Dr. habil. Heike Diefenbach

„Peer Reviewed“ – kein Qualitätssiegel!

ScienceFiles Blaue Reihe Band 7

©2020; Dr. habil. Heike Diefenbach
http://sciencefiles.org

Zitate und auszugsweise Verwendung von Teilen dieses Buches sind nur unter Angabe der Quelle erlaubt. Der Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung ist ohne Zustimmung des Autoren unzulässig. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmung und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Zitiervorschlag:
Diefenbach, Heike (2020). „Peer Reviewed“ – kein Qualitätssiegel. Llanelli: ScienceFiles, Blaue Reihe, Band 7.

Inhalt

1. „Peer reviewed” – kein Qualitätssiegel!

Im Zusammenhang mit der These vom menschengemachten Klimawandel und neuerdings im Zusammenhang mit Untersuchungen zu Herkunft und Verlauf von Erkrankungen durch den neuen Corona-Virus, der nunmehr von der WHO die Bezeichnung SARS-CoV-2 erhalten hat, wird wieder verstärkt bemerkt, dass es sich bei einem Text um einen „peer reviewed“ Text handle oder ein Text nicht „peer reviewed“ sei. Suggeriert werden soll damit, dass „peer reviewed“ ein Qualitätssiegel sei, d.h. dasjenige, was in Texten steht, die „peer reviewed“ sind, wäre zuverlässig oder zumindest glaubhaft oder zuverlässigerer oder glaubhafter als dasjenige, was in Texten steht, die nicht „peer reviewed“ sind. Tatsächlich hat sich die Formel „peer reviewed“ zu einer „heiligen Kuh“ der Wissenschaftlichkeit entwickelt, wie de Vries (2001: 231) es (in englischer Sprache) ausdrückt, zeigt das Merkmal „peer reviewed“ doch an, dass der Text, der das Merkmal trägt, einem „peer-reviewing“ oder einem Prozess der Begutachtung durch Fachkollegen unterzogen worden ist.

Was genau dieser Prozess, sofern er tatsächlich stattfindet, für einen bestimmten Text beinhaltet hat, ist aber nicht nur den Fachvertretern normalerweise unbekannt, sondern auch und vor allem all denen, die nicht Fachvertreter sind und die im Text dargestellten Erkenntnisse als „Normal“bürger ohne besondere fachliche Kenntnis als wahr und relevant akzeptieren sollen, d.h. glauben sollen, was im Text steht. Insofern ist das Label „peer reviewed“ ein Prototyp des Fehlschlusses ad auctoritatem, bei dem etwas nur deshalb für richtig, zuverlässig oder zumindest glaubhaft gehalten wird, weil es von einer Institution, hier: den Herausgebern von Fachzeitschriften, für „gut“ befunden wurde, die ihrerseits Anspruch auf Zuverlässigkeit oder Glaubhaftigkeit erhebt – und Zuverlässigkeit oder Glaubhaftigkeit von vielen Menschen gewohnheitsmäßig zugesprochen bekommt.

2. „Peer reviewing“: die Idee

Bei einem Text, der „peer reviewed“ ist, ist die Quelle, der der Text seine Zuverlässigkeit oder Glaubwürdigkeit verdanken soll, der Herausgeber bzw. der Herausgeberkreis, der den Text publiziert hat. Dabei wird angenommen, dass der/die Herausgeber ein „peer reviewing“ durchgeführt haben.

„Peer reviewing“ bezeichnet im Kern die Prüfung eines Manuskriptes vor seiner Veröffentlichung durch sogenannte „peers“, d.h. gewöhnlich Fachkollegen, und besonders durch solche Fachkollegen, die über dasselbe Thema oder ein ähnliches Thema gearbeitet haben oder arbeiten wie das, das im Manuskript behandelt wird. Die Annahme dahinter ist, dass solche „peers“ auf der Höhe der Diskussion über das Thema sind, die relevanten Theorien und Daten kennen, die Methoden im Umgang mit Forschung über dieses Thema und speziell mit den Daten, die für das Thema relevant sind, beherrschen und daher die Güte eines Textes beurteilen können, der dieses Thema bearbeitet. „Güte“ soll dabei nicht nur Freiheit von offensichtlichen Fehlern in den Daten oder in der Logik der vorgebrachten Argumentation bedeuten, sondern auch Verständlichkeit, Vollständigkeit der Darstellung vorhergehender Theoriebildung und Datenauswertungen zum Thema u.a.m.

Manuskripte von nicht ausreichender Güte sollen durch die Prüfung durch Fachkollegen als solche identifiziert und von einer Veröffentlichung – zumindest einer Veröffentlichung in der vorliegenden Form – ausgeschlossen werden. Die Idee ist also, dass Manuskripte, die es bis zum Druck oder zur digitalen Veröffentlichung geschafft haben, solche von hoher Güte oder wissenschaftlicher Qualität sind, eben weil sie den kritischen Begutachtungsprozess durch Personen mit entsprechender Kompetenz „überlebt“ haben.

Ein „peer reviewing“ wird normalerweise von Herausgebern von wissenschaftlichen Fachzeitschriften (bzw. solchen, die es sein wollen,) verwendet, aber nicht nur von ihnen. Es kann z.B. verwendet werden, wenn es darum geht, zu entscheiden, welche Papiere aus einer Menge bei einem Tagungsveranstalter eingereichter Papiere auf der Tagung vorgetragen werden sollen und welche nicht, oder wenn Herausgeber eines Sammelbandes unter einer Menge eingereichter Beiträge eine bestimmte Anzahl von Beiträgen auswählen müssen, die im Sammelband untergebracht werden kann. In jedem Fall soll das „peer reviewing“ sicherstellen, dass Texte, die ein Herausgeber(-/kreis) in der

von ihm herausgegebenen Fachzeitschrift druckt bzw. digital veröffentlicht, oder Vortragspapiere, die auf einer Fachtagung vorgetragen werden, qualitätvolle Arbeiten sind, d.h. bestimmten Ansprüchen gerecht werden.

Die Veröffentlichung möglichst vieler „peer reviewed“ Arbeiten galt früher als das „sine qua non“ wissenschaftlicher Leistung, in einer wissenschaftlichen Welt des „publish or perish“, wie Hojat, Gonnella und Caelleigh (2003: 75) sagen, aber seit etwa der zweiten Hälfte der 1980er-Jahre (s. z.B. Bailar & Patterson 1985; Hargens 1988; Horrobin 1982) und insbesondere seit dem Ende der 1990er-Jahre ist das „peer reviewing“ verstärkt selbst zum Untersuchungsgegenstand gemacht worden, und das vermeintliche Qualitätsmerkmal „peer reviewed“ hat aufgrund der Ergebnisse dieser Untersuchungen stark gelitten. Für das Verständnis der Probleme, die diese Untersuchungen mit Bezug auf das „peer reviewing“ identifiziert haben, ist es wichtig, dass man eine Vorstellung davon hat, wie die Praxis oder korrekt: die Praktiken des „peer reviewing“ aussehen.

3. „Peer reviewing“: die Praxis

Wenn ein Manuskript bei einem Herausgeber(-/kreis) einer Fachzeitschrift mit einem „peer review“-Verfahren zur Veröffentlichung eingereicht wird, dann teilt er den Autoren des Manuskriptes mit, dass das Manuskript eingegangen ist und nunmehr zur Prüfung an „peers“ weitergegeben wird – normalerweise, denn es ist auch möglich, dass Herausgeber aufgrund ihrer ersten Durchsicht eingereichter Manuskripte einige verwerfen, ohne sie überhaupt nach außen, an „peers“, zur Begutachtung weiterzugeben.

Wieviele „peers“ das Manuskript prüfen, bleibt normalerweise unbestimmt. Manche Herausgeber(-/kreise) von Fachzeitschriften geben an, Manuskripte zur Prüfung an mindestens zwei, möglichst drei oder mehr, Fachkollegen zu übergeben, andere schweigen sich diesbezüglich aus. Der/die Herausgeber verschicken dann das zu prüfende Manuskript an „peers“ mit der Bitte um Begutachtung, oder sie schreiben einen Fachkollegen zunächst an, um seine Bereitschaft zu erfragen, als Gutachter allgemein oder für ein spezielles Manuskript tätig zu werden; wenn der Fachkollege seine Bereitschaft erklärt, bekommt er das Manuskript geschickt.

Wenn der Fachkollege das Manuskript erhält, trägt es keine/n Autorennamen mehr, und gewöhnlich erhält er auch einen Hinweis von dem/den Herausgeber/n darauf, dass diejenigen Publikationen aus der Literaturliste im Manuskript entfernt wurden, die von dem Autor bzw. den Autoren des Manuskriptes stammen. Umgekehrt wird dem/den Autoren des Manuskriptes nicht mitgeteilt, welche Fachkollegen es sind, die sein/ihr Manuskript zur Prüfung zugeschickt bekommen haben. Gegenseitige Anonymität oder „double masking“ oder „blinding“ soll sicherstellen, dass die prüfenden Fachkollegen, Gutachter genannt, bei der Prüfung nur das Manuskript prüfen und nichts anderes, d.h. sich nicht von Sympathien oder Antipathien gegenüber den Autoren leiten lassen. Die gegenseitige Anonymität soll auch ausschließen, dass berufsstrategische Erwägungen zum Tragen kommen, z.B. dann, wenn ein Gutachter einen Autor als direkten Konkurrenten um Positionen oder Fördergelder wahrnimmt und er verhindern möchte, dass dieser Autor etwas Neues oder Wichtiges veröffentlicht oder überhaupt einen weiteren Eintrag in seine Veröffentlichungsliste bekommt, oder wenn ein Nachwuchswissenschaftler ein Manuskript zur Prüfung erhält, das von einem „gestandenen“ Fachkollegen verfasst wurde, der für die weitere berufliche Karriere des Nachwuchswissenschaftlers noch wichtig sein könnte, so dass der Nachwuchswissenschaftler sich scheuen könnte, das Manuskript des „gestandenen“ Kollegen so zu kritisieren, wie er das eigentlich für angemessen halten würde.

Es sei hier vorweggenommen, dass Justice et al. eine Studie vorgelegt haben, die gezeigt hat, dass „[m]asking reviewers to author identity as commonly practiced does not improve quality of reviews“ (Justice et al. 1998: 240). Auch van Rooyen et al. (1998: 234) haben in ihrer Untersuchung festgestellt, dass

“[b]linding and unmasking made no editorially significant difference to review quality, reviewers' recommendations, or time taken to review. Other considerations should guide decisions as to the form of peer review adopted by a journal, and improvements in the quality of peer review should be sought via other means”.

Dies stimmt mit den Befunden von Godlee, Gale und Martyn (1998: 237) überein:

“Neither blinding reviewers to the authors and origin of the paper nor requiring them to sign their reports had any effect on rate of detection of errors. Such measures are unlikely to improve the quality of peer review reports”, während Isenberg, Sanchez und Zafran (2009: 881) aus ihrer Studie berichten, dass immerhin „[d]ouble-masking may [!] improve the quality of biomedical publishing or at least reduce reviewer bias for effectively masked manuscripts”.

Ich habe es in meiner Funktion als Gutachterin für verschiedene deutsch- oder englischsprachige Fachzeitschriften sowie die Deutsche Forschungsgemeinschaft und andere Forschungsförderer immer so gehalten, dass ich dem/den Herausgeber/n gegenüber erklärt habe, dass ich keinen Wert auf meine Anonymität lege, weil ich glaube, dass ein direkter, offener Austausch deutlich nutzenbringender ist als ein anonymisiertes feedback in „Einbahnstraßen“-Manier, aber ich weiß von keinem Fall, in dem man meinem Angebot entsprochen hätte, weder in deutsch- noch in englischsprachigen Fachzeitschriften.

Gewöhnlich haben „peers“ einige Wochen Zeit, um das Manuskript zu prüfen. Bei manchen Fachzeitschriften kann die Prüfung bis heute mehr oder weniger formlos erfolgen, und tatsächlich war das in vergangenen Jahrzehnten die Regel. In den letzten Jahren haben aber immer mehr Fachzeitschriften die Notwendigkeit erkannt, „peers“ für die Prüfung der Manuskripte Kriterien vorzugeben, auf die die „peers“ zumindest kurz schriftlich eingehen sollen. Solche Kriterien können u.a. sein: „Relevanz der Forschung“, „Neuheitswert der Forschung“, „Hinreichende Aufarbeitung von Vorgängerforschung“, „Stringenz der Argumentation“. Immer werden „peers“ um eine Empfehlung darüber gebeten, ob das Manuskript in dieser Form bzw. mit kleineren Änderungen in der Fachzeitschrift veröffentlicht werden soll, ob es im Prinzip in der Fachzeitschrift veröffentlicht werden soll, aber größere Überarbeitungen erfordert, oder ob es nicht in der Fachzeitschrift veröffentlicht werden soll bzw. vom „peer“ als nicht zur Veröffentlichung in der Fachzeitschrift geeignet eingestuft wird.

„Peers“ verfassen ihr Gutachten und schicken es an den/die Herausgeber, die die Gutachten idealerweise lesen und die Empfehlungen der Gutachter zur Grundlage ihrer Entscheidung darüber machen, ob sie das in Frage stehende Manuskript veröffentlichen wollen oder nicht, und falls ja, ob sie sich Änderungen vom Autor/von den Autoren

wünschen oder sie gar zur Bedingung für die Veröffentlichung machen wollen oder nicht. Herausgeber sind nicht an die Empfehlungen der Gutachter gebunden. Es liegt in ihrem eigenen Ermessen, ob sie den Empfehlungen in welcher Weise folgen wollen. Wenn sich Herausgeber über die Empfehlungen von Gutachtern hinwegsetzen, dann häufig in dem Fall, dass mehrere Gutachter am Prozess beteiligt waren und widersprüchliche Empfehlungen gegeben haben, aber sie können sich auch in anderen Fällen über Empfehlungen von Gutachtern hinwegsetzen und machen von dieser Möglichkeit auch Gebrauch.

Wenn in einem Herausgeberkreis über die Veröffentlichung eines Manuskriptes entschieden werden muss, dann kann das im Rahmen sogenannter Herausgebersitzungen oder sonstwie als gemeinsame Entscheidung der Herausgeber passieren, oder einem Herausgeber wird die (Haupt-/)Zuständigkeit für bestimmte Manuskripte zugeteilt.

Wie genau Herausgeber sich über die Veröffentlichung oder Nicht-Veröffentlichung eines Manuskriptes verständigen, welchen Stellenwert Gutachten von Fachkollegen tatsächlich haben, ob der Stellenwert der Gutachten bestimmter Fachkollegen bei bestimmten Herausgebern höher ist als der anderer, wie ggf. der Diskussionsprozess zwischen Herausgebern aussieht, dies alles ist sehr weitgehend unbekannt.

Unbekannt bleibt in aller Regel auch, wie Herausgeber „peers“ auswählen. Die Vermutung, dass die Auswahl von „peers“ in Entsprechung zum genauen Thema des Manuskriptes erfolgt, also als tatsächliche „peers“ im engen Sinn, und sonst nichts, ist plausibel, mag aber ein rationalistisches Vorurteil sein. Man kann sich viele Gründe, vor allem „netzwerk“strategische, dafür vorstellen, warum Herausgeber bestimmten Fachkollegen Manuskripte bestimmter Autoren zur Begutachtung vorlegen oder sie bestimmten Gutachten ggf. mehr Gewicht beimessen als anderen.

„Peers“ wird gewöhnlich nicht offiziell mitgeteilt, zu welcher Entscheidung der/die Herausgeber gekommen sind, aber natürlich kann ein „peer“ diesbezüglich später seine eigenen Schlüsse ziehen, wenn er beobachtet, ob das Manuskript, das er geprüft hat, in der Fachzeitschrift veröffentlicht wurde oder nicht (oder woanders veröffentlicht wurde). Ich weiß auch von „peer reviewing“-Verfahren, bei denen die „peers“ der/die Herausgeber

selbst war/waren, also keine Weitergabe des Manuskriptes an Fachkollegen außerhalb des Herausgeberkreises stattgefunden hat.

Schon diese kurze Beschreibung der Praxis sollte deutlich gemacht haben, dass es sehr einfältig ist, von „dem“ „peer reviewing“-Verfahren zu sprechen. In der Realität gibt es viele verschiedene „peer reviewing“-Verfahren, die sich von Zeitschrift zu Zeitschrift und von Fall zu Fall voneinander unterscheiden können. Wenn man weiß, dass ein Text „peer reviewed“ ist, dann weiß man daher ziemlich wenig darüber, wie und von wem und woraufhin der Text tatsächlich wie gut geprüft wurde.

Selbst dann, wenn ein mit bester Absicht und optimal durchgeführtes „peer reviewing“-Verfahren durchgeführt wurde, hängt die Güte des Ergebnisses des Verfahrens davon ab, wie kompetent, aufmerksam und aufrichtig die „peers“ und die Herausgeber selbst sind. Aber „Who Reviews the Reviewers?“ (Baxt et al. 1998).

4. „Peer-reviewing“ auf dem empirischen Prüfstand

Die Untersuchungen über „peer-reviewing“, die bislang vorliegen, zeigen erhebliche Mängel der entsprechenden Verfahren in verschiedenen Hinsichten. Im Folgenden berichte ich empirische Befunde aus einer Reihe von Studien, die relativ bekannt sind und relativ einfach gefunden werden können, so dass der Leser sie leicht finden und selbst lesen kann. Untersuchungen über „peer reviewing“ sind inzwischen so zahlreich, dass es kaum möglich ist – jedenfalls nicht mir im Rahmen dieses Textes – eine weitgehend vollständige Darstellung aller existierenden Studien oder Befunde zum Thema „peer reviewing“ zu geben. Der interessierte Leser sei daher auf eigene weiterführende Recherche und Lektüre verwiesen. Er wird ggf. dabei aber feststellen, dass die weiteren Befunde, die er zusammentragen wird, sehr weitgehend im Einklang mit denen stehen, die ich im Folgenden ansprechen werde.

Eine weitere Konsequenz aus der Vielzahl der Studien zum „peer reviewing“, die inzwischen vorliegen, ist, dass man, wenn man einen Überblick über die Befunde geben möchte, sie um der Übersichtlichkeit willen zu ordnen versuchen muss, d.h. sie

verschiedenen Themenbereichen oder Aspekten des „peer reviewing“ zuzuordnen (wobei man die ein oder andere Studie ebenso gut einem anderen Bereich bzw. mehr als einem Bereich zuordnen könnte; es geht hier wirklich nur um Effizienz der Darstellung). In der folgenden Darstellung werden fünf Bereiche unterschieden: Der erste Bereich betrifft die hohe Inkonsistenz der Beurteilung von Manuskripten durch Herausgeber oder Gutachter, der zweite Bereich betrifft „biases“ oder Verzerrungseffekte, die durch persönliche Eigenschaften oder Präferenzen von Gutachtern oder Herausgebern zustande kommen oder durch journalistisches statt wissenschaftliches Denken bzw. weitverbreitete Vorurteile darüber, was berichtenswert sei und was nicht. Die mangelnde Fähigkeit von Herausgebern oder Gutachtern, innovative Arbeiten als solche zu erkennen und zu würdigen, ist der dritte Bereich, der im Folgenden betrachtet wird. Der vierte betrifft die mangelnde Fähigkeit von Gutachtern oder Herausgebern, Unsinn als solchen zu erkennen und von der Veröffentlichung auszuschließen, und der letzte Bereich betrifft den – nicht immer leicht zu markierenden– Übergang vom Mangel zum Betrug.

4.1 „Peer reviewing“ gleicht einer Lotterie: Erhebliche Inkonsistenz bei der Beurteilung von Manuskripten

Baxt et al. (1998), die die Qualität von Gutachten durch den Einsatz eines absichtlich mit Fehlern, darunter schwerwiegenden methodischen Fehlern (u.a. eine fehlerhafte statistische Analyse), behafteten, fiktiven Manuskriptes, das an alle Gutachter der Annals of Emergency Medicine, verschickt wurde, überprüft haben, haben u.a. festgestellt, dass in den 203 Gutachten, die sie erhielten,

„[o]nly 9 reviewers identified the 2 existing reports on the use of propranolol for migraine headaches, and only 3 reviewers identified the [two] fictitious references. Thirty-one percent of the reviewers identified statistical errors, and 14.8% of the reviewers (30) misspelled propranolol throughout their reviews” (315),

und

“[t]he number of years since training, the number of other journals reviewed for, and the number of reviews over the last year were not associated with identifying a greater number of major or minor errors. The only statistically significant difference detected was that

reviewers at the assistant professor level identified more minor errors than associate professors or professors (Baxt et al. 1998: 315).

In ihrer zusammenfassenden Schlussfolgerung schreiben Baxt et al.:

„On the basis of the results of this study, 1 set of reviewers from 1 specialty failed to identify the majority of major errors placed in such a manuscript and 68% failed to realize that the conclusions were not supported by the data” (Baxt et al. 1998: 316).

Aber „[i]t is not clear that these results can be generalized” (Baxt et al. 1998: 316). Um zu prüfen, inwieweit diese Befunde verallgemeinerbar sind, müssten entsprechende Untersuchungen für Fachzeitschriften bzw. Gutachter in vielen verschiedenen wissenschaftlichen Disziplinen und Sub-Disziplinen durchgeführt werden. Bislang gilt aber, dass es vor allem die Medizin ist, in der man sich bemüht, die Qualität des vermeintlich per se qualitätvollen „peer reviewing“ zu untersuchen und Verbesserungsvorschläge zu formulieren und zu testen, und dies vor allem in den USA und anderen englischsprachigen Ländern der Fall ist. Eine offene Diskussion über den Zustand des „peer reviewing“ im eigenen Fach hat m.W. bislang ebenfalls nur oder vor allem in der Medizin stattgefunden. Nicht nur die Vielzahl der Publikationen zum Thema in fachmedizinischen Zeitschriften zeigen das, sondern u.a. auch die Tatsache, dass mehrere internationale Kongresse über „Peer Review in Biomedical Publications“ vom Journal of the American Medical Association (JAMA) und der British Medical Journal Publishing Group organisiert und durchgeführt wurden (Hojat et al. 2003: 78).

Untersuchungen zum „peer reviewing“ im Fachbereich Medizin ergeben regelmäßig ein sehr ernüchterndes Bild von der Qualität des „peer reviewing“, so z.B. die die Studie von Kravitz et al., die aus dem Jahr 2010 stammt. In der Studie untersuchten die Autoren die Empfehlungen von 5.881 Gutachtern für 2.264 Manuskripte – für die meisten Manuskripte wurden drei Gutachten eingeholt –, die beim Journal of General Internal Medicine (JGIM) in den Jahren 2004 und 2008 eingereicht wurden und von den Herausgebern an externe Gutachten weitergegeben wurden – das waren nur 36 Prozent aller in den beiden Jahren 2004 und 2008 bei der Zeitschrift eingereichten Manuskripte! Kravitz et al. errechneten statistische Maße für die Übereinstimmung der

Gutachterempfehlungen mit Bezug auf das jeweils selbe Manuskript und kamen zum folgenden Ergebnis:

„Among the 2264 manuscripts reviewed during the study period, just under half received reviews that were in complete agreement not to reject (i.e., all reviewers recommended accept/revise), less than 10% received reviews that were in complete agreement to reject, and the balance received reviews with conflicting recommendations … The editors rejected 48% of 2264 manuscripts sent out for external peer-review. If all reviewers recommended not to reject, editors rejected the manuscript 20% of the time. If all reviewers recommended ‘reject’, editors rejected 88% of the time. And if reviewers were divided, editors rejected the manuscript 70% of the time … „The results of this analysis suggest that reviewers for JGIM agreed on the disposition of manuscripts at a rate barely exceeding what would be expected by chance” (Kravitz et al. 2010: 3).

Die Befunde, die bislang aus anderen Fachbereichen, vorliegen, sprechen dafür, dass dort die Qualität des „peer reviewing“ – zumindest! – nicht besser ist als in der Medizin. So haben z.B. Peters und Ceci (1982) in einem frühen Experiment 12 Texte von Forschern ausgewählt, die an bekannten Psychologie-Abteilungen verschiedener Universitäten arbeiteten und diese Texte bereits in psychologischen Fachzeitschriften veröffentlicht hatten, die sich hohen Ansehens erfreuten und eine Ablehnungsrate von Manuskripten von 80 Prozent hatten. Peters und Ceci reichten dieselben Texte nach 18 bis 32 Monaten bei denselben Zeitschriften wieder ein, verwendeten dabei aber fiktive Namen und Einrichtungen. Sie stellten fest, dass von den insgesamt 38 Herausgebern oder Gutachtern, die mit den wiedereingereichten Texten beschäftigt waren, nur drei (bzw. 8 Prozent) die Texte als Wiedereinreichungen erkannten, und von den verbleibenden neun Texten, d.h. den Texten, die nicht als Wiedereinreichung unter anderen Autorennamen erkannt wurden, wurden acht von den Gutachtern oder Herausgebern abgelehnt, meist aufgrund schwerwiegender methodischer Mängel (Peters & Ceci 1982: 187) – und dies bei Texten, die ja bereits in genau dieser Zeitschrift veröffentlicht worden waren, ohne dass sich die Herausgeber oder Gutachter hierüber bewusst waren.

Dieser Befund lässt nicht nur Zweifel daran aufkommen, dass Repräsentanten eines Faches tatsächlich zumindest mehrheitlich die einschlägigen Fachzeitschriften lesen und ggf. die Inhalte erinnern, sondern weist auch darauf hin, dass es – wie in der Studie von

Kravitz et al. (2010) – einer Lotterie ähnelt, ob man in einer bestimmten Zeitschrift einen Text veröffentlicht bekommt oder nicht, je nachdem, auf welche/n Gutachter man trifft.

Das „Lotterie-Element“ im „peer reviewing“ haben auch Neff und Olden, diesmal mit Bezug auf Fachzeitschriften im Bereich der Biologie, beobachtet:

„Here we use probability theory to model the peer-review process, focusing on two key components: (1) editors' prescreening of submitted manuscripts and (2) the number of referees polled. The model shows that the review process can include a strong “lottery” component, independent of editor and referee integrity. Focusing on journal publications, we use a Bayesian approach and citation data from biological journals to show that top journals successfully publish suitable papers—that is, papers that a large proportion of the scientific community would deem acceptable—by using a prescreening process that involves an editorial board and three referees; even if that process is followed, about a quarter of published papers still may be unsuitable. The element of chance is greater if journals engage only two referees and do no prescreening (or if only one editor prescreens); about half of the papers published in those journals may be unsuitable. Furthermore, authors whose manuscripts were initially rejected can significantly boost their chances of being published by resubmitting their papers to other journals” (Neff & Olden 2006: 333).

Justice et al. (1994) haben mit Hilfe einer 10-Punkte-Skala das Ausmaß der Übereinstimmung zwischen Gutachtern, zufällig ausgewählten Lesern und Experten für klinische Forschungsmethoden mit Bezug auf 113 Manuskripte, die bei den Annals of Internal Medicine eingereicht wurden, untersucht. Die Autoren stellten fest:

„Readers and peers gave high grades (77% and 73% gave a grade of 5 or better, respectively), while experts were more critical (52% gave a grade of 5 or better; $P < .0001$). Agreement was relatively high among judge groups (in all cases, > 69%) *but agreement beyond chance was poor* (kappa < 0.04). One third of readers (33%) thought that the manuscript had little relevance to their work” (Justice et al. 1994: 117; Hervorhebung d.d.A.).

Schließlich sei auf die Studie von Rothwell und Martyn (2000) hingewiesen, die ebenfalls festgestellt haben, dass die Auswahl von Manuskripten durch peer reviewing einer Lotterie gleicht. Die Autoren haben in ihrer Untersuchung der Gutachten zu Manuskripten, die bei zwei verschiedenen Zeitschriften aus dem Bereich der klinischen Neurowissenschaft eingereicht wurden, beobachtet, dass

„[a]greement between reviewers as to whether manuscripts should be accepted, revised or rejected was not significantly greater than that expected by chance … for 179 consecutive papers submitted to Journal A, and was poor …. for 116 papers submitted to Journal B” (Rothwell & Martyn 2000: 1964).

Die Autoren prüften darüber hinaus die Übereinstimmung zwischen Gutachte(r)n mit Bezug auf Zusammenfassungen von Arbeiten, mit denen sich deren Verfasser um einen entsprechenden Vortrag auf Fachkonferenzen bewarben:

„Abstracts submitted for presentation at the conferences were given a score of 1 (poor) to 6 (excellent) by multiple independent reviewers. For each conference, analysis of variance of the scores given to abstracts revealed that differences between individual abstracts accounted for only 10-20% of the total variance of the scores” (Rothwell & Martyn 2000: 1964),

während

„[o]ver a quarter of the variance in abstract scores (27% for Meeting A and 32% for Meeting B) could be accounted for by the tendency for some reviewers to give higher or lower scores than others” (Rothwell & Martyn 2000: 1966.

Die Autoren kommen aufgrund dieser Befunde zu der folgenden Schlussfolgerung:

„Thus, although recommendations made by reviewers have considerable influence on the fate of both papers submitted to journals and abstracts submitted to conferences, agreement between reviewers in clinical neuroscience was little greater than would be expected by chance alone” (Rothwell & Martyn 2000: 1964).

4.2 Verzerrungseffekte („biases“) bei Gutachtern oder Herausgebern

Mit der Tendenz bestimmter Gutachter, Manuskripte insgesamt eher positiv oder negativ zu bewerten, die Rothwell und Martyn beobachtet haben, ist schon ein Beleg dafür erbracht, dass die Beurteilung von Manuskripten keineswegs (nur) aufgrund objektivierbarer und replizierbarer sachlicher Kriterien erfolgt, sondern (auch) von den Vorlieben oder Abneigungen von Gutachtern geprägt ist. Sie resultieren in Verzerrungseffekten oder „biases“ bei der Begutachtung von Manuskripten, weil sie „... results [produzieren] that depart systematically from the true values” (Murphy 1976: 239), wobei „true values“ von Manuskripten hier als die Bewertung aufzufassen ist, die Manuskripte erhalten würden, wenn sie allein aufgrund objektivierbarer und replizierbarer sachlicher Kriterien beurteilt würden.

Owen (1982) hat eine ganze Reihe von solchen Verzerrungseffekten (nicht nur bei Gutachtern, sondern Lesern von Fachaufsätzen allgemein,) aufgelistet, die man in verschiedene Bereiche unterteilen bzw. entsprechend zusammenfassen kann (vgl. hierzu die kurze Zusammenfassung bei Weller 2002: 208), u.a. in methodische Verzerrungseffekte, die z.B. dann vorliegen, wenn ein Gutachter/Leser bestimmte Auswertungsverfahren bevorzugt oder ablehnt, in statusbezogene Verzerrungseffekte wie den sogenannten Matthäus-Effekt, der vorliegt, wenn ein Gutachter/Leser sich z.B. in seiner Einschätzung davon leiten lässt, ob ein Text von einem bekannten oder von einem unbekannten Wissenschaftler stammt oder aus einer Projektgruppe an einer anerkannten Forschungseinrichtung oder an einer eher randständigen Einrichtung, und in persönliche Verzerrungseffekte, die in der Person des Gutachters oder Herausgebers liegen, zu denen auch Verzerrungseffekte aufgrund der ideologischen Orientierung eines Gutachters oder Herausgebers gehören, der z.B. bestimmte Positionen aus weltanschaulichen Gründen ablehnt und Manuskripte, die für diese Position sprechen, negativ bewertet, oder Manuskripte positiv bewertet, wenn sie zu Schlussfolgerungen kommen, die dem Gutachter oder Herausgeber weltanschaulich sympathisch sind.

Was ideologisch begründete Verzerrungseffekte betrifft, so berichten Hojat, Gonnella und Caelleigh (2003: 82) von dem Streit um Jay Belskys Untersuchungen über die möglichen negativen Wirkungen, die die Betreuung von Kindern in Tagesbetreuungseinrichtungen auf die Kinder haben kann, wie folgt:

„Another example of the issue [d.h. Verzerrung aufgrund ideologischer Überzeugungen] is Jay Belsky's study of possible negative effects of day care experiences on children that was rejected for publication on the ground that reporting such findings can generate anxiety among working mothers … When the study was finally published in a less broadly read journal … it generated outrage among proponents of day care centers that continued for a long while …".

Einer anderen Klasse von persönlichen Verzerrungseffekten widmet sich die Studie von Siegelman (1991), der fünf Gruppen von Gutachtern identifiziert hat, die im Zeitraum von November 1985 bis Mai 1990 als Gutachter für die Zeitschrift Radiology tätig waren, und zwar nach dem Kriterium, wie weit und in welche Richtung ihre Bewertungen von Manuskripten vom arithmetischen Mittel der Bewertungen aller Gutachter insgesamt gesehen abweichen; in der Beschreibung von Siegelman selbst:

„The mean ratings for referees who had been sent 10 or more manuscripts (n = 660) during the period of investigation were computed. The standard deviation of the mean ratings was calculated. On the basis of the deviation from the mean score, reviewers were classified into five categories: zealots, pushovers, mainstream, demoters, and assassins" (Seligman 1991: 637).

Die Untersuchung von Seligman hat gezeigt, dass sich die Existenz der fünf Gruppen von Gutachtern – unter Kontrolle von Zufallseffekten, die dafür gesorgt haben könnten, dass bestimmte Gutachter tatsächlich nur oder überwiegend besonders qualitätvolle oder besonders mangelhafte Manuskripte zu begutachten hatten – nachweisen lässt, d.h. dass es Gutachter gibt, die bei der Begutachtung von Manuskripten Standards haben, die deutlich höher oder deutlich niedriger sind als die Standards der Mehrheit der Gutachter. Dies verweist wieder auf den Lotterie-Charakter des „peer reviewing“:

„Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage" (Seligman 1991: 642).

Darüber hinaus hat Seligman beobachtet, dass „assassins“ und „zealots“, also besonders kritische oder besonders wohlwollende Gutachter, in allen thematischen Bereichen vorkommen, die in der Zeitschrift abgedeckt werden, aber dass sie sich nicht gleichmäßig über alle thematische Bereiche verteilen. So waren im Bereich der Nuklearmedizin und der Computeranwendungen in der Radiologie keine besonders wohlwollenden Gutachter vertreten, während in den Bereichen Ultraschall, Cardiovaskulare Interventionsmedizin und Neuroradiologie am häufigsten wohlwollende oder besonders wohlwollende Gutachter („zealots“ oder „pushovers“) vertreten waren (Seligman 1991: 640, Table 4). Das weist darauf hin, dass bestimmte Subdisziplinen oder Themenbereiche innerhalb einer Disziplin ihre eigene Anspruchs“kultur“ haben, in der Ansprüche durchschnittlich höher oder niedriger sind als in anderen Subdisziplinen oder Themenbereichen. Seligman (1991: 642) hält fest:

„Editors should be aware of reviewer variation. Editors of journals with a small corps of referees undoubtedly will recognize their assassins and zealots and will manage to deal with the disparities. For large journals with numerous reviewers, there is a danger that authors will be treated unfairly if no effort is made to record and to recognize differences in reviewer standards”.

Eine frühe Studie von Mahoney (1977) hat gezeigt, dass es Gutachter gibt, die einem Verzerrungseffekt unterliegen, den man als Bestätigungseffekt bezeichnen kann: Sie haben die Tendenz, Manuskripte positiv zu bewerten bzw. Forschungsergebnisse als zuverlässig zu bewerten, wenn sie mit den derzeit weithin akzeptierten Überzeugungen über den in Frage stehenden Sachverhalt im Einklang stehen, und Manuskripte negativ zu bewerten bzw. Forschungsergebnisse als wenig verlässlich zu bewerten, wenn sie dem angeblichen oder tatsächlichen „Konsens“ widersprechen. Dieser Befunde mag alt sein, aber keineswegs veraltet, sondern – im Gegenteil – von großer Aktualität vor dem Hintergrund der Beschwörung eines angeblich existierenden „Konsenses“ unter sogenannten Klimawissenschaftlern.

Die Ergebnisse der Studie von Peters und Ceci (1982), über die oben schon berichtet wurde, wurden von den Autoren als Beleg für den Matthäus-Effekt interpretiert, weil sie bei ihrer Wiedereinreichung der bereits in der Zeitschrift veröffentlichten Texte erfundene Autorennamen verwendeten, die angeblich an Einrichtungen beschäftigt waren, die

ebenfalls erfundene Bezeichnungen trugen und von denen deshalb die Gutachter und Herausgeber der Zeitschrift noch niemals etwas gehört haben konnten.

Eine weitere wichtige Klasse von Verzerrungseffekten hängt mit der gesamten Wissenschaftskultur zusammen, in der Arbeiten weithin als uninteressant angesehen werden, wenn es sich bei ihnen um Replikationsstudien handelt oder um Arbeiten, die die Hypothese, die in der Arbeit geprüft wird, nicht bestätigen.

Dass Letztere eine niedrigere Chance haben, veröffentlicht zu werden, als Arbeiten, die die untersuchte Hypothese bestätigen, belegen u.a. die Arbeiten von Sterling (1959), von Greenwald (1975) und von Dickersin (1990). Der Effekt tritt vermutlich besonders häufig in Fällen auf, in denen der im Manuskript berichtete Nicht-Zusammenhang mit ideologischen Überzeugungen oder sozialpolitischen Positionen konfligiert, wie dies der Fall gewesen ist bei Untersuchungen, die keinen Zusammenhang zwischen dem Kokainkonsum schwangerer Frauen und der Entwicklung des Fötus im Mutterleib feststellen konnten:

„To examine whether studies showing no adverse effects of cocaine in pregnancy have a different likelihood of being accepted for presentation by a large scientific meeting, all abstracts submitted to the Society of Pediatric Research between 1980 and 1989 were analysed. There were 58 abstracts on fetal outcome after gestational exposure to cocaine. Of the 9 negative abstracts (showing no adverse effect) only 1 (11%) was accepted, whereas 28 of the 49 positive abstracts were accepted (57%). This difference was significant. Negative studies tended to verify cocaine use more often and to have more cocaine and control cases. Of the 8 rejected negative studies and the 21 rejected positive studies, significantly more negative studies verified cocaine use, and predominantly reported cocaine use rather than use of other drugs. This bias against the null hypothesis may lead to distorted estimation of the teratogenic risk of cocaine and thus cause women to terminate their pregnancy unjustifiably” (Koren et al. 1989: 1440).

Was den Verzerrungseffekt betrifft, der auf die Ablehnung von Replikationsstudie zurückgeht, so berichten Martin und Clarke (2017: 3), dass psychologische Fachzeitschriften die Einreichung von Manuskripten, die über Replikationsstudien berichten, normalerweise nicht ermutigen. Von 1.151 psychologischen Fachzeitschriften

tun dies nur 33; das sind gerade einmal 2,9 Prozent. Tatsächlich gibt es Zeitschriften, die unmissverständlich klar machen, dass Replikationsstudien in ihnen nicht veröffentlicht werden, wie Martin und Clarke 2017: 3) berichten:

„A typical statement is that provided by the International Journal of Behavioral Development, for example: “Studies whose sole purpose is to replicate well-established developmental phenomena in different countries or (sub) cultures are not typically published in the International Journal of Behavioral Development.” This prescription is not unique to this journal”.

Makel, Plucker und Hegarty (2012) haben untersucht, wie häufig Replikationsstudien in den 100 Psychologie-Fachzeitschriften mit den höchsten „impact factors“ seit dem Jahr 1900 tatsächlich vertreten sind. Sie fassen ihre Ergebnisse wie folgt zusammen:

„This investigation revealed that roughly 1.6% of all psychology publications used the term replication in text. A more thorough analysis of 500 randomly selected articles revealed that only 68% of articles using the term replication were actual replications, resulting in an overall replication rate of 1.07%. Contrary to previous findings in other fields, this study found that the majority of replications in psychology journals reported similar findings to their original studies (i.e., they were successful replications). However, replications were significantly less likely to be successful when there was no overlap in authorship between the original and replicating articles. Moreover, despite numerous systemic biases, the rate at which replications are being published has increased in recent decades” (Makel. Plucker & Hegarty 2012: 537).

Replikationsstudien werden also sehr selten gedruckt, und der “bias” gegen Replikationsstudien tritt häufig in Kombination mit dem “bias” gegen Manuskripte auf, die die Hypothese, die in der Arbeit geprüft wird, nicht bestätigen.

Für den Bereich der Sozialwissenschaften und speziell der Management Studies haben Kerr, Tolliver und Petree bereits im Jahr 1977 eine Studie durchgeführt, in deren Rahmen 50 Items umfassende Fragebögen an 19 bekannte sozialwissenschaftliche Fachzeitschriften und Fachzeitschriften aus dem Bereich der Management Studies verschickt wurden, um die wichtigsten Gründe für die Akzeptanz oder Ablehnung von

Manuskripten zur Veröffentlichung in diesen Zeitschriften zu identifizieren. Die Autoren berichteten mit Bezug auf Replikationsstudien:

„Item 17 investigated the reaction of reviewers to replicative studies. Generally, it appears that such studies are not considered favourably by the majority of those responding” (Kerr, Tolliver & Petree 1977: 138).

Nur eine einzige der 19 untersuchten Fachzeitschriften war “… not particularly negative toward replication“ (Kerr, Tolliver & Petree 1977: 140).

Aus einer Studie von Neuliep und Crandall (1993), deren Ergebnisse in einer Fachzeitschrift gedruckt wurde, obwohl man sie als Replikationsstudie auf diejenige von Kerr, Tolliver und Petree betrachten kann, berichten die Autoren:

„80 social science journal reviewers responded to questionnaires regarding their reviewing history and attitudes toward replication studies. Results indicate that reviewers were biased against replication studies. Many reviewers regarded studies demonstrating some new effect as more worthwhile and publishable than those studies either replicating an effect or failing to replicate an effect” (Neuliep & Kandall 1993: 21).

Wenn ausgerechnet Replikationsstudien besonders häufig von einer Publikation in einer Fachzeitschrift ausgeschlossen werden, ist das ein großes Problem, denn Replikation ist

„… at the heart of any science. In all science, replication serves at least two purposes: First, to establish the reliability of previous findings and, second, to determine the generality of these findings under differing conditions. These goals, of course, are intrinsically interrelated. Each time that certain results are replicated under different conditions, this not only established generality of findings, but also increases confidence in the reliability of these findings” (Hersen & Barlow 1976: 317).

Durch Replikationsstudien können also die Zuverlässigkeit und die Generalisierbarkeit von Zusammenhängen (oder Nicht-Zusammenhängen), die in „one shot“-Studien beobachtet wurden, geprüft werden, oder anders ausgedrückt: Irrtümer oder unzulässige

Verallgemeinerungen können durch Replikationsstudien als solche festgestellt und ggf. korrigiert werden.

Was die Generalisierbarkeit von Beobachtungen in „one shot“-Studien angeht, so hat Lamal diesbezüglich festgehalten:

„We would doubtless all agree that the person who concluded that all dogs are black after seeing only one dog that happened to be black, was foolish” (Lamal 1990: 34).

Seltsamerweise haben ausgerechnet Leute, die Wissenschaftler sein wollen, anscheinend kein Problem damit, nach Beobachtung eines einzigen Hundes, der zufällig ein schwarzer Hund gewesen ist, zu akzeptieren, dass alle Hunde schwarz seien bzw. sein müssten. Für den Bereich der Marketing-Fachzeitschriften stellen Evanschitzky et al. (2007) dementsprechend und entmutigenderweise fest:

„Researchers express concern over a paucity of replications. In line with this, editorial policies of some leading marketing journals now encourage more replications. This article reports on an extension of a 1994 study to see whether these efforts have had an effect on the number of replication studies published in leading marketing journals. Results show that the replication rate has fallen to 1.2%, a decrease in the rate by half. As things now stand, practitioners should be skeptical about using the results published in marketing journals as hardly any of them have been successfully replicated, teachers should ignore the findings until they receive support via replications and researchers should put little stock in the outcomes of one-shot studies” (Evanschitzky et al. 2007: 411; Hervorhebung d.d.A.).

Würden Replikationsstudien als höchst wünschenswert gewertschätzt, regelmäßig durchgeführt und deren Ergebnisse veröffentlicht, unabhängig davon, ob sie den in Frage stehenden Zusammenhang bestätigen oder nicht bestätigen können, wäre dies ein deutliches Signal an alle Forscher, ihre Daten und ihre Verfahrensweisen so transparent wie möglich zu machen und so gründlich wie möglich zu arbeiten – eben weil sie mit einer Replikationsstudie mit Bezug auf die eigene Arbeit rechnen müssen. So betrachtet ist die mangelnde Wertschätzung von Replikationsstudien in der Wissenschaft ein Indikator für mangelnden Willen zur Transparenz und zur Überprüfung der eigenen Arbeit durch

Fachkollegen. Wer mag warum Angst vor Replikationsversuchen der eigenen Arbeit haben?!

4.3 Ignoranz gegenüber Innovationen

Wie oben schon im Zusammenhang mit der Studie von Mahoney (1977) zum Bestätigungseffekt bemerkt wurde, gibt es bei Gutachtern die Tendenz, Manuskripte positiv zu bewerten bzw. Forschungsergebnisse als zuverlässig zu bewerten, wenn sie mit dem, was gerade als „Konsens“ gilt, übereinstimmen. Auch, wenn wir nicht wissen, wie stark diese Tendenz ausgeprägt ist und ob sie z.B. in verschiedenen Fachdisziplinen unterschiedlich stark ausgeprägt ist, so lässt sich doch festhalten, dass es sie gibt und sie, wo es sie gibt, als Innovationshemmer wirkt.

Als Innovationshemmer wirkt aber nicht nur der Bestätigungseffekt als solcher, d.h. der Bestätigungseffekt als Zustimmungs- oder Plausibilitäts-bias. Manche oder vielleicht auch viele Gutachter scheinen darüber hinaus außer Stande zu sein, das Potential oder die mögliche hohe Relevanz erkennen zu können, die ein „abweichendes“ Manuskript für das Fach haben kann. Wenn das Interesse, das ein Text bei Fachkollegen findet, bzw. die Häufigkeit, mit der ein Text von Fachkollegen zitiert wird, als Merkmale dafür akzeptiert werden, dass ein Text insofern qualitätvoll ist, dass er für das Fach wichtig oder zumindest anregend ist, dann muss man festhalten, dass „peer reviewing“ auch diesbezüglich versagen kann:

Campanario (1996) hat auf der Basis einer Auswertung von 205 Kommentaren von Autoren einiger der meistzitierten Texte aller Zeiten festgestellt, dass immerhin in 22 Kommentaren bzw. knapp 11 Prozent der Fälle die Autoren davon berichten, auf Schwierigkeiten oder Widerstand mit Bezug auf die Durchführung ihrer Forschung oder die Veröffentlichung ihrer Forschungsergebnisse im entsprechenden Manuskript gestoßen zu sein. Drei der Artikel, mit denen es solche Probleme gab, haben sich nach ihrer Veröffentlichung als die meistzitierten in der Fachzeitschrift, in der sie schließlich gedruckt wurden, erwiesen.

Bei Gans und Shepherd (1994: 167, Tabelle 1), die 140 Ökonomen – inklusive aller damals lebenden Gewinner des Nobel-Preises für Ökonomie – darum gebeten haben, von ihren Erfahrungen mit der Ablehnung ihrer Manuskripte durch Fachzeitschriften zu berichten, findet man eine Aufstellung von sehr bekannten und viel zitierten Texten berühmter Ökonomen, für die allesamt eine Veröffentlichung in Fachzeitschriften zunächst abgelehnt wurde, die aber später veröffentlicht wurden. Ein großer Teil davon darf heute als ein Klassiker der Ökonomie bezeichnet werden, wie z.B. „The Market for Lemons: Quality, Uncertainty and the Market Mechanism“ von George Akerlof, „A Theory of the Allocation of Time“ von Gary S. Becker und „Increasing Returns, Monopolistic Competition, and International Trade" von Paul Krugman.

Dies alles bedeutet nicht notwendigerweise, dass Arbeiten, die besonders innovativ sind, regelmäßig auf Ablehnung bei Förderern, Herausgebern oder Gutachtern stoßen, weil sie innovativ sind, aber dass dies der Grund oder ein Grund für die Ablehnung eines Manuskriptes sein kann, zeigt der Befund aus Campanarios Studie, nach dem einige dieser später veröffentlichten, meistzitierten Manuskripte abgelehnt wurden, weil Gutachter fanden, dass sie nicht von hinreichender Wichtigkeit seien oder gängigen Vorstellungen widersprächen oder ungewöhnliche Methoden benutzten.

Auch Morton berichtet davon, dass Manuskripte auf Ablehnung stoßen, weil die relativ neuen statistischen Methoden, die die Autoren verwendet hatten, den Gutachtern unbekannt waren oder sie sie nicht vollständig verstanden haben. Morton kommt zu dem Ergebnis, dass

„… for authors to comply with these guidelines [i.e., ‚guidelines for reporting statistics in American Physiological Society journals], the initial challenge is to have a team of reviewers who are also willing to accept the unfamiliar. Indeed, the opinions of reviewers who are ill informed about relatively novel statistical methods and recommended reporting practices may have implications for the final editorial decision on the suitability of submitted manuscripts for publication” (Morton 2009: 7).

Dies verweist auf das notwendigerweise vorhandene Problem, dass gerade mit Bezug auf innovative oder methodisch vergleichsweise anspruchsvolle Studien nicht viele Fachkollegen die Offenheit und das Vorstellungsvermögen oder die Kompetenz haben, die

sie als „peers“ für diesen Bereich qualifizieren würden. Fachkollegen ohne entsprechende Kompetenz können die Unverständlichkeit eines Manuskriptes für sie vielleicht fälschlich als Ergebnis mangelnder Qualität des Manuskriptes auffassen statt als ein Ergebnis mangelnder eigener Kompetenz.

Darüber, wie Vieles der Wissenschaft dadurch verloren gegangen ist, dass Förderer, Herausgeber oder Gutachter die Originalität und Relevanz für sie ungewöhnlicher Arbeiten nicht erkannten oder diesen Arbeiten wegen ihrer Abweichung vom Gewöhnlichen verständnislos gegenüberstanden, läßt sich – wie Campanario bemerkt – nur spekulieren.

Bis hierhin sollte klar geworden sein, dass der Hinweis darauf, dass eine Veröffentlichung „peer reviewed“ ist, keinesfalls ohne Weiteres als ein Qualitätsmerkmal oder mehr: als Unbedenklichkeitsbescheinigung gelten kann. Wenn

- „peer reviewing“ einer Lotterie gleicht, wenn
- „peer reviewing“ nicht verhindern kann, dass fehlerhafte Manuskripte veröffentlicht werden,
- oder vor allem solche, die das bereits Bekannte rekapitulieren oder bestätigen, und wenn
- „peer reviewing“ Innovationen eher im Weg stehen als befördern,

sind sie dann wenigstens dazu geeignet, den größten Unsinn zu identifizieren und auszusondern, bevor er publiziert wird? Leider nein, wie der nächst Abschnitt zeigen wird.

4.4 Unfähigkeit des „peer reviewing“, Unsinn zu identifizieren und auszusondern

Die Verwendung fiktiver Manuskripte wie bei Baxt et al. (1998; s.o.) auch in anderen Fachbereichen als der Medizin hat für diese Fachbereiche gezeigt, dass es nicht gelingt, durch das „peer reviewing“ die Veröffentlichung von reinen Unsinns-Publikationen zu verhindern. Das berühmteste Beispiel hierfür ist wahrscheinlich der Text von Alan Sokal mit dem Titel „Transgressing the Boundaries: Towards a Transformative Hermeneutics of Quantum Gravity“, der im Jahr 1996 in der Zeitschrift Social Text veröffentlicht wurde,

einer Zeitschrift, die den sogenannten „Cultural Studies“ verschrieben ist, die ihrerseits vom Postmodernismus geprägt sind, und vorzugsweise Texte über Geschlecht, Hautfarbe und Umwelt, also die üblichen „Verdächtigen“, druckt.

Um einen Eindruck vom Inhalt dieses Textes zu vermitteln, seien zwei mehr oder weniger typische Absätze aus dem Text hier zitiert:

„Here my aim is to carry these deep analyses one step further, by taking account of recent developments in quantum gravity: the emerging branch of physics in which Heisenberg's quantum mechanics and Einstein's general relativity are at once synthesized and superseded. In quantum gravity, as we shall see, the space-time manifold ceases to exist as an objective physical reality; geometry becomes relational and contextual; and the foundational conceptual categories of prior science – among them, existence itself – become problematized and relativized. This conceptual revolution, I will argue, has profound implications for the content of a future postmodern and liberatory science” (Sokal 1996: 218).

Und

“As Irigaray anticipated, an important question in all of these theories is: can the boundary be transgressed (crossed), and if so, what happens then? Technically, this is known as the problem of boundary conditions (b.c.). At a purely mathematical level, the most salient aspect of boundary conditions is the great diversity of possibilities: for example, "free b.c." (no obstacle to crossing), "reflecting b.c." (specular reflection as in a mirror), "periodic b.c." (re-entrance in another part of the manifold), and "antiperiodic b.c." (re-entrance with 180-degree twist). The question posed by physicists is: of all these conceivable boundary conditions, which ones actually occur in the representation of quantum gravity? Or perhaps, do all of them occur simultaneously and on an equal footing, as suggested by the complementarity principle?” (Sokal 1996: 226).

Nachdem der Text in der Zeitschrift veröffentlicht worden war, erklärte Sokal, dass er den Text als eine Parodie auf die Wissenschaft imitierende Sprache des Postmodernismus verfasst habe und er als solche nichts anderes sei als eine Aneinanderreihung weitgehend sinnloser und teilweise absurder Sätze. Wie leicht vorhersehbar war, wurde Sokal

daraufhin von entsprechend geneigter ideologischer Seite heftig kritisiert, teilweise beschimpft, und es wurde ihm vorgeworfen, unaufrichtig zu sein und mit seinem Text „kritische“ oder Grenzen überschreitende „Forschung“ zu diskreditieren, er damit der Sache der „Rechten“ zuarbeite. Kritik an Wissenschaftsparodie wurde also sehr schnell politisiert bzw. ideologisiert, ganz so, wie „Linke“ und Postmoderne, das auch heute zu tun pflegen. Demgegenüber trat die Tatsache, dass Sokals Text auf einen grundlegenden Mißstand im Wissenschaftsbetrieb, nämlich dessen Unfähigkeit, wissenschaftliche Standards einzuhalten und als solche zu verteidigen gegen Angriffe von Personen, die Wissenschaft lediglich, um politischer Zielsetzungen willen zu imitieren versuchen, in den Hintergrund.

In dem Buch mit dem Titel „Fashionable Nonsense: Postmodern Intellectuals‘ Abuse of Science“, das Sokal gemeinsam mit dem belgischen Physiker Jean Bricmont im Jahr 1998 veröffentlicht hat und das auch in deutschsprachiger Übersetzung unter dem nicht ganz treffenden Titel „Eleganter Unsinn“ gedruckt wurde, unterziehen die Autoren das Verhältnis zwischen Postmodernismus und Wissenschaft einer detaillierteren Betrachtung, zitieren in diesem Zusammenhang Prosa postmoderner Autoren ausführlich und erläutern, wo und warum das, was dort geschrieben steht, Unsinn ist. Sokal und Bricmont demonstrieren, dass die von postmodernen Autoren habituell vorgebrachte Kritik an der Vernunft und der wissenschaftlichen Methode verfehlt ist und dass und wie postmoderne Autoren wissenschaftliche Konzepte häufig missbrauchen oder einfach missverstehen, was Sokal bereits durch seinen Schwindel mit dem absichtlich in weiten Teilen unsinnigen, in anderen Teilen banalen, Text, der in Social Text gedruckt wurde, illustriert hatte. Die Tatsache, dass Sokals Unsinns-Text in einer Zeitschrift gedruckt wurde, die ein angeblich wissenschaftliches Fach, die sogenannten Cultural Studies, repräsentieren, hat zumindest gezeigt, dass das „peer reviewing“ als Verfahren immer dann versagt und versagen muss, wenn ein Text kein wissenschaftlicher ist, d.h. in ihm das allen Wissenschaften mehr oder weniger gemeinsame Arbeitsprogramm, das auf einer mehr oder weniger geteilten Epistemologie basiert, aufgekündigt wird – aus Inkompetenz oder bewusst, im Zuge des Versuchs, eine Alternative zu Wissenschaft zu schaffen, die dennoch Wissenschaft sein will.

Die Prüfung der Qualität von wissenschaftlichen Zeitschriften und Tagungen bzw. Tagungsleitungen, Herausgebern oder Gutachten durch die Einreichung von „Schein“-

Manuskripten scheint inzwischen fast ein Standard-Instrument der diesbezüglichen Evaluationsforschung geworden zu sein. Das Instrument ist seit den 1990er-Jahren aber stark modernisiert und automatisiert worden. So hat eine Gruppe von Absolventen von CASIL, dem Computer Science and Artificial Intelligence Laboratory am MIT, ein Computerprogramm namens SCIgen entwickelt, das zufällig sinnlose Texte samt Abbildungen und Zitationen aus dem Bereich der Computerwissenschaft generiert. Was der Postmodernismus für die Sozialwissenschaften ist, ist SCIgen für die Computerwissenschaft insofern auch SCIgen mit Schlagworten operiert wie z.B. „Byzantine fault tolerance“ und „distributed hash tables“.

Um ihr Programm zu prüfen, haben die Programm-Entwickler ein Manuskript mit dem Titel „Rooter: A Methodology for the Typical Unification of Access Points and Redundancy“ erstellen lassen, und das Manuskript wurde im Jahr 2005 als Vortragsmanuskript für WMSCI akzeptiert, die World Multiconference on Systemics, Cybernetics and Informatics, ohne einer Prüfung durch Gutachter unterzogen worden zu sein. Die Programmentwickler machten ihren Schwindel bekannt, zogen das Unsinns-Papier zurück, und die Konferenz verlor ihre Sponsoren.

SCIgen wird seitdem sowohl zur Identifizierung von Tagungsveranstaltern oder Zeitschriften mit niedrigen (oder gar keinen) wissenschaftlichen Standards benutzt, aber auch zur Generierung von Manuskripten, die Konferenzveranstaltern oder Fachzeitschriften untergeschoben werden (sollen):

„… in 2013 IEEE [Institute of Electrical and Electronics Engineers] and Springer Publishing removed more than 120 papers from their sites after a French researcher’s analysis determined that they were generated via SCIgen“ (Conner-Simons 2015),

und gleichzeitig

“[t]he creators [of SCIgen] continue to get regular emails from computer science students proudly linking to papers they’ve snuck into conferences, as well as notes from researchers urging them to make versions for other disciplines” (Conner-Simons 2015).

Bohannon (2013), ein ehemaliger (auch) für Science tätiger Journalist, hat mit Hilfe eines Schwindel-Manuskriptes, in das offensichtliche Fehler eingebaut waren und das sträfliche Verletzungen wissenschaftlicher Standards enthielt, bzw. mit Hilfe leicht verschiedener Varianten desselben Manuskriptes geprüft, wie es um das „peer reviewing“ bei open access-Zeitschriften bestellt ist, die wissenschaftliche Arbeiten digital online und ohne Zugangsbeschränkungen für Leser verteilen. Er hat festgestellt:

„Of the 255 papers that underwent the entire editing process to acceptance or rejection, about 60% of the final decisions occurred with no sign of peer review. For rejections, that’s good news: It means that the journal’s quality control was high enough that the editor examined the paper and declined it rather than send it out for review. But for acceptances, it likely means that the paper was rubber-stamped without being read by anyone. Of the 106 journals that discernibly performed any review, 70% ultimately accepted the paper. Most reviews focused exclusively on the paper’s layout, formatting, and language … Only 36 of the 204 submissions generated review comments recognizing any of the paper’s scientific problems. And 16 of those papers were accepted by the editors despite the damming reviews” (Bohannon 2013: 64).

Ein weiteres Ergebnis, das Bohannon mit seinem Schwindel-Manuskript erzielt hat, ist, dass es nicht möglich ist, sein Vertrauen in Publikationen in einer Zeitschrift durch die Reputation der Zeitschrift oder hier: die Reputation der Art von Zeitschrift begründen zu wollen. Z.B.

„[s]ome open-access journals that have been criticized for poor quality control provided the most rigorous peer review of all. For example, the flagship journal of the Public Library of Science, PLOS ONE, was the only journal that called attention to the paper’s potential ethical problems, such as its lack of documentation about the treatment of animals used to generate cells for the experiment. The journal meticulously checked with the fictional authors that this and other prerequisites of a proper scientific study were met before sending it out for review. PLOS ONE rejected the paper 2 weeks later on the basis of its scientific quality” (Bohannon 2013: 61).

Ein Fall aus neuerer Zeit, nämlich den Jahren 2017 und 2018, ist der Fall von „reflexiver Ethnographie“ („reflexive ethnography“), die von Helen Pluckrose, James A. Lindsay und

Peter Boghossian unternommen wurde, um die Funktionsweise der „grievance studies“ zu testen, „….which is corrupting academic research“

Unter der Bezeichnung „grievance studies“ fassen die drei

“… fields of scholarship loosely known as “cultural studies” or “identity studies” (for example, gender studies) or “critical theory” … “

zusammen,

“…because of their common goal of problematizing aspects of culture in minute detail in order to attempt diagnoses of power imbalances and oppression rooted in identity”

Die drei Autoren verfassten 20 Manuskripte, die alle auf Absurditäten oder zutiefst Unethischem beruhten, darunter auf Hitlers „Mein Kampf“, aus dem die drei Passagen entnommen und in eine feministische Kampfschrift transformiert hatten. Dieser Text wurde von den drei Autoren mit den Namen fiktiver Autorinnen versehen, die angeblich an einer Einrichtungen namens Feminist Activist Collective for Truth (FACT) tätig sein sollten, die aber ebenso wenig wie die angegeben Autorinnen existierte.

Nicht nur dieser Text, der den Titel „Our Struggle Is My Struggle: Solidarity Feminism as an Intersectional Reply to Neoliberal and Choice Feminism“ trug, wurde von einer Zeitschrift aus dem Bereichen der „grievance studies“ zur Veröffentlichung akzeptiert und gedruckt, nämlich von Affilia: Journal of Women and Social Work. Von den 20 Test- Manuskripten, die die Autoren verfassten, wurden sieben zur Veröffentlichung durch Fachzeitschriften akzeptiert, sieben weitere waren noch im Begutachtungsprozess, als die drei beschlossen, ihre Schwindel bekannt zu machen, und nur sechs waren abgelehnt worden. Die Autoren berichten

„1 paper (the one about rape culture in dog parks) gained special recognition for excellence from its journal, Gender, Place, and Culture, a highly ranked journal that leads the field of feminist geography. The journal honored it as one of twelve leading pieces in feminist geography as a part of the journal’s 25th anniversary celebration”

Was am Schwindel von Pluckrose, Lindsay und Boghossian bedenklich ist, ist nicht nur die Tatsache, dass die Gutachter oder Herausgeber von Zeitschriften aus dem Bereich der „grievance studies“ sich nicht nur nicht an erheblichen methodischen Mängeln, völlig unplausiblen Daten und „non sequiturs“ störten (oder diese Mängel gar nicht erkannten), sondern auch und vielleicht besonders, dass schiere Ausmaß der Perversionen, die Gutachtern und Herausgebern von Zeitschriften aus dem Bereich der „grievance studies“ akzeptabel erschienen:

„Many papers advocated highly dubious ethics including training men like dogs (“Dog Park”), punishing white male college students for historical slavery by asking them to sit in silence on the floor in chains during class and to be expected to learn from the discomfort (“Progressive Stack”), celebrating morbid obesity as a healthy life-choice (“Fat Bodybuilding”), treating privately conducted masturbation as a form of sexual violence against women (“Masturbation”), and programming super-intelligent AI with irrational and ideological nonsense before letting it rule the world (“Feminist AI”). There was also considerable silliness including claiming to have tactfully inspected the genitals of slightly fewer than 10,000 dogs whilst interrogating owners as to their sexuality (“Dog Park”), becoming seemingly mystified about why heterosexual men are attracted to women (“Hooters”), insisting there is something to be learned about feminism by having four guys watch thousands of hours of hardcore pornography over the course of a year while repeatedly taking the Gender and Science Implicit Associations Test (“Porn”), expressing confusion over why people are more concerned about the genitalia others have when considering having sex with them (“CisNorm”), and recommending men to anally self-penetrate in order to become less transphobic, more feminist, and more concerned about the horrors of rape culture (“Dildos”). None of this, except that Helen Wilson recorded one “dog rape per hour” at urban dog parks in Portland, Oregon, raised so much as a single reviewer eyebrow, so far as their reports show”

Damit dürfte ein vorläufiger Tiefstpunkt mit Bezug auf Gutachter- und Herausgeberleistungen erreicht sein.

4.5 Vom Mangel zum Betrug: Bewusste Manipulationen des „peer reviewing”

Was bislang dargestellt wurde, fällt sehr weitgehend in den Bereich dessen, was am „peer reviewing“ mangelhaft ist und die Herausgeber- oder Gutachterseite des Prozesses betrifft. Bewusste Täuschung, bewusster Betrug durch Herausgeber oder Gutachter ist im Rahmen des üblichen „peer reviewing“ selbstverständlich möglich, aber bislang wurde es vor allem als Problem angesehen, dass Herausgeber oder Gutachter nicht im Stande sind – oder aufgrund praktischer Probleme, z.B. mit dem Zugang zu Daten, die einem Manuskript zugrunde liegen, gar nicht im Stande sein können –, Täuschung oder Betrug auf Seiten der Autoren zu identifizieren, der in der Erfindung oder Fälschung von Daten bestehen kann, im Plagiieren, in der Fälschung von Autorenschaft u.a.m. bestehen kann.

Organisationen wie VroniPlag in Deutschland und Computerprogramme zur Prüfung von Manuskripten daraufhin, ob es sich bei ihnen vollständig oder teilweise um Plagiate handelt, helfen diesbezüglich sicherlich weiter, aber ihr Nutzen scheint vor allem in einer Abschreckungswirkung zu bestehen, denn nicht alle plagiierten Sätze oder Absätze in Manuskripten werden von solchen Programmen korrekt identifiziert (s. z.B. Stapleton 2012 mit Bezug auf das weit verbreitete Programm Turnitin).

Die kriminelle Energie von Autoren erschöpft sich auch keineswegs in dem, was einige Autoren neuerdings als „traditional fraud and misconduct“ (Biagioli & Lippman 2020: 2) wie es z.B. das Plagiieren ist, bezeichnen, denn inzwischen gibt es neue Formen des Betrugs oder der Manipulation, die teilweise direkt in einen speziellen „peer review“-Prozess eingreifen oder stattfinden, nachdem ein Manuskript zur Publikation akzeptiert worden ist.

So kommt es z.B. vor, dass Autoren, wenn sie von der Identität der Gutachter, die ihr Manuskript begutachten sollen, erfahren, oder wenn sie selbst mögliche Gutachter vorgeschlagen haben, die von den Herausgebern als Gutachter in diesem Fall akzeptiert wurden, falsche e-mail-Adressen einrichten und als falsche Gutachter ihre eigenen Papiere begutachten – selbstverständlich positiv. Es liegt in der Natur der Sache, dass man hiervon nur dann erfährt, wenn der Versuch fehlschlägt, und über die tatsächliche Häufigkeit, mit der gefälschte Gutachten die Grundlage von Entscheidungen über die Veröffentlichung eines Manuskriptes in einer Fachzeitschrift bieten, lässt sich nur spekulieren.

Mit Hilfe der der Datenbank, die von Adam Marcus und Ivan Oransky, den Gründern und Betreibern des blogs Retraction Watch, seit 2010 geführt wird, sowie Datenbanken von PubMed und Google Scholar konnten Qi, Deng und Guo (2017) in ihrer Studie 250 Texte identifizieren, die aufgrund gefälschter Gutachten zurückgezogen werden mussten. Dass es sich bei diesen Texten vor allem um Texte aus den „life sciences“ handelt, hat mit der Ausrichtung von Retraction Watch zu tun, das sich (bislang) stark auf die „life sciences“ konzentriert, was Marcus und Oransky wie folgt begründen:

“… 1) we’re both medical reporters in our day jobs, so our sources and knowledge base are both deeper in the life sciences and 2) there are more papers published in the life sciences than in other areas”.

Die von Qi, Deng und Guo identifizierten Texte waren in 48 verschiedenen Fachzeitschriften von fünf verschiedenen Verlegern veröffentlicht worden. Die meisten dieser Texte, insgesamt 57 Prozent der 250 identifizierten Texte, waren in Fachzeitschriften gedruckt worden, die von SAGE (mit 77 oder 31 Prozent der zurückgezogenen Texten in ihren Fachzeitschriften) oder von Springer (mit 66 oder 26 Prozent der zurückgezogenen Texten in ihren Fachzeitschriften) verlegt wurden. Zehn der Fachzeitschriften, in denen Texte, für die Gutachten gefälscht wurden, veröffentlicht wurden, hatten mehr als fünf solcher Texte veröffentlicht (Qi, Deng & Guo 2017: 499). Drei Viertel der 250 zurückgezogenen Texte waren von chinesischen Autoren verfasst worden (Qi, Deng & Guo 2017: 500). Das Jahr mit der größten Anzahl von aufgrund gefälschter Gutachten zurückgezogener Texte war das Jahr 2014, auf das 102 oder 40,8 Prozent der Texte entfielen.

Es ist unklar, was genau diese Verteilungen bedeuten, d.h. ob sie z.B. bedeuten, dass bestimmte Verleger oder bestimmte Fachzeitschriften mit Bezug auf die Begutachtung von Manuskripten besonders einfach zu täuschen sind oder es auf andere Faktoren oder schlichten Zufall zurückzuführen ist, dass Texte, für die gefälschte Gutachten erstellt wurden, bei diesen Verlegern oder Fachzeitschriften besonders häufig aufgeflogen sind. Festhalten lässt sich aber:

“In conclusion, there is a disproportionate distribution of retracted papers due to faked peer reviews among different journals and countries. Journal editors should greatly improve the peer review mechanism, ….” (Qi, Deng & Guo 2017: 502).

Es kommt ebenfalls vor, dass die Datenbanken von Zeitschrift gehackt werden, um dafür zu sorgen, dass das eigene Manuskript veröffentlicht wird, z.B. indem falsche, positive Gutachten eingestellt werden, oder um den eigenen Namen in der Autorenzeile für ein Manuskript, das zum Druck vorbereitet wird, ein- oder anzufügen (Biagioli & Lippman 2020: 2). Der Eintrag als Mitautor kann auch zum Kauf bzw. angeboten bzw. erkauft werden.

Es gibt außerdem Manipulationen, die das „peer reviewing“ nicht direkt betreffen, aber Herausgeber und ggf. Gutachter, auch, wenn sie insgesamt gesehen kompetent sind und aufrichtig agieren, zu täuschen geeignet sind, z.B. dann, wenn Autoren sogenannte Zitierringe bilden, d.h. vereinbaren, sich gegenseitig zu zitieren, um damit die Zahl der eigenen Zitationen künstlich in die Höhe zu treiben. Dies kann über statusbezogene Verzerrungseffekte auf Herausgeber und Gutachter wirken, wenn sie Entscheidungen über Manuskripte solcher Autoren treffen sollen.

Darüber hinaus können Herausgeber oder Gutachter selbst Zitationen der eigenen Arbeiten oder von Arbeiten von Autoren, mit denen Herausgeber oder Gutachter Zitierringe unterhalten, den Autoren, die ein Manuskript eingereicht haben, abpressen oder zumindest empfehlen, wobei Autoren, die ihr Manuskript veröffentlicht sehen wollen, gut daran tun, der Empfehlung nachzukommen (Teixeira da Silva 2017).

5. Schlussfolgerung: „Peer reviewed“ ist kein Qualitätssiegel

Die empirischen Befunde, die in Abschnitt 3 berichtet wurden (und viele andere, die in den vorliegenden Text nicht aufgenommen wurden, aber bei entsprechender Recherche leicht gefunden werden können), sollten hinreichend belegt haben, dass es keinen Grund dafür gibt, die schlichte Tatsache, dass über einen Text gesagt wird, er sei „peer reviewed“ als Qualitätssiegel aufzufassen bzw. aus dieser schlichten Tatsache zu schließen, dass der Text ein Text von hoher Qualität sei, sei es im Hinblick auf faktische Richtigkeit, auf wissenschaftliche Relevanz oder was auch immer.

Das Label „peer reviewed“ wird auf einen Text manchmal nur deshalb übertragen, weil die Zeitschrift, in der er veröffentlicht wurde, angibt, dass sie Manuskripte zur Veröffentlichung aufgrund eines „peer reviewings“ auswähle. Das mag in bestimmten Fällen oder regelmäßig, bei bestimmten Zeitschriften vielleicht sogar immer, der Fall sein, vielleicht aber auch nicht. Wenn es der Fall ist, bleibt in aller Regel unbekannt, wer das Manuskript begutachtet hat – ein Herausgeber, mehrere Herausgeber, mindestens ein Herausgeber und ein, zwei, drei oder mehr Gutachter –, wie genau Gutachter ggf. ausgewählt wurden, wie genau das „peer reviewing“ verlaufen ist, ob z.B. bestimmte Begutachtungskriterien vorgegeben wurden oder nicht, wie Herausgeber über die Veröffentlichung von Manuskripten entscheiden, besonders in dem Fall, in dem Gutachter oder Gutachter und Herausgeber zu unterschiedlichen Einschätzungen des Manuskriptes kommen.

Was wir aus empirischen Studien wissen, ist, dass „peer reviewing“ aufgrund sehr starker Variation in der Einschätzung ein und derselben Manuskripte durch verschiedene Gutachter einer Lotterie gleicht, dass „peer reviewing“ nicht verhindern kann, dass fehlerhafte Manuskripte veröffentlicht werden oder vor allem solche, die das bereits Bekannte rekapitulieren oder bestätigen, und dass „peer reviewing“ Innovationen eher im Weg steht als befördert. Wir wissen darüber hinaus, dass es „peer reviewing“ nicht gelingt, Unsinn als solchen zu identifizieren und auszusondern, bevor er publiziert wird, und dass „peer reviewing“ direkt oder indirekt, d.h. in Zusammenhängen, die dem eigentlichen Gutachterprozess vor- oder nachgelagert sind, manipuliert werden kann und wird.

Auf der Basis der vorliegenden empirischen Studien muss man deshalb zu dem Urteil kommen, zu dem bereits Cowley (2015) gekommen ist:

“Peer-review is neither reliable, fair, nor a valid basis for predicting ‘impact’: as quality control, peer-review is not fit for purpose” (Cowley 2015: 1).

Es sei hier nur am Rande ergänzend bemerkt, dass es selbst dann, wenn „peer reviewed“ tatsächlich anzeigen würde, dass ein Text von hoher Qualität sein müsse, der Text deshalb keine Überlegenheit per se gegenüber Texten, die nicht „peer reviewed“ sind, für sich beanspruchen könnte. Warum nicht? Weil es ein Fehlschluss der Verneinung des antecedens wäre, wenn man meinen würde:

> Wenn ein Text „peer reviewed“ ist, muss er von hoher Qualität sein.
> Dieser Text ist nicht „peer reviewed“.
> DAHER: Dieser Text ist nicht von hoher Qualität

(vgl. hierzu z.B. Salmon 1983: 58-59).

Es gibt durchaus Texte von hoher Qualität (und von höherer Qualität als Texte, die „peer reviewed“ sind), die nicht „peer reviewed“ sind (und ich hoffe, dass der vorliegende Text zu diesen Texten zu zählen sein wird). Es mag auch Texte geben, die von hoher Qualität sind, aber nach erfolgter Publikation zurückgezogen wurden, weil für sie vielleicht unnötigerweise ein gefälschtes positives Gutachten vorgelegt wurde. Nichts spricht dagegen, dass es in der Realität nicht alle möglichen Konstellationen von Textqualität, Gutachterleistung, und Manipulationsstrategien gibt, die zur Veröffentlichung qualitätvoller Texte führen, zur Veröffentlichung von Texten niedriger Qualität, zur Nicht-Veröffentlichung qualitätvoller Texte oder zur Nicht-Veröffentlichung von Texten niedriger Qualität. Wenn ein Text, der „peer reviewed“ ist, inhaltlich einem Text widerspricht, der nicht „peer reviewed“ ist (oder umgekehrt), was bedeutet das dann also für die Inhalte in den beiden Texten? Nichts.

In jedem Fall ist der Leser bei seiner Lektüre eines Textes, sei er in einer Fachzeitschrift oder sonstwo veröffentlicht oder nicht, auf den Text selbst zurückgeworfen. Er muss also sozusagen selbst Gutachter sein. In vielen Fällen ist man mit der Prüfung von Texten insofern überfordert als man nicht beurteilen kann, was z.B. die für die Daten angemessenen Verfahrensweisen sind, ob im Text die für das Thema relevanten

Vorarbeiten berücksichtigt wurden u.v.m. Das bedeutet aber nicht, dass man den Text nicht in den Teilen beurteilen kann oder darf, die einem zugänglich sind, also z.B. mit Bezug auf die Stringenz seiner Argumentation, die Abwesenheit von Fehlschlüssen oder unangemessenen Generalisierungen. D.h. die guten alten Mittel der Logik und des kritischen Denkens tun heute und hier wie früher und überall ihren Dienst, wenn man sie nur einzusetzen bereit ist und sie kompetent einsetzen kann.

Wie die in Abschnitt 3 genannten empirische Studien gezeigt haben, erreichen Gutachter oft keine größere Übereinstimmung untereinander mit Bezug auf ein und dasselbe Manuskript als man es per Zufall erwarten würde. Es gibt daher keinen guten Grund, sich bei der Beurteilung von Arbeiten von Gutachtern oder Herausgebern mittels des Labels „peer reviewed“ leiten zu lassen. Und angesichts der Lotterie, die „peer reviewing“ ist, ist es auch nicht notwendig, mit der eigenen Beurteilung von Texten aufgrund dessen, was man mit den Mitteln der Logik und des kritischen Denkens und, sofern vorhanden, mit dem, was man an statistischem oder methodischem Wissen hat, allzu zurückhaltend zu sein. Selbst dann, wenn die eigene Beurteilung eines Textes am Ende dieselbe Qualität haben sollte wie eine zufällige Beurteilung, so befindet man sich damit doch in der Gesellschaft von Gutachtern, die immerhin in der Regel Fachvertreter sind.

Wer heute als Wissenschaftler tätig ist, wer selbst hinreichend Erfahrung mit „peer reviewing“, entweder als Autor oder als Herausgeber oder Gutachter, hat, wer es mit Wissenschaft halbwegs ernst meint, dem können die Mängel des „peer reviewing“ schwerlich entgangen sein, und dies wirft die Frage auf, wer es ist, der als Wissenschaftler auftritt und gleichzeitig – bzw. dennoch – so tut, als könne der Hinweis darauf, dass ein Text „peer reviewed“ sei, als ein Qualitätssiegel für den entsprechenden Text aufgefasst werden. M.E. handelt es sich bei diesen Menschen vor allem um solche, die tatsächlich keine Wissenschaftler sind, sondern „professionals“ anderer Art, die ein Interesse an bestimmten Ergebnissen haben, weil sie (auch weiterhin) ein Auskommen in der Nische, in der sie tätig sind, ermöglichen. Ein Anstellungsverhältnis an einer eigentlich wissenschaftlichen Einrichtung oder in einem eigens geschaffenen Projekt an irgendeinem An-Institut einer eigentlich wissenschaftlichen Einrichtung kann dazu genutzt werden, eine bestimmte Nische einzurichten, zu legitimieren oder ihre Bedeutung zu erhöhen; man denke nur an die politisch gewollte und gesteuerte Etablierung von „Gender Studies“ an Hochschulen oder die Inszenierung des Klima-Schwindels durch die Fälschung oder

Unterschlagung von empirischen Daten (wie im berühmt gewordenen „climategate“) durch angeblich wissenschaftliches Personal, das viel Wert darauf legt, als Wissenschaftler gelten zu können, Texte zu veröffentlichen, die „peer reviewed“ sind oder z.B. – wie Michael E. Mann – einen Nobelpreis erhalten zu haben, wenn dies in der Realität eine zumindest sehr fragwürdige Angelegenheit ist.

De Vries hat das schon vor rund 20 Jahren ähnlich gesehen und formuliert:

“Peer reviewed articles are the outer shell of the learned world. They apotheosize the power elites and are instrumental in the distribution of funds. To advance the insider’s knowledge they are less essential. Science as the expression of human curiosity and industrial impetus will not be doomed if peer review dissipates” (De Vries 2001: 239).

Auflösungserscheinungen zeigt das “peer reviewing” in seiner traditionellen Form, aber in einer modernisierten und demokratisierten Form hat es in der vergangenen Jahren geradezu eine Blüte erlebt, nämlich in Form von blogs oder Organisationen, die als „watchdogs“ fungieren wie z.B. Retraction Watch, oder die es – wie z.B. PubPeer – Wissenschaftlern ermöglichen, ihre Manuskripte vor oder nach Veröffentlichung zu verteilen oder sie von „peers“ kommentieren zu lassen oder selbst die Manuskripte von „peers“ zu kommentieren, wobei sich die „peers“ selbst rekrutieren: wer meint, sich kompetent zu einem Manuskript oder einem Inhalt im Manuskript äußern zu können und etwas hinreichend Nennenswertes dazu zu sagen zu haben, tut es.

„This new generation of watchdogs is successfully making up for their lack of resources by mobilizing hundreds of scientists – some named, but mostly anonymous – who are willing to read texts, evaluate images, run through statistical analyses for a publication’s data, and share their findings and views on websites, blogs, wikis, and social media … And though they lack legal authority, these new watchdogs can be very effective through their ability to maximize the visibility of these issues, which may force the authorities to intervene … Their modus operandi is that of traditional peer review but, through the adoption of a crowd-sourcing model, it operates on a scale and is able to draw expertise from a population that is an order of magnitude larger than that of traditional peer review as practiced by journals” (Biagioli & Lippman 2020: 17).

Aufgrund der großen Anzahl von selbstrekrutierten „Gutachtern” ist es schwierig, in diesen Medien persönliche Vorlieben oder Abneigungen auszuleben, muss man doch mit Antworten auf die eigenen Kommentare rechnen, mit der Aufforderung, seine Einschätzungen zu begründen, und mit Entgegnungen und Einwänden gegen die eigenen Begründungen nicht nur durch einen oder wenige Gesprächspartner, sondern möglicherweise durch eine Vielzahl von Gesprächspartnern sehr unterschiedlicher Ausrichtung und Expertise. In einem solchen Diskussionsforum tut man gut daran, selbst sein kritischster und aufmerksamster „Gutachter“ zu sein!

Diese neuen Formen des “peer reviewing” markieren deshalb nach Biagioli und Lippman (2020: 17) eine Transformation “… from top-down to bottom-up knowledge production“ und reihen sich damit ein in die Demokratisierungsprozesse, die wir derzeit – trotz teilweise heftigen Widerstands auf Seiten derer, die sich nicht zutrauen, in einer solchem Diskussionsumgebung bestehen zu können, – in verschiedenen gesellschaftlichen Bereichen, aber insbesondere in der Medienlandschaft, erleben.

6. Literatur:

Bailar, John C. & Patterson, Kay, 1985: Journal Peer Review: The need for a Research Agenda. New England Journal of Medicine 312 (March 7): 654–657.

Baxt, William G., Waeckerle, Joseph F., Berlin, Jesse A. & Callaham, Michael L., 1998: Who Reviews the Reviewers? Feasibility of Using a Fictitious Manuscript to Evaluate Peer Reviewer Performance. Annals of Emergency Medicine 32(3): 289-403.

Biagioli, Mario & Lippman, Alexandra, 2020: Introduction: Metrics and the New Ecologies of Academic Misconduct, S. 1-23 in: Biagioli, & Lippman, (Hrsg.): Gaming the Metrics: Misconduct and Manipulation in Academic Research. Cambridge (Mass.): The MIT Press.

Bohannon, John, 2013: Who’s Afraid of Peer Review? Science 342(6154): 60-65.

Campanario, Juan Miguel, 1996: Have Referees Rejected Some of the Most-cited Articles of All Times? Journal of the American Society for Information Science 47(4): 302-310.

Chen, Chaomei, Ibekwe-SanJuan, Fidelia & Hou, Jianhua, 2010: The Structure and Dynamics of Cocitation Clusters: A Multiple-perspective Cocitation Analysis. Journal of the American Society for Information Science and Technology 61(7): 1386-1409.

Conner-Simons, Adam, 2015: How Three MIT Students Fooled the World of Scientific Journals. MIT News, April 14, 2015.
http://news.mit.edu/2015/how-three-mit-students-fooled-scientific-journals-0414

Cowley, Stephen J., 2015: How Peer-review Constrains Cognition: on the Frontline in the Knowledge Sector. Frontiers in Psychology 6, Article 1706.
doi: 10.3389/fpsyg.2015.01706.

De Vries, Jaap, 2001: Peer Review: The Holy Cow of Science, S. 231-244 in: Fredriksson, Einar H. (Hrsg.): A Century of Science Publishing: A Collection of Essays. Amsterdam: IOS Press.

Dickersin, Kay, 1990: The Existence of Publication Bias and Risk Factors for its Occurrence. Journal of the American Medical Association 263(10): 1385–1389.

Evanschitzky, Heiner, Baumgarth, Carsten, Hubbard, Raymond & Armstrong, J. Scott, 2007: Replication Research’s Disturbing Trend. Journal of Business Research 60(4): 411-415.

Gans, Joshua S. & Shepherd, George B., 1994: How Are the Mighty Fallen: Rejected Classic Articles by Leading Economists. The Journal of Economic Perspectives 8(1): 165-179.

Godlee, Fiona., Gale, Catharine R. & Martyn, Christopher N., 1998: Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports: A Randomized Controlled Trial. Journal of the American Medical Association (JAMA) 280(3): 237-240.

Greenwald, Anthony G. 1975: Consequences of Prejudice Against the Null Hypothesis. Psychological Bulletin 82(1): 1–20.

Hargens, Lowell L., 1988: Scholarly Consensus and Journal Rejection Rates. American Sociological Review 53(1): 139-151.

Hersen, Michel & Barlow, David H., 1976: Single-case Experimental Designs: Strategies for Studying Behavior Change. New York (NY): Pergamon Press.

Hojat, Mohammadreza, Gonnella, Joseph S. & Caelleigh, Addeane S., 2003: Impartial Judgment by the “Gatekeepers” of Science: Fallibility and Accountability in the Peer Review Process. Advances in Health Sciences Education 8(1): 75-96.

Horrobin, David F., 1982: Peer Review: A Philosophically Faulty Concept Which Is Proving Disastrous for Science. The Behavioral and Brain Sciences 5(2): 217-218.

Isenberg, Sherwin. J., Sanchez, Elizabeth & Zafran, Karyn. Cook., 2009; The Effect of Masking Manuscripts for the Peer-review Process of an Ophthalmic Journal. British Journal of Ophthalmology 93(7): 881-884.

Justice Amy C., Berlin Jesse A., Fletcher, Suzanne W., Fletcher Robert H. & Goodman Steven N., 1994: Do Readers and Peer Reviewers Agree on Manuscript Quality? Journal of the American Medical Association (JAMA) 272(2): 117-119.

Justice, Amy C., Cho, Mildred K., Winker, Margaret A., Berlin, Jesse A. & Rennie, Drummond, 1998: Does Masking Author Identity Improve Peer Review Quality? A

Randomized Controlled Trial. Journal of the American Medical Association (JAMA) 280(3): 240-242.

Koren, Gideon, Shear, Heather, Graham, Karen & Einarson, Tom, 1989: Bias Against the Null Hypothesis: The Reproductive Hazards of Cocaine. The Lancet 334(8677): 1440-1442.

Kravitz, Richard L., Franks, Peter, Feldman, Mitchell D., Gerrity, Martha et al., 2010: Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care? PLoS Online 5(4): e10072. https://doi.org/10.1371/journal.pone.0010072

Lamal, Peter A., 1990: On the Importance of Replication. Journal of Social Behavior and Personality 5(4): 31-35.

Mahoney, Michael J., 1977: Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. Cognitive Therapy and Research1(2): 161-175.

Makel, Matthew C., Plucker, Jonathan A. & Hegarty, Boyd, 2012: Replications in Psychology Research: How Often Do They Really Occur? Perspectives on Psychological Science 7(6): 537–542.

Martin, Gabrielle N. & Clarke Richard M., 2017: Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. Frontiers in Psychology 8, Article 523. doi: 10.3389/fpsyg.2017.00523.

Morton, James P., 2009: Reviewing Scientific Manuscripts: How Much Statistical Knowledge Should a Reviewer Really Know? Advances in Physiology Education 33(1): 7-9.

Murphy, Edmond A., 1976: The Logic of Medicine. Baltimore: Johns Hopkins University Press.

Neff, Bryan D. & Olden, Julian D., 2006: Is Peer Review a Game of Chance? BioScience 56(4): 333-340.

Neuliep, James W. & Crandall, Rick, 1993: Reviewer Bias Against Replication Research. Journal of Social Behavior & Personality 8(6): 21-29.

Owen 1982: Reader Bias. Journal of the American Medical Association (JAMA) 247(18): 2533-2534.

Peters, Douglas P. & Ceci, Stephen J., 1982: Peer-review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again. Behavioral and Brain Sciences 5(2): 187-255.

Qi, Xingshun, Deng, Han & Guo, Xiaozhong, 2017: Characteristics of Retractions Related to Faked Peer Reviews: An Overview. Postgraduate Medical Journal 93(1102): 499-503.

Rothwell, Peter M. & Martyn, Christopher N., 2000: Reproducibility of Peer Review in Clinical Neuroscience: Is Agreement Between Reviewers Any Greater Than Would Be Expected by Chance Alone? Brain 123(9): 1964-1969.

Salmon, Wesley C., 1989: Logik. Stuttgart: Philipp Reclam jun.

Seligman, Stanley S., 1991: Assassins and Zealots: Variations in Peer Review. Radiology 178(3): 637-642.

Shatz, David, 2004: Peer Review: A Critical Inquiry. Lanham: Rowman & Littlefield.

Smith, 2008: Peer Review: A Flawed Process at the Heart of Science and Journals.

Sokal, Alan D., 1996: Transgressing the Boundaries: Towards a Transformative Hermeneutics of Quantum Gravity. Social Text, no. 46/47: 217-252.

Sokal, Alan D. & Bricmont, Jean, 1998: Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science. New York: Picador.

Stapleton, Paul, 2012: Gauging the Effectiveness of Anti-plagiarism Software: An Empirical Study of Second Language Graduate Writers. Journal of English for Academic Purposes 11(2): 125-133.

Teixeira da Silva, Jaime A., 2017: The Ethics of Peer and Editorial Requests for Self-citation of Their Work and Journal. Medical Journal of the Armed Forces India 73(2):181–183.

Van Rooyen, Susan, Godlee, Fiona, Evan, Stephen, Black, Nick & Smith, Richard, 1998: Effect of Blinding an Unmasking on the Quality of Peer Review: A Randomized Trial. Journal of the American Medical Association (JAMA) 280(3): 234-237.

Weller, Ann C., 2002: Editorial Peer Review: Its Strengths and Weaknesses. Medford (NJ): Information Today.