables and figures are not part of the analysis.

To analyze the use of standard effect measures (Cohen's, η^2, beta coefficients), the corpus of studies is limited in each case to those that cite appropriate procedures (t-test, ANOVA, regression).

Because *JATSdecoder*'s extraction heuristic of the α level cannot distinguish between nominal and corrected values, the extracted maximum value is analyzed in articles that contain extractable p-values. For articles that use multiple α levels, only the maximum value is extracted for analysis.

Since the the Bayesian information criterion (BIC) is used for model selection and not for statistical inference, it is removed from the extracted statistical methods before analyzing changes in the use of Bayesian inference methods.

In an evaluation study [48] the accuracy of all applied extraction heuristics for methodological study characteristics was high (from.85 for α levels to .993 for directed testing).

Global changes in study characteristics of content published before and after 2015 are reported with 99.9% confidence intervals for differences in proportions and 99.9% bootstrap confidence intervals for differences in medians (20,000 resamples). Journal-specific analyses are not reported with p-values nor confidence intervals because the respective samples represent a complete and not a random sample of articles from the respective journal.

To facilitate the analysis of changes in the country of origin the *countrycode* [51] package is used to convert country names to continents, which are then transformed to WEIRD and non-WEIRD involved countries of origin (except for Israel and New Zealand, which are manually set to WEIRD). Multicore processing is performed using the R package *future.apply* [52].

Results

The final article collection comprises 57,909 articles. 5,900 out of 63,809 initial articles were removed, as they contain one of the exclusion search terms within the title or do not contain any extractable statistical result. The journal-based distribution of yearly released research articles is shown in Fig 1. Both full open access journals (*Frontiers in Psychology* and *PLoS ONE*) published 77% of all included articles. Along with a steady increase of yearly published articles by *Frontiers in Psychology* and *PLoS One*, *PLoS One* shows to have a slump in psychological research content in 2015. The closed access journals appear to have a more constant number of released research articles per year.

The article set is divided into two 6-year publication intervals to analyze global changes in study characteristics of psychological research. Table 1 shows the relative use of the here targeted study characteristics for the selected articles that were released before and after 2015, as well as for the full article collection.

Overall, the proportion of articles that report p-values has decreased from 92% in and before 2015 to 82% in publications after 2015. The proportion of articles that report multiple studies has decreased from 17% to 13%. The median number of reported p-values per study in articles with p-values has decreased from 14 to 12. Compared to publications in and before 2015, fewer articles contain results that are reported in a manner, that enables a recomputation (69% vs. 58%) and a consistency check of p-values (67% vs. 55%). Overall, the median proportion of reported p-values that are below.05 is constantly high (69%) and even higher (74%) when only computable p-values are considered. That is, 50% of the articles contain more than these proportions of significant results at α = .05. Hardly any use of α levels below.05 is observed.

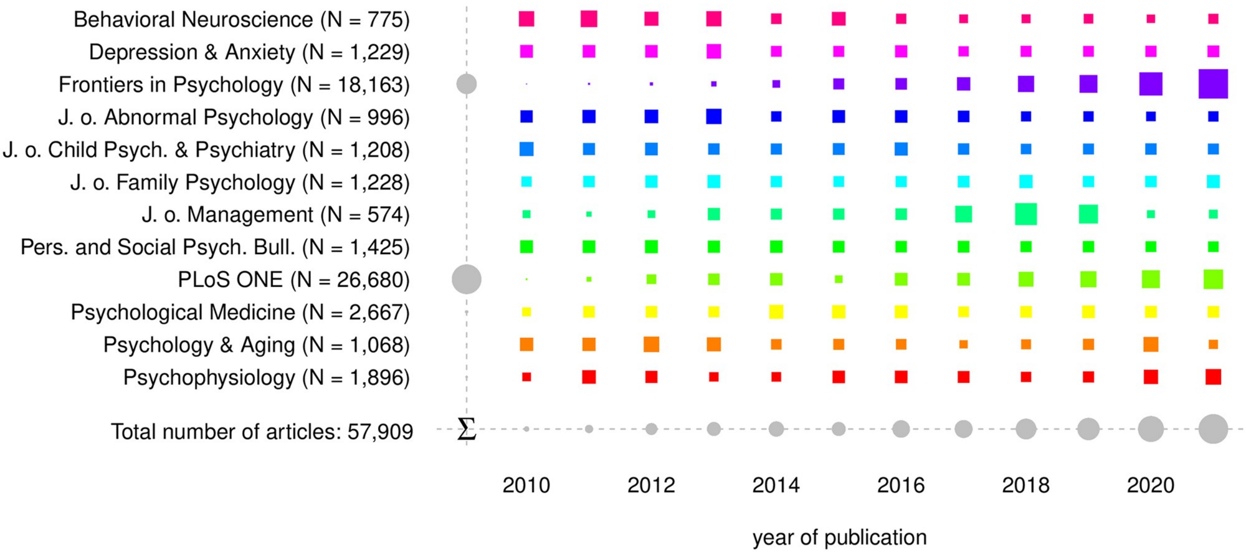


Fig 1. Number of yearly released research articles by journal and year.

https://doi.org/10.1371/journal.pone.0283353.g001

Within all included articles, the reporting of confidence intervals has increased from 21% to 32%. Reports of power analysis and the application of Bayesian inference have more than doubled, from 5% to 11% and from 2% to 5%, respectively. Corrections for multiple testing are reported in about one quarter of all articles, with a slight decrease from 27% to 23%, although multiple testing is applied in almost every article. Directed testing is rarely reported, with a

Table 1. Change in study characteristics in research articles with statistical results before and after 2015.

Feature	≤ 2015	> 2015	total	Δ	99.9% CI for Δ
N articles in initial selection	21,238	42,571	63,809	21,333	
N empirical research articles	19,443	38,466	57,909	19,023	
proportion of articles with with p-value/s	.92	.82	.85	-.098	[-.108; -.088]
→ proportion of articles with multiple studies	.17	.13	.15	-.047	[-.058; -.036]
→ median number of p-values per study	14	12	13	-2	[-2; -1]
→ proportion of articles with recomputable p-value	.69	.58	.62	-.11	[-.125; -.095]
→ proportion of articles with checkable p-value	.67	.55	.60	-.12	[-.135; -.105]
→ median proportion of reported p < .05	.69	.69	.69	-.006	[-.018; .007]
→ median proportion of computable p < .05	.73	.75	.74	-.011	[-.026; .018]
→ proportion of articles with alpha level < .05	.03	.02	.02	-.006	[-.011; -.002]
→ proportion of articles with alpha level < .01	.01	.01	.01	-.004	[-.007; -.001]
proportion of articles with confidence interval	.21	.32	.28	.106	[.092; .119]
proportion of articles with power analysis/value	.05	.11	.09	.062	[.054; .07]
proportion of articles with Bayesian analysis	.02	.05	.04	.023	[.017; .028]
proportion of articles with correction for multiple testing	.27	.23	.24	-.045	[-.058; -.031]
proportion of articles with one sided test	.05	.03	.04	-.019	[-.025; -.013]
proportion of articles with extractable sample size	.83	.77	.79	-.063	[-.075; -.05]
→ median of extracted sample sizes	105	190	151	85	[74; 95]

Note: CIs for Δ represent confidence intervals for differences in proportions and bootstrap confidence intervals for differences in medians based on 20,000 resamples.

https://doi.org/10.1371/journal.pone.0283353.t001

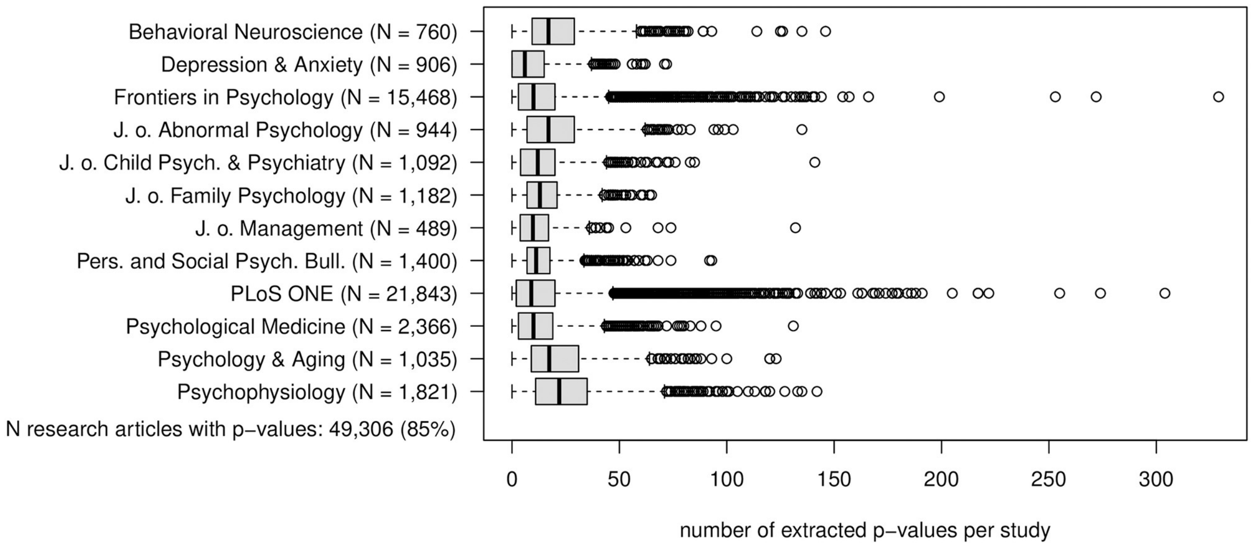


Fig 2. Number of articles with p-values and distribution of number of extracted p-values per study by journal.

https://doi.org/10.1371/journal.pone.0283353.g002

decrease from 5% to 3%. In 79% of all articles, a sample size is estimated by *JATSdecoder*. Articles published after 2015 show to have much higher median sample sizes (190) than articles published in and before 2015 (105).

More detailed insight per journal and year is given for every feature in the following sections.

The reporting of p-values

Extractable p-values are detected in 49,306 (85%) articles. The full distribution of the number of extracted p-values per study is displayed for every journal in Fig 2. It is evident that many studies report so many p-values that a correction for multiple testing is required to control the false positive rate. Every here included journal contains several articles with more than 50 p-values per study. Some studies in articles published by *Frontiers in Psychology* and *PLoS One* even contain more than 200 p-values.

A more precise insight to the amount and properties of reported p-values is displayed for every journal in Table 2. It contains the proportion of articles that report p-values within text, the median and total number of extractable, computable and checkable p-values per journal, as well as the proportion of computable p-values that fall below .05, .01 and .001.

Overall, 85% of all analyzed articles report at least one p-value. The lowest rate is observed for *Depression & Anxiety* (74%). The median number of p-values per study within articles that report p-values ranges between 10 in *Depression and Anxiety* and 22 in *Psychophysiology*. The median number of checkable p-values per study is lower for every journal. Still, 50% of all articles that contain p-values report more than 8 computable and checkable p-values. In total, 67% of all computable p-values (two-sided) are smaller than.05, 49% are below.01 and 35% below.001.

The reporting of effect sizes

Table 3 displays the relative frequencies of reported standard effect measures (Cohen’s d, η^2, $\hat{\beta}$/OR) within articles using the corresponding statistical procedure (t-test, ANOVA,

Table 2. Journal specific properties of extracted raw, computable and checkable p-values within articles that contain any statistical result.

Journal	N articles	% with p-value	median number of p-values per study			total number of p-values			proportion of computable p-values		
			p-value	comp.	check.	p-value	comp.	check.	p<.05	p<.01	p<.001
Behavioral Neuroscience	775	98%	17	12.8	12	22,427	15,199	14,336	63%	43%	29%
Depression & Anxiety	1,229	74%	10	6	6	12,005	4,980	3,827	65%	45%	32%
Frontiers in Psychology	18,163	85%	12.5	8	8	303,221	163,286	150,908	67%	50%	37%
J. Abnormal Psychology	996	95%	18	10	10	21,108	10,081	9,447	66%	47%	33%
J. Child Psych. & Psychiatry	1,208	90%	13	7	7	17,387	7,485	6,906	67%	45%	31%
J. Family Psychology	1,228	96%	13.5	6	5	19,326	7,371	6,643	64%	45%	30%
J. Management	574	85%	11	4	3.5	8,032	2,349	1,779	73%	61%	51%
Pers. and Social Psych. Bull.	1,425	98%	11.5	7.7	7.2	48,804	32,877	31,057	69%	48%	35%
PLoS ONE	26,680	82%	12	7.5	7	409,800	146,175	133,748	67%	49%	35%
Psychological Medicine	2,667	89%	11	6	6	35,153	11,185	10,125	65%	45%	32%
Psychology & Aging	1,068	97%	18	13	11.9	27,818	18,067	16,432	72%	54%	40%
Psychophysiology	1,896	96%	22	16	15	51,121	34,240	32,226	70%	50%	35%
Total	57,909	85%	13	8	8	976,202	453,295	417,434	68%	49%	35%

https://doi.org/10.1371/journal.pone.0283353.t002

regression) and their increase factors for articles published before and after 2016. The proportions of articles reporting a standard effect measure are calculated based on the number of articles in which *JATSdecoder* identified an appropiate statistical method (t-test, ANOVA, Regression).

The reporting of effect sizes has increased for all measures involved and differs greatly between journals. All journals show inceased reporting rates for Cohen's d in articles that report the use of t-tests. Wheras 4% of the articles by *Behavioural Neuroscience* that report t-test results and that where published before 2016 report Cohen's d, 50% of articles published by *Personality and Social Psychology* after 2015 report it within the text corpus. Global textual reporting rates for η^2 in articles with ANOVA and $\hat{\beta}$/OR in articles with regression analysis have also mostly increased. Each of the effect measures presented here is most commonly

Table 3. Proportion of articles that report standard effect sizes in studies with t-test, ANOVA and regression analysis.

Journal	Cohen's d in t-test			η^2 in ANOVA			$\hat{\beta}$/OR in regression		
	$\leq$ 2015	$\geq$ 2015	factor	$\leq$ 2015	$\geq$ 2015	factor	$\leq$ 2015	$\geq$ 2015	factor
Behavioral Neuroscience	0.04	0.13	3.09	0.04	0.18	4.72	0.09	0.11	1.20
Depression & Anxiety	0.21	0.24	1.18	0.16	0.12	0.74	0.25	0.23	0.92
Frontiers in Psychology	0.21	0.30	1.43	0.49	0.55	1.14	0.37	0.47	1.28
J. Abnormal Psychology	0.30	0.50	1.65	0.23	0.40	1.69	0.21	0.28	1.29
J. Child Psych. & Psychiatry	0.22	0.30	1.36	0.29	0.28	0.94	0.37	0.40	1.08
J. Family Psychology	0.16	0.19	1.18	0.08	0.16	1.92	0.30	0.43	1.45
J. Management	0.08	0.15	1.95	0.08	0.16	2.12	0.48	0.65	1.36
Pers. and Social Psych. Bull.	0.40	0.57	1.45	0.57	0.73	1.28	0.84	0.84	0.99
PLoS ONE	0.12	0.16	1.42	0.26	0.31	1.19	0.25	0.22	0.91
Psychological Medicine	0.15	0.21	1.41	0.12	0.17	1.41	0.25	0.29	1.16
Psychology & Aging	0.26	0.39	1.50	0.39	0.58	1.49	0.24	0.37	1.50
Psychophysiology	0.16	0.34	2.11	0.37	0.41	1.09	0.34	0.44	1.29
Total	0.15	0.24	1.54	0.32	0.42	1.33	0.40	0.44	1.09

https://doi.org/10.1371/journal.pone.0283353.t003

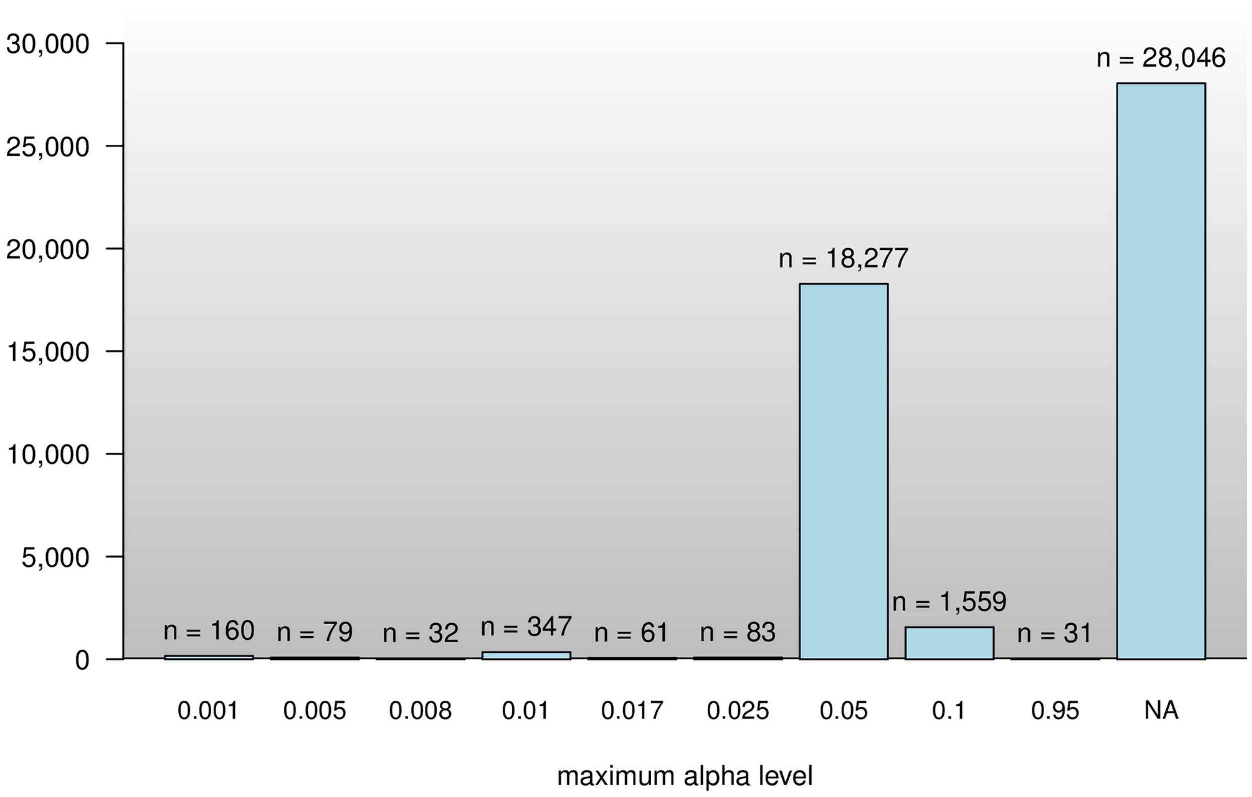


Fig 3. Distribution of maximum α levels extracted over 30 times.

https://doi.org/10.1371/journal.pone.0283353.g003

reported in *Personality and Social Psychology* with 84% of studies using an regression also reporting $\hat{\beta}$ or OR within the text corpus.

α level

Most of the articles (57%) that contain p-values do not contain an extractable report of an α level but mostly make implicit use of the standard α level level of .05. In order to rule out incorrect and mostly corrected α levels, Fig 3 shows the distribution of the extracted maximum α levels that were extracted over 30 times. α levels other than.05 are extremely rare. The second most common detected maximum α level is 10%. Standard α levels of .01 and .001 are hardly detected in any article (N = 347 and N = 160). Most of the extracted values of .005 are corrected α levels in studies that analyze voxels in fMRI analyzes with a nominal α of .05. Some articles (N = 31) report using a 95% significance level, but actually use a.05 α level.

Table 4 shows the relative frequency of articles with p-values that report the use of a maximum α level below.05. There do not seem to be any notable changes in the use of the α level over time. The highest proportion of articles with p-values below.05 is detected in the articles of the *Journal of Management* in 2014 (7%). However, it remains unclear whether this is due to corrected or nominal levels of α.

Table 4. Relative frequency of extracted maximum α levels < .05 in articles with p-values.

Journal	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Sum
Behavioral Neuroscience	.04	.02	.00	.00	.00	.04	.00	.04	.00	.02	.00	.02	.02
Depression & Anxiety	.00	.00	.04	.04	.06	.02	.01	.00	.00	.02	.05	.03	.03
Frontiers in Psychology	.01	.03	.02	.03	.03	.02	.03	.02	.03	.02	.02	.02	.02
J. Abnormal Psychology	.04	.02	.03	.02	.06	.02	.00	.01	.03	.01	.00	.04	.02
J. Child Psych. & Psychiatry	.03	.01	.01	.02	.00	.04	.01	.00	.02	.02	.00	.03	.02
J. Family Psychology	.00	.00	.01	.03	.00	.01	.00	.00	.02	.00	.00	.00	.01
J. Management	.00	.00	.04	.00	.07	.03	.00	.00	.00	.00	.00	.03	.01
Pers. and Social Psych. Bull.	.01	.02	.01	.01	.02	.01	.00	.01	.00	.01	.00	.01	.01
PLoS ONE	.02	.03	.04	.03	.03	.03	.03	.02	.02	.02	.01	.01	.02
Psychological Medicine	.03	.02	.01	.03	.04	.02	.01	.01	.03	.02	.00	.02	.02
Psychology & Aging	.00	.02	.02	.01	.01	.01	.01	.07	.03	.01	.00	.04	.02
Psychophysiology	.03	.00	.01	.01	.01	.01	.01	.03	.02	.03	.02	.01	.01
Total	.02	.02	.03	.02	.03	.02	.02	.02	.02	.02	.02	.02	.02

https://doi.org/10.1371/journal.pone.0283353.t004

Use of confidence intervals

Within articles that contain any extractable result, the proportion of articles that report confidence intervals has increased (from 21% before 2016 to 32% after 2016) with a peak of 35% in 2021. All journals report more confidence intervals today than they did before 2016. Table 5 shows the relative frequency of detected reports of confidence intervals in articles with p-values. The highest rate of use is observed in *Personality and Social Psychology Bulletin* (90%) which seems to have changed reporting standards in 2015, when reporting rates increased strongly. Between 2010 and 2021, the relative use of confidence intervals increased the most in the *Personality and Social Psychology Bulletin* (from 19% to 90% and the *Journal of Abnormal Psychology* (from 22% to 65%). The lowest rate of articles with confidence intervals is found in *Behavioral Neuroscience* (6% in total, 13% in 2021).

Table 5. Relative frequency of confidence interval use in articles with p-values.

Journal	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Behavioral Neuroscience	.01	.01	.01	.05	.07	.11	.05	.17	.06	.08	.11	.13	.06
Depression & Anxiety	.33	.30	.33	.40	.40	.45	.48	.43	.57	.49	.54	.57	.44
Frontiers in Psychology	.15	.08	.10	.10	.12	.16	.21	.25	.25	.29	.29	.32	.25
J. Abnormal Psychology	.22	.31	.24	.27	.28	.42	.32	.50	.52	.46	.41	.65	.37
J. Child Psych. & Psychiatry	.27	.25	.36	.40	.44	.42	.42	.48	.50	.45	.53	.55	.42
J. Family Psychology	.22	.22	.22	.28	.24	.36	.32	.35	.43	.44	.35	.41	.32
J. Management	.25	.14	.13	.14	.18	.41	.34	.35	.39	.35	.44	.56	.32
Pers. and Social Psych. Bull.	.19	.24	.25	.42	.55	.87	.90	.86	.80	.88	.89	.90	.63
PLoS ONE	.11	.12	.16	.20	.21	.24	.24	.26	.29	.31	.34	.34	.26
Psychological Medicine	.44	.48	.47	.47	.44	.49	.46	.52	.44	.47	.51	.47	.47
Psychology & Aging	.12	.20	.15	.16	.24	.20	.28	.24	.36	.45	.29	.44	.25
Psychophysiology	.09	.06	.09	.09	.09	.13	.20	.16	.22	.29	.33	.38	.19
Total	.20	.18	.18	.21	.22	.25	.27	.29	.30	.33	.34	.35	.28

https://doi.org/10.1371/journal.pone.0283353.t005

Table 6. Relative frequency of power values or mentions of power analysis by journal and year in articles with p-values.

Journal	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Behavioral Neuroscience	.00	.00	.00	.01	.00	.03	.00	.02	.06	.06	.04	.13	.02
Depression & Anxiety	.04	.03	.06	.05	.09	.08	.07	.07	.08	.07	.08	.14	.07
Frontiers in Psychology	.04	.02	.02	.03	.04	.04	.05	.07	.09	.12	.12	.12	.09
J. Abnormal Psychology	.02	.03	.04	.05	.08	.04	.07	.05	.08	.17	.15	.12	.07
J. Child Psych. & Psychiatry	.07	.04	.04	.09	.06	.07	.10	.13	.14	.14	.12	.11	.09
J. Family Psychology	.04	.03	.03	.05	.06	.04	.05	.06	.05	.10	.08	.13	.06
J. Management	.00	.05	.10	.04	.00	.00	.02	.01	.01	.00	.09	.03	.02
Pers. and Social Psych. Bull.	.02	.01	.00	.02	.05	.22	.37	.34	.53	.63	.63	.62	.27
PLoS ONE	.02	.04	.05	.06	.06	.08	.07	.08	.10	.11	.12	.12	.09
Psychological Medicine	.07	.06	.06	.07	.06	.06	.05	.10	.03	.08	.07	.05	.06
Psychology & Aging	.05	.01	.03	.05	.06	.05	.06	.08	.23	.23	.26	.33	.11
Psychophysiology	.03	.02	.04	.04	.05	.04	.05	.12	.12	.19	.26	.30	.11
Total	.03	.03	.04	.05	.05	.06	.07	.09	.10	.12	.13	.13	.09

https://doi.org/10.1371/journal.pone.0283353.t006

Power analysis

The use of power analysis in articles that contain extractable p-values has more than quadrupled, from 3% in 2010 to 13% in 2021, but even so heavily varies between journals. In every journal, an increase of power analytical concepts is observed. Table 6 shows the relative amount of articles that report any power value by journal and year. Whereas comparatively few articles published by the *Personality and Social Psychology Bulletin* report the use of power analysis before 2015, the proportion of articles with power analysis peaked at 63% in 2019 and 2020. Also, articles in *Psychology and Aging* show a strong increase in reports of power analyses, with 33% of all articles in 2021. Lowest increase rates are observed for the *Journal of Management* and *Psychological Medicine*. The lowest overall rates of power analysis are observed for *Behavioral Neuroscience* and the *Journal of Management* (both with 2% overall rates).

Since *JATSdecoder* does not discriminate between a priori and post hoc power values, Table 7 shows the absolute and relative frequencies of the first detected power value of articles that report power as categorized values. More than half of the 3,094 articles that report power values (56%) seem to have made use of the a priori set standard value of .8. The values that are either .9 or .95 (10% each) could have been listed as a priori or a posteriori values. Since studies with power below .8 would not be very meaningful, it is probable that all values that fall within an interval were calculated post hoc.

Bayesian statistics

The overall rate of articles mentioning Bayesian inferential methods (extractions of the Bayesian information criterion have been removed) has quadrupled from 1% in 2010 to 4% in 2021. Table 8 shows the relative amount of articles that mention Bayesian inferential methods by journal and year. The highest rate is observed in the *Journal of Management* in 2015 (26%),

Table 7. Absolute (h(x)) and relative (f(x)) frequency distribution of the first detected and categorized power value per article.

Power	[0; .8)	.8	(.8; .85)	.85	(.85; .9)	.9	(.9; .95)	.95	(.95; 1]	Total
h(x)	443	2,335	81	69	65	411	81	429	274	4,188
f(x)	.11	.56	.02	.02	.02	.10	.02	.10	.07	1

https://doi.org/10.1371/journal.pone.0283353.t007

Table 8. Relative frequency of application of Bayesian inferential statistics by journal and year.

Journal	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Behavioral Neuroscience	.00	.00	.00	.01	.02	.00	.00	.08	.04	.04	.04	.07	.02
Depression & Anxiety	.01	.01	.01	.02	.01	.02	.01	.02	.03	.00	.04	.03	.02
Frontiers in Psychology	.02	.01	.05	.04	.04	.05	.05	.04	.05	.05	.05	.04	.04
J. Abnormal Psychology	.02	.02	.00	.01	.01	.03	.07	.04	.11	.10	.15	.12	.05
J. Child Psych. & Psychiatry	.00	.00	.00	.02	.03	.03	.01	.02	.00	.03	.04	.03	.02
J. Family Psychology	.00	.02	.00	.02	.01	.02	.03	.01	.01	.02	.04	.01	.02
J. Management	.00	.00	.00	.00	.00	.26	.02	.03	.02	.06	.00	.03	.04
Pers. and Social Psych. Bull.	.00	.00	.00	.00	.01	.03	.03	.01	.04	.04	.07	.09	.02
PLoS ONE	.03	.02	.02	.03	.02	.02	.04	.03	.05	.05	.05	.04	.04
Psychological Medicine	.01	.01	.01	.01	.00	.00	.01	.02	.03	.03	.04	.06	.02
Psychology & Aging	.03	.01	.01	.02	.01	.04	.07	.10	.12	.17	.13	.17	.07
Psychophysiology	.00	.01	.01	.02	.00	.03	.03	.08	.04	.11	.14	.13	.05
Total	.01	.01	.02	.02	.02	.03	.04	.04	.04	.05	.05	.04	.04

https://doi.org/10.1371/journal.pone.0283353.t008

which published a special issue with the title *Bayesian Probability and Statistics in Management Research* in February. Still, great differences can be observed between the journals. While in 2021 17% of all empirical research articles published by *Psychology and Aging* use inferential Bayesian methods, the rate in the *Journal of Family Psychology* is 1%.

Preregistration and multiverse analyses

The 'preregistration revolution' called for by Nosek et al. [37] has not yet taken place. A clear distinction of preregistered and registered reports and protocols cannot be made. In six of the 12 journals, a text search in the title and abstract with the regular expression pattern 'registered report|registered replication|registered stud[yi]' identified only 82 articles (0.14%), that were preregistered or registered. Table 9 shows the absolute frequency of detected preregistered or registered reports by journal and year. The first preregistered reports were published by *Frontiers in Psychology* in 2013. Overall, *PLoS One* has published most preregistered psychological reports or protocols, with 40 articles. A text search with the pattern 'replicat' in the title and abstract reveals that 25 of the 82 articles (30%) are replications. In total, the same search task

Table 9. Absolute frequency of preregistered or registered reports by journal and year.

Journal	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Behavioral Neuroscience	0	0	0	0	0	0	0	0	0	0	0	0	0
Depression & Anxiety	0	0	0	0	0	0	0	0	0	0	0	0	0
Frontiers in Psychology	0	0	0	2	1	2	1	0	1	3	11	7	28
J. Abnormal Psychology	0	0	0	0	0	0	0	0	1	0	0	1	2
J. Child Psych. & Psychiatry	0	0	0	0	0	0	0	0	0	0	0	0	0
J. Family Psychology	0	0	0	0	0	0	0	0	0	0	0	0	0
J. Management	0	0	0	0	0	0	0	0	0	0	0	0	0
Pers. and Social Psych. Bull.	0	0	0	0	0	0	0	0	1	2	0	2	5
PLoS ONE	0	0	0	0	0	0	0	1	3	6	8	22	40
Psychological Medicine	0	0	0	0	0	0	0	0	0	0	0	0	0
Psychology & Aging	0	0	0	0	0	0	0	0	0	1	0	1	2
Psychophysiology	0	0	0	0	0	0	0	0	0	0	2	3	5
Total	0	0	0	2	1	2	1	1	6	12	21	36	82

https://doi.org/10.1371/journal.pone.0283353.t009

identified 1,992 (3.4%) replications in the entire collection of articles. Thus, among the replications, preregistration is more common, although still quite rare (1.3%).

Multiverse analyses are performed even less frequently than preregistrations in the literature analyzed here. A search task on the title, abstract, and extracted statistical methods with the regular expression '[Mm]ulti[-]*verse' yielded only ten articles mentioning multiverse analyses. The first four mentions are observed in 2018, one of which is is a non-excluded theoretical paper by Gelman [5] 'The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It'. In each of the following three years, two articles were identified that used multiverse analyses.

Correction for multiple testing

Although most of the articles that contain p-values use multiple testing, only 24% report the use of a corrected α level. There is great variability between journals regarding to the use of procedures that correct the α level for multiple testing. Whereas half of all included articles with p-values by *Behavioral Neuroscience* report the use of corrections for multiple testing, with a peak of 62% in 2019, only 2% of all articles by the *Journal of Management* report a correction of α levels. Except for *PLoS ONE* and *Frontiers in Psychology*, all other journals contain an increased amount of articles with multiple test corrections since 2016, although most changes are rather weak (see Table 10).

Fig 4 shows the relation of the relative amount of articles that report the use of correction procedures for multiple testing and the number of extracted p-values from text for every journal. In general, the proportion of articles that make use of α level correction procedures increases with the number of reported p-values in almost every journal. Except for articles with more than 80 p-values, articles published in *Behavioral Neuroscience* make the most intensive use of α level correction procedures, probably because they are an implemented standard in fMRI data analysis software. The lowest correction rates are observed for the *Journal of Management*, which, at the same time, has the lowest maximum number of 74 extractable p-values within text. Seven out of eight articles with more than 200 p-values, which are published by *PLoS One* and *Frontiers in Psychology*, report the use of α level corrections.

Test direction

Although the extraction heuristics showed to have a high accuracy, an explicit report of one-tailed testing is found in only 4% of all articles. At least one one-tailed test is detected in 5% of all articles published in and before 2015 and in 3% of all articles published after 2015 (see Table 1). Table 11 shows the relative frequency of detected test direction by journal. The highest amount of articles using one-tailed tests is found in *Psychophysiology* (8%).

Estimated sample size

The estimated sample sizes vary greatly between journals. Table 12 displays the journal specific median and .75-quanile of the estimated sample sizes in and before 2015 and after 2015 as well as the corresponding increase factors. Except for *Behavioral Neuroscience*, the median sample sizes have increased in every journal. The .75-quantiles of sample sizes have also increased in every journal. In articles published after 2015 in *Depression & Anxiety* and *Psychological Medicine*, more than 25% of all estimated sample sizes are greater than 2,000.

A more detailed overview of the median estimated sample sizes by journal and year, as well as global medians, is presented in Table 13. More than half of all annual estimated sample sizes in *Behavioral Neuroscience* are below 75 and in *Psychophysiology* below 85. All other journals

Table 10. Relative frequency of at least one detected multiple test correction procedure by journal and year in articles with p-values.

Journal	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Behavioral Neuroscience	.48	.48	.43	.52	.48	.55	.50	.46	.61	.62	.59	.36	.50
Depression & Anxiety	.09	.21	.16	.16	.24	.19	.16	.21	.28	.27	.20	.26	.20
Frontiers in Psychology	.21	.27	.26	.28	.27	.26	.25	.26	.25	.23	.21	.18	.23
J. Abnormal Psychology	.19	.14	.16	.15	.14	.18	.15	.16	.21	.28	.28	.31	.19
J. Child Psych. & Psychiatry	.16	.18	.17	.24	.17	.13	.25	.24	.16	.19	.23	.25	.20
J. Family Psychology	.16	.07	.03	.08	.06	.10	.12	.08	.09	.12	.07	.10	.09
J. Management	.00	.00	.00	.06	.02	.02	.04	.03	.00	.01	.00	.03	.02
Pers. and Social Psych. Bull.	.07	.09	.05	.04	.06	.05	.08	.08	.09	.06	.10	.11	.07
PLoS ONE	.33	.34	.34	.35	.33	.31	.29	.27	.23	.23	.20	.19	.26
Psychological Medicine	.19	.23	.19	.23	.22	.29	.33	.22	.30	.28	.35	.31	.26
Psychology & Aging	.13	.11	.15	.16	.16	.14	.12	.16	.20	.18	.22	.20	.16
Psychophysiology	.28	.28	.31	.36	.40	.31	.34	.38	.38	.46	.45	.46	.37
Total	.22	.26	.27	.30	.29	.26	.26	.25	.24	.23	.21	.19	.24

https://doi.org/10.1371/journal.pone.0283353.t010

show to have much higher median sample sizes with a maximum of 732 in the Journal of Child Psychology & Psychiatry in 2021.

Table 14 shows the median of the estimated sample sizes for articles with and without a repeated measures analysis (RMA) and the proportions of articles with RMA. The partitioning

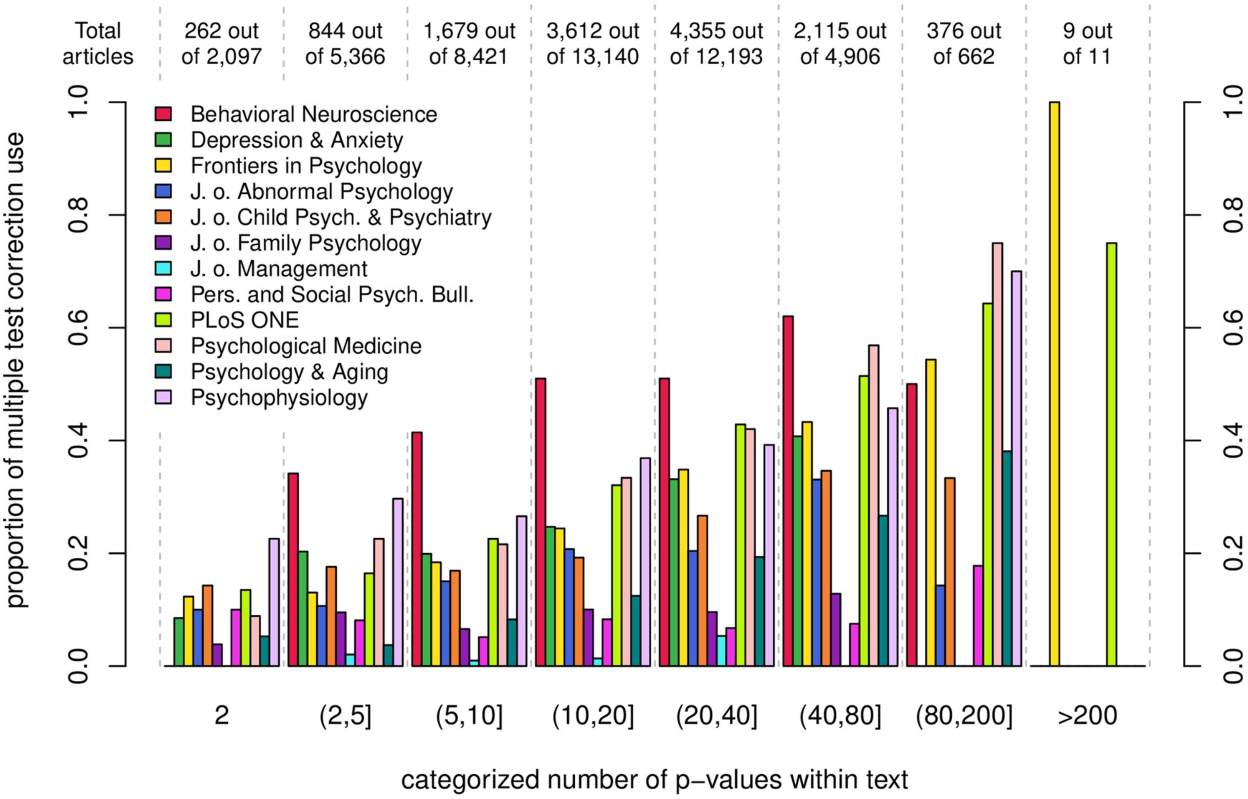


Fig 4. Absolute frequencies of articles with correction procedures by categorized number of p-values (numbers on top) and journal-wise relation of number of extracted p-values from text and use of correction procedures for multiple testing (bars).

https://doi.org/10.1371/journal.pone.0283353.g004

Table 11. Relative frequencies of detected test direction by journal.

Journal	no detection	one and two sided	one sided	two sided
Behavioral Neuroscience	.85	.02	.03	.10
Depression & Anxiety	.79	.01	.01	.19
Frontiers in Psychology	.89	.01	.03	.07
J. Abnormal Psychology	.88	.01	.02	.08
J. Child Psych. & Psychiatry	.86	.01	.02	.11
J. Family Psychology	.94	.00	.02	.04
J. Management	.80	.03	.04	.13
Pers. and Social Psych. Bull.	.89	.01	.03	.06
PLoS ONE	.84	.01	.02	.13
Psychological Medicine	.81	.01	.01	.17
Psychology & Aging	.91	.01	.03	.05
Psychophysiology	.80	.03	.05	.12
Total	.86	.01	.03	.11

https://doi.org/10.1371/journal.pone.0283353.t011

was performed with a search in the extracted statistical methods using the regular expression *'repeated measure|cross lagged|panel |longitud'*. While the proportion of articles with RMA decreased from 25% to 16%, the median of the estimated sample size increased for both types of designs (68 to 89 in articles with RMA, 129 to 232 in articles without RMA). Both, the absolute and relative increase in median estimated sample size is higher for articles without measurement repetition.

Country of origin

Fig 5 shows the change in relative frequency of country involvement for articles published before and after 2015 for the most frequently detected countries of origin after 2015. Whereas the relative amount of articles from the United States has decreased most (26.6% to 19.6%), the relative amount of articles from China has nearly doubled (4.2% to 7.9%.) In relative terms, Norway, Poland and Portugal, three comparably small distributors in and before 2015, have increased their output the most.

Table 12. Median and .75-quantile of estimated sample size before and in 2015 and after 2015 by journal.

Journal	median			.75-quantile		
	$\leq$ 2015	$>$ 2015	factor	$\leq$ 2015	$>$ 2015	factor
Behavioral Neuroscience	58	56	0.97	116.75	134	1.15
Depression & Anxiety	217	390.5	1.8	1164	2068.25	1.78
Frontiers in Psychology	77	189.5	2.46	179	525	2.93
J. Abnormal Psychology	207.5	262	1.26	726.75	900	1.24
J. Child Psych. & Psychiatry	265.5	461.5	1.74	1186.5	1920.75	1.62
J. Family Psychology	246	300	1.22	593	628	1.06
J. Management	243	304	1.25	562.5	800.75	1.42
Pers. and Social Psych. Bull.	161	356	2.21	305	900	2.95
PLoS ONE	82	170	2.07	317.25	724.5	2.28
Psychological Medicine	311	326	1.05	2075	2705.25	1.3
Psychology & Aging	156	201	1.29	448	994	2.22
Psychophysiology	61	74	1.21	129	143	1.11

https://doi.org/10.1371/journal.pone.0283353.t012

Table 13. Median estimated sample size by journal and year.

Journal	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Behavioral Neuroscience	56	72	47	60	55	64	74	45	59	50	56	64	57
Depression & Anxiety	260	151	274	273	205	240	368	182	344	334	474	519	285
Frontiers in Psychology	46	53	52	73	78	95	114	150	136	175	228	272	149
J. Abnormal Psychology	208	214	202	237	127	228	199	321	162	198	571	462	229
J. Child Psych. & Psychiatry	256	190	285	244	263	404	399	286	208	648	518	732	365
J. Family Psychology	297	230	210	228	279	288	285	269	305	247	331	306	268
J. Management	184	389	194	282	221	374	379	300	241	329	373	301	287
Pers. and Social Psych. Bull.	132	141	132	183	170	245	234	317	356	328	406	557	233
PLoS ONE	49	59	76	77	98	106	114	127	166	173	201	282	127
Psychological Medicine	192	248	490	318	421	258	224	273	530	410	419	331	318
Psychology & Aging	152	181	111	150	254	186	189	175	172	481	164	295	175
Psychophysiology	51	52	61	52	79	73	72	59	71	81	84	80	68
Total	114	97	97	100	108	119	129	148	165	190	220	271	151

https://doi.org/10.1371/journal.pone.0283353.t013

Fig 6 shows the relative frequency of article origin from WEIRD and non-WEIRD countries of origin for the publication period focused on here. The amount of articles from non-WEIRD countries has steadily increased within the last 12 years. Whereas in 2010 a relatively small amount of articles (12%) was published or co-published by authors from non-WEIRD countries, non-WEIRD country involvement increased to 43.7% in 2020. In 2021, 21.4% of the articles with extractable country were published solely by authors from non-WEIRD countries, whereas 56.3% were published solely by authors from WEIRD countries.

Discussion

This study contributes to the ongoing discussion about the replication crisis in psychology and is the first application of *JATSdecoder* to gain detailed insights into the research practice of a subject. Surprisingly, most of the long-proposed methodological reforms listed above are implemented only hesitantly or not at all.

In psychological research, the use of nil-null hypothesis testing is pervasive. Interval hypotheses are almost never tested. The global median of the reported number of p-values per study is 13. The analysis supports Meehl's and Cohen's statement about the predictably high rejection rate of point-null hypotheses. Regardless of the type of test and sample size, 68% of the recalculated p-values are significant at $\alpha = .05$. In half of the articles, more than two-thirds of the extracted and three-quarters of the recalculated p-values are below.05.

Given the high rejection rates of null hypotheses and the evidence from the literature that point null hypotheses are all too easily rejected, a critical examination of the common practice of undirected nil-null hypothesis testing seems necessary. In practice, thresholds other than

Table 14. Median estimated sample sizes in articles with and without repeated measures analysis and proportion of articles with repeated measures design.

Design type	≤2015	>2015	δ	99.9% CI for δ
with repeated measures	68	89	21	[13; 27]
without repeated measures	129	232	103	[88; 116]
proportion with repeated measures	0.25	0.16	-0.085	[-0.103; -0.068]

Note: CIs for δ represent bootstrap confidence intervals for differences in medians based on 20,000 resamples and a confidence interval for difference in proportions.

https://doi.org/10.1371/journal.pone.0283353.t014

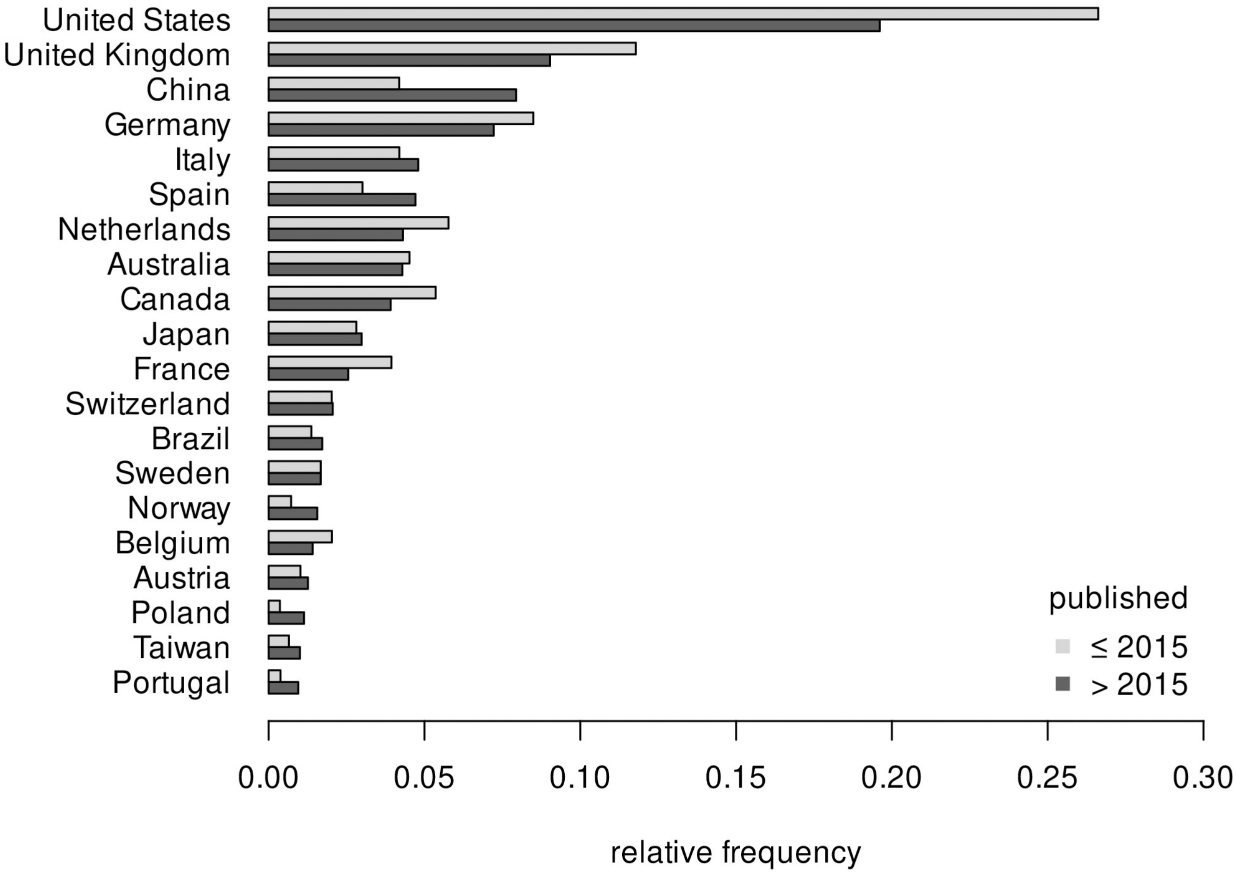


Fig 5. Change in relative country involvement before and after 2015.

https://doi.org/10.1371/journal.pone.0283353.g005

zero would need to be discussed in the community and adjusted over time, but could serve as an indicator of relevance or quality in a particular research area or domain.

Departing from the ritualized nil-null hypothesis test would certainly be associated with a higher rate of 'unsuccessful' studies, but at the same time the focus would shift to the relevance of the observed effects and the resilience of the underlying theories. A threshold-based approach can also produce non-significant results as strong evidence for a theory when non-zero effects are tested with sample sizes that yield high power. In addition, a significant result would then have to be considered as a refutation or limitation of a theory, which is not the case with non-significant results in low-powered studies.

The methodological study features focused on here vary greatly between journals. Some of these features have changed within the last 12 years. Certainly the underlying effects for the observed changes are multicausal and dynamic. Besides general concerns about reproducibility and a general greater awareness of methodological issues, the causes may be strongly driven by journal requirements, the adoption of changing standards by researchers, and editors and reviewers enforcing these standards.

The reporting of correction procedures for multiple testing has decreased in articles published after 2015. Although it is not indicated to strictly correct for the cumulation of α errors of every test carried out in a study [53], the large number of hypothesis tests per study indicates that at least some sort of α error adjustment is required in many cases. The simple call to set

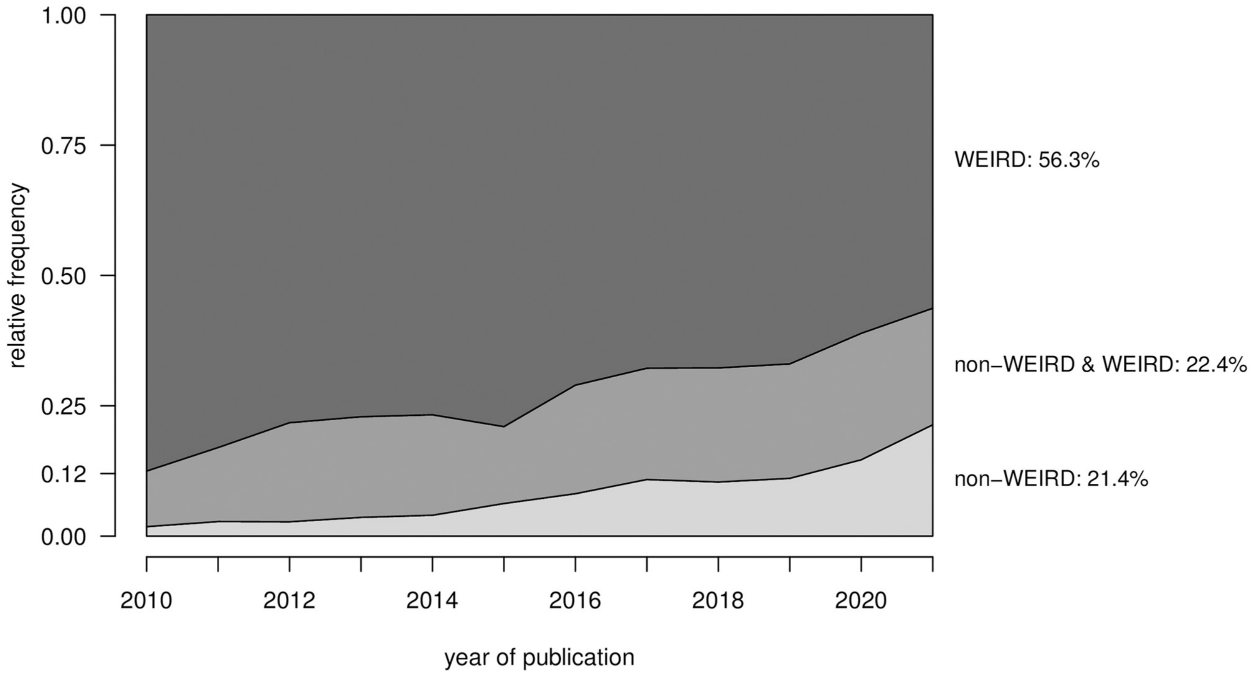


Fig 6. Change in WEIRD and non-WEIRD country of origin over time.

https://doi.org/10.1371/journal.pone.0283353.g006

the α level lower than.05 is not followed. Instead, α levels greater than.05 were found more frequently than values below.05. Although corrections for multiple testing and lower α errors are associated with lower power, their use should be considered more frequently in psychology.

Standard effect sizes are increasingly often reported in nearly every journal, which is beneficial for meta-analyses that accumulate effect sizes. Also, the reporting of confidence intervals has steadily increased, from 20% in all articles in 2010 to 35% in 2021. In the journal *Personality and Social Psychology Bulletin*, reporting effect sizes and confidence intervals has become the general standard. This is not surprising since the publication guidelines require their reporting regardless of the significance level. However, it remains unclear whether confidence intervals are mainly interpreted in terms of their width or as a simple substitute for a significance test, which would not be any information gain.

The estimated sample sizes have increased within the examined period from a median of 105 in articles published in and before 2015 to 190 in articles published after 2015. The trend toward larger samples is observed in both repeated-measures and single-measures designs. Assuming that the effect sizes of interest have not changed, this is an indication of increased test power and a good sign. However, it is not yet clear whether this effect is related to the more frequent use of power analyses or to the increasing ease of obtaining samples through online surveys or other sources.

Although several journals request a power analysis in their current guidelines, realized sample sizes are rarely justified with power analysis. The proportion of reported power analyzes has increased from only 3% in 2010 and 2011 to 13% in 2020 and 2021. This could be due to the high cost that a power analysis would indicate for collecting a sample needed to demonstrate a realistically estimated effect size. An intriguing question would be with what effect sizes the reported power analyses were conducted, and whether there is a stringent relationship with the realized sample sizes.

Besides the fact that any good theory implies the direction of an effect, directed testing represents a simple and cost-saving way to increase the power of a statistical test. Because one-tailed testing is generally a rare practice, doubts about post-hoc decisions on test sidedness (which surely is a questionable research practice) can be dispelled by preregistration.

None of the the journals analyzed here requires preregistration, and preregistration is rarely used in practice. This would be particularly desirable for confirmatory studies, as preregistration increases the credibility of results by minimizing questionable research practices. However, to ensure that non-significant results are included in the literature, preregistered reports should be peer reviewed in advance to data collection and published regardless of the outcome.

The origin of psychological research has turned to a less WEIRD one, mostly due to the many recent publications from China. It is reasonable to conclude, that also the origin of samples has diversified. However, it remains to be investigated whether the samples have become more multicultural on the article level and whether other sample characteristics have changed, for example whether students are less over-represented in the samples. Digital questionaires and worker markets have made it very easy to invite international participants to surveys. How much psychological research is based international samples nowadays remains an open question for future research. From a population-based perspective, research from African and Latin American countries is still severely underrepresented, which may be due to less research funding in those countries.

Here, mainly global comparisons were performed. More detailed analyses of specific subgroups of studies and correlations of study characteristics are still pending. Study characteristics in all open access journals that are part of the PubMed Central database can be easily downloaded or analyzed via an interactive web application, accessible at: www.scianalyzer.com. It allows identification and analysis of individual article selections by journal, topic, author, affiliation, time ranges and study features extracted with *JATSdecoder*. Further, the data export provided enables individual analyses of the study properties for selected articles.

Limitations

Although a rather big sample of articles was included for this analysis, the article collection represents a selective sample by highly renowned journals with high standards. The results should therefore not be generalized to other journals or to psychology in general. As no weighting has been performed, the observed global changes are heavily influenced by the high number of articles by both here included open access journals, that supply 77% of the analyzed articles. The selection process for original research articles may have resulted in false positive and negative inclusions. However, this should have little to no impact on the reported results.

The *JATSdecoder* algorithms are precise, but not error-free. Extraordinary or only implicitly inferred study features cannot be extracted, resulting in a negative bias. Additionally, journal- and time period-specific extraction biases may have occurred, limiting the comparability of extracted study chararacteristics. Since statistical results within tables were not extracted, the total number of test results is certainly higher. Since the conversion of special characters in PDF documents can be error-prone and results reported in tables and figures are not extracted with *JATSdecoder* the frequencies of reported test results and standard effect sizes may have been underestimated.

Acknowledgments

The author thanks Marcella Dudley for the linguistic revision of the manuscript.

Author Contributions

Conceptualization: Ingmar Böschen.

Data curation: Ingmar Böschen.

Formal analysis: Ingmar Böschen.

Methodology: Ingmar Böschen.

Software: Ingmar Böschen.

Writing – original draft: Ingmar Böschen.

References

1. Meehl PE. Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science. 1967; 34(2):103–115. https://doi.org/10.1086/288135
2. Berger JO, Sellke T. Testing a point null hypothesis: The irreconcilability of p values and evidence. Journal of the American Statistical Association. 1987; 82(397):112–122. https://doi.org/10.2307/2289139
3. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods. 2000; 5(2):241–301. https://doi.org/10.1037/1082-989X.5.2.241 PMID: 10937333
4. Gigerenzer G, Krauss S, Vitouch O. The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. In: Handbook on Quantitative Methods in the Social Sciences. Sage, Thousand Oaks, CA. Citeseer; 2004. p. 391–408.
5. Gelman A. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. Personality and Social Psychology Bulletin. 2018; 44(1):16–23. https://doi.org/10.1177/0146167217729162 PMID: 28914154
6. Cohen J. The statistical power of abnormal-social psychological research: a review. The Journal of Abnormal and Social Psychology. 1962; 65(3):145–153. https://doi.org/10.1037/h0045186 PMID: 13880271
7. Sedlmeier P, Gigerenzer G. Do studies of statistical power have an effect on the power of studies? Psychological Bulletin. 1992; 105(2):309–316. https://doi.org/10.1037/0033-2909.105.2.309
8. Arnett JJ. The neglected 95American. American Psychologist. 2008; 63(7):602–614. https://doi.org/10.1037/0003-066X.63.7.602
9. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behavioral and Brain Sciences. 2010; 33(2-3):61–83. https://doi.org/10.1017/S0140525X0999152X PMID: 20550733
10. Gigerenzer G. Mindless statistics. The Journal of Socio-Economics. 2004; 33(5):587–606. https://doi.org/10.1016/j.socec.2004.09.033
11. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician. 2006; 60(4):328–331. https://doi.org/10.1198/000313006X152649
12. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011; 22(11):1359–1366. https://doi.org/10.1177/0956797611417632 PMID: 22006061
13. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349(6251):943–953. https://doi.org/10.1126/science.aac4716
14. Ioannidis JP. Why most published research findings are false. PLoS Medicine. 2005; 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
15. Gelman A, Carlin J. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science. 2014; 9(6):641–651. https://doi.org/10.1177/1745691614551642 PMID: 26186114
16. Greenland S. Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values. The American Statistician. 2019; 73(sup1):106–114. https://doi.org/10.1080/00031305.2018.1529625
17. Cohen J. Things I have learned (so far). American Psychologist. 1990; 45(12):1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304
18. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London Series A, Mathematical and Physical Sciences. 1937; 236(767):333–380.
19. Cox DR, Hinkley DV. Theoretical statistics. CRC Press; 1979.

20. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological methods. 1996; 1(2):115–129. https://doi.org/10.1037/1082-989X.1.2.115

21. Cumming G. The new statistics: Why and how. Psychological Science. 2014; 25(1):7–29. https://doi.org/10.1177/0956797613504966 PMID: 24220629

22. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science. 2018; 1(2):259–269. https://doi.org/10.1177/2515245918770963

23. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. Trials. 2011; 12(1):1–12. https://doi.org/10.1186/1745-6215-12-106 PMID: 21539749

24. Johnson VE. Revised standards for statistical evidence. Proceedings of the National Academy of Sciences. 2013; 110(48):19313–19317. https://doi.org/10.1073/pnas.1313476110 PMID: 24218581

25. Ioannidis JP. The proposal to lower P value thresholds to .005. Jama. 2018; 319(14):1429–1430. https://doi.org/10.1001/jama.2018.1536 PMID: 29566133

26. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2018; 2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z PMID: 30980045

27. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic bulletin & review. 2009; 16(2):225–237. https://doi.org/10.3758/PBR.16.2.225 PMID: 19293088

28. Wagenmakers EJ, Wetzels R, Borsboom D, Van Der Maas HL. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). Journal of Personality and Social Psychology. 2011; 100(3):426–432. https://doi.org/10.1037/a0022790 PMID: 21280965

29. Goodman SN. Of P-values and Bayes: a modest proposal. Epidemiology. 2001; 12(3):295–297. https://doi.org/10.1097/00001648-200105000-00006 PMID: 11337600

30. Woolston C. Psychology journal bans P values. Nature. 2015; 519(7541):9–9. https://doi.org/10.1038/519009f

31. Lakens D. So you banned p-values, how's that working out for you?; 2016. Available from: https://daniellakens.blogspot.com/2016/02/so-you-banned-p-values-hows-that.html.

32. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biology. 2015; 13(3):e1002106. https://doi.org/10.1371/journal.pbio.1002106 PMID: 25768323

33. Kerr NL. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review. 1998; 2(3):196–217. https://doi.org/10.1207/s15327957pspr0203_4 PMID: 15647155

34. Rosenthal R. The file drawer problem and tolerance for null results. Psychological Bulletin. 1979; 86 (3):638–641. https://doi.org/10.1037/0033-2909.86.3.638

35. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science. 2012; 23(5):524–532. https://doi.org/10.1177/0956797611430953 PMID: 22508865

36. Fox N, Honeycutt N, Jussim L. Better Understanding the Population Size and Stigmatization of Psychologists Using Questionable Research Practices. Meta-Psychology. 2022; 6. https://doi.org/10.15626/MP.2020.2601

37. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proceedings of the National Academy of Sciences. 2018; 115(11):2600–2606. https://doi.org/10.1073/pnas.1708274114 PMID: 29531091

38. Scheel AM, Schijen MR, Lakens D. An excess of positive results: Comparing the standard Psychology literature with Registered Reports. Advances in Methods and Practices in Psychological Science. 2021; 4(2):25152459211007467. https://doi.org/10.1177/25152459211007467

39. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science. 2016; 11(5):702–712. https://doi.org/10.1177/1745691616658637 PMID: 27694465

40. Cramer AO, van Ravenzwaaij D, Matzke D, Steingroever H, Wetzels R, Grasman RP, et al. Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. Psychonomic bulletin & review. 2016; 23(2):640–647. https://doi.org/10.3758/s13423-015-0913-5 PMID: 26374437

41. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics. 1979; p. 65–70.

42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995; 57(1):289–300. https://doi.org/10.1038/519009f

43. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. Journal of research on educational effectiveness. 2012; 5(2):189–211. https://doi.org/10.1080/19345747.2011.618213

44. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: https://www.R-project.org/.

45. Böschen I. JATSdecoder: A Metadata and Text Extraction and Manipulation Tool Set; 2022. Available from: https://CRAN.R-project.org/package=JATSdecoder.

46. Böschen I. Software review: The JATSdecoder package—extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed Central's open access database. Scientometrics. 2021. https://doi.org/10.1007/s11192-021-04162-z PMID: 34720253

47. Böschen I. Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. Scientific Reports. 2021; 11. https://doi.org/10.1038/s41598-021-98782-3 PMID: 34593888

48. Böschen I. Evaluation of the extraction of methodological study characteristics with JATSdecoder. Scientific Reports. 2023; 13. https://doi.org/10.1038/s41598-022-27085-y PMID: 36599903

49. Tkaczyk D, Szostek P, Fedoryszak M, Dendek PJ, Bolikowski Ł. CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR). 2015; 18(4):317–335. https://doi.org/10.1007/s10032-015-0249-8

50. National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM). Journal Publishing Tag Library—NISO JATS Draft Version 1.1d2; 2014. https://jats.nlm.nih.gov/publishing/tag-library/1.1d2/index.html.

51. Arel-Bundock V, Enevoldsen N, Yetman C. countrycode: An R package to convert country names and country codes. Journal of Open Source Software. 2018; 3(28):848. https://doi.org/10.21105/joss.00848

52. Bengtsson H. future.apply: Apply Function to Elements in Parallel using Futures. 2020; R package version 1.4.0.

53. Rubin M. When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. Synthese. 2021; 199(3):10969–11000. https://doi.org/10.1007/s11229-021-03276-4

14 Danksagung

Für das Vertrauen und die finanzielle Unterstützung durch ein Stipendium des Arbeitsbereichs Forschungsmethoden und Statistik des Fachbereichs Psychologie möchte ich mich ganz herzlich bei Prof. Martin Spieß bedanken. Ohne seinen Rückhalt und die Unterstützung wären weder diese Arbeit, noch das R-Paket *JATSdecoder* realisiert worden. Für die präzisen Überarbeitungen der englischsprachigen Veröffentlichungen bin ich Marcella Dudley zu äußerstem Dank verpflichtet. Weiterhin gilt mein Dank Dennis Warnholz, der bei der Einrichtung der öffentlich zugänglichen Server und der Kommunikation mit dem Rechenzentrum nie die Geduld verlor und für meine technischen Problemen stets ein offenes Ohr hatte.

Insbesondere möchte meiner Familie, Laura, Jette und Tammo danken, die mir, in den uns allen belastenden Pandemie-Zeiten, mit nicht enden wollenden Lockdown-Maßnahmen und Homeschooling, immer auch genügend Zeit für die Arbeit an meinem Projekt gaben, mich motivierten und stets für anregende Abwechslung zur einsamen und teils frustrierenden Computerarbeit sorgten. Zuletzt möchte ich meinen Eltern dafür danken, dass sie mich stets darin bestärkt haben, meine selbstgesteckten Ziele zu verwirklichen.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR PSYCHOLOGIE UND
BEWEGUNGSWISSENSCHAFT
Institut für Bewegungswissenschaft
Institut für Psychologie

Erklärung gemäß *(bitte Zutreffendes ankreuzen)*

☐ **§ 4 (1c) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 5 (4d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich,

Ingmar Böschen (Vorname, Nachname),

dass ich mich an einer anderen Universität oder Fakultät noch keiner Doktorprüfung unterzogen oder mich um Zulassung zu einer Doktorprüfung bemüht habe.

Hamburg, den ____________ ____________________

Ort, Datum Unterschrift

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Fakultät für
Psychologie und
Bewegungswissenschaft

Institut für Bewegungswissenschaft
Institut für Psychologie

Eidesstattliche Erklärung nach *(bitte Zutreffendes ankreuzen)*

□ **§ 7 (4) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 9 (1c und 1d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich an Eides statt,

1. dass die von mir vorgelegte Dissertation nicht Gegenstand eines anderen Prüfungsverfahrens gewesen oder in einem solchen Verfahren als ungenügend beurteilt worden ist.
2. dass ich die von mir vorgelegte Dissertation selbst verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die wörtlich oder inhaltlich übernommenen Stellen habe ich als solche kenntlich gemacht.

Hamburg, den ____________________ ____________________________

Ort, Datum Unterschrift