Peer-Review

DOI: 10.1002/ange.200800513

Die Effektivität des Peer-Review-Verfahrens: Übereinstimmungsreliabilität und Vorhersagevalidität der Manuskriptbegutachtung bei der *Angewandten Chemie*

Lutz Bornmann und Hans-Dieter Daniel**

Manuskriptbegutachtung · Peer-Review · Übereinstimmungsreliabilität · Vorhersagevalidität

1. Einleitung

In der modernen Wissenschaft ist das Peer-Review zum wichtigsten Bewertungsverfahren avanciert.[1] Für die Befürworter des Verfahrens ist es unverzichtbar, weil nur aktive Wissenschaftler aus demselben Forschungsgebiet – die „Peers" – in der Lage seien, die Forschung ihrer Fachkollegen auf ihre wissenschaftliche Güte hin zu überprüfen.[2] Die Kritiker des Peer-Reviews sehen die Schwächen des Verfahrens insbesondere darin, dass erstens mehrere Gutachter selten in der Bewertung ein und derselben wissenschaftlichen Arbeit übereinstimmen (Problem der Übereinstimmungsreliabilität) und dass zweitens Empfehlungen und Entscheidungen im Peer-Review systematische Urteilsverzerrungen – „Biases" – aufweisen, sodass der Zusammenhang zwischen der Publikationsentscheidung und dem Impact der begutachteten Arbeit nach der Publikation gering sei (Problem der Vorhersagevalidität).[3,4]

Die Peer-Review-Forschung, die sich in den letzten Jahren mit der Kritik am Verfahren beschäftigt hat, bezieht sich zum größten Teil auf das Peer-Review bei Zeitschriften,[5–8] etwas seltener auf das Peer-Review bei Forschungs- und Stipendienanträgen.[9–11] Für Godlee und Jefferson „*the biggest surprise is how little we know about its workings. For a system that demands ever increasing rigor and levels of proof from scientists, it remains itself remarkably untouched by the rigors of science*".[12] Viele Peer-Review-Studien weisen methodische Schwächen auf und „*most of the publications on journal peer review are more opinion than research, often the ruminations of a former editor. Likewise, most of the many letters to editors on the topic, the comments of one kind or another are predominantly opinion*".[13]

In der vorliegenden Studie wird geprüft, ob die Kritik am Peer-Review-Verfahren bei einer Zeitschrift, die wissenschaftliche Originalartikel publiziert, gerechtfertigt ist. Daniel[14] hatte mit der Studie „Guardians of science" das Peer-Review der Zeitschrift *Angewandte Chemie* Mitte der 1980er Jahre evaluiert. Mit einem optimierten Studiendesign und modernen Verfahren der statistischen Datenanalyse haben wir erneut die Güte des Peer-Reviews der *Angewandten Chemie* nach den Kriterien der Übereinstimmungsreliabilität und der Vorhersagevalidität untersucht.

Für die Untersuchung der Reliabilität haben wir das Ausmaß der Übereinstimmung zwischen den Gutachtern bestimmt.[15] Urteile gelten dann als reliabel (zuverlässig), wenn die Empfehlungen mehrerer unabhängiger Gutachter über ein und dasselbe Manuskript übereinstimmen.[16]

Für die Überprüfung der Vorhersagevalidität des Peer-Reviews werden in den Studien zur Manuskriptbegutachtung – aus Mangel an anderen operationalisierbaren Indikatoren – Zitierhäufigkeiten von angenommenen und abgelehnten, jedoch andernorts publizierten Zeitschriftenartikeln verwendet.[17] Nach van Raan geben Zitierungen „*a good to even very good quantitative impression of at least one important aspect of quality, namely international impact*".[18] Nach Lindsey sind Zitierungen „*our most reliable convenient measure of quality in science—a measure that will continue to be widely used*".[19] Wissenschaftliche Urteile über die Publikationswürdigkeit von Zeitschriftenmanuskripten gelten dann als prädiktiv valide, wenn sich die Zitierhäufigkeiten von angenommenen und abgelehnten, jedoch andernorts publizierten Manuskripten statistisch signifikant unterscheiden.[14]

2. Methode

Die *Angewandte Chemie* ist Eigentum der Gesellschaft Deutscher Chemiker (GDCh, Frankfurt am Main) und wird von ihr beim Verlag Wiley-VCH (Weinheim) herausgegeben. Eine „Zuschrift" (Communication, Letter), die bei der Zeitschrift eingereicht wird, wird in der Regel einer internen und einer externen Begutachtung unterzogen.

[*] Dr. L. Bornmann, Prof. Dr. H.-D. Daniel
ETH Zürich, Professur für Sozialpsychologie und Hochschulforschung
Zähringerstraße 24, 8092 Zürich (Schweiz)
Fax: (+41) 44-632-12-83
E-Mail: bornmann@gess.ethz.ch
Prof. Dr. H.-D. Daniel
Universität Zürich
Leiter der Evaluationsstelle
Mühlegasse 21, 8001 Zürich (Schweiz)
Fax: (+41) 44-634-43-79
E-Mail: daniel@evaluation.uzh.ch

Hintergrundinformationen zu diesem Beitrag sind im WWW unter http://dx.doi.org/10.1002/ange.200800513 zu finden.

Zunächst prüft ein Redakteur, inwieweit eine Zuschrift für die Entwicklung eines Teilgebietes der Chemie von großer Bedeutung ist (interne Begutachtung). Wenn diese Bedeutung nach Ansicht des Redakteurs gegeben ist, wird die Zuschrift an mehrere (in der Regel drei)[20] unabhängige, externe Gutachter mit der Bitte gesandt, eine Beurteilung auf einem Evaluationsbogen (auf dem unter anderem die Frage steht: „Do you recommend acceptance of the Communication?“) vorzunehmen und diese auf einem Kommentarblatt zu begründen. Liegen die Beurteilungen dem Redakteur vor, entscheidet er auf ihrer Grundlage über die Annahme oder Ablehnung einer Zuschrift.[21]

Der Überprüfung des Peer-Reviews der *Angewandten Chemie* liegen Informationen zu insgesamt 1899 Zuschriften zugrunde, die im Jahr 2000 begutachtet wurden. Die Informationen wurden dem Archivmaterial der Zeitschrift entnommen, das im Verlagshaus der *Angewandten Chemie* elektronisch erfasst wurde. Mit den Zuschriften aus dem Jahr 2000 ist für die Studie eine etwas „ältere“ Manuskriptkohorte als Datengrundlage ausgewählt worden, da für die Überprüfung der Vorhersagevalidität ein Zeitraum von mehreren Jahren zwischen der Begutachtung einer Zuschrift und der Messung des Indikators für wissenschaftliche Qualität (hier: Zitierungen) liegen sollte.[22] Die 1899 Zuschriften haben neben der internen Begutachtung durch die Redakteure insgesamt 4593 externe Gutachten auf einem Evaluationsbogen und/oder Kommentarblatt erhalten. Im Durchschnitt liegen demnach für eine Zuschrift 2.4 externe Gutachten vor. Auf der Grundlage dieser Gutachten sind von den 1899 Zuschriften 46 % ($n = 878$) für die Publikation in der *Angewandten Chemie* angenommen und 54 % ($n = 1021$) abgelehnt worden.

3. Ergebnisse

3.1. Übereinstimmungsreliabilität des Peer-Reviews

3.1.1. Übereinstimmung zwischen den Empfehlungen der Gutachter

In Tabelle 1 sind Kappa-Koeffizienten zum Ausmaß der Übereinstimmung zwischen zwei oder mehr Gutachtern je Zuschrift bei der Antwort auf die Frage „Do you recommend acceptance of the Communication?“ dargestellt (Antwortvorgaben: „Yes, without alterations“, „Yes, after minor alterations“, „Yes, but only after major alterations“ und „No“). Kappa (κ) ist ein statistisches Maß für die Höhe der Übereinstimmung zwischen zwei oder mehr Gutachtern.[23] κ hat den Wert 1, wenn die Antworten der Gutachter vollständig übereinstimmen; ist der Wert dagegen nahe 0, ist das beobachtete nicht viel größer als das zu erwartende (d. h. das zufällige) Ausmaß der Übereinstimmung. „*Multiplied by 100, κ indicates the percentage by which two raters' agreement exceeds the agreement that could be expected from chance*“.[23] Wie Tabelle 1 zeigt, liegen für das Peer-Review der *Angewandten Chemie* die Kappa-Koeffizienten je nach κ-Statistik (gewichtet oder ungewichtet) und Anzahl der Gutachter, die eine Zuschrift beurteilt haben (zwischen zwei und fünf), im Bereich von 0.10 bis 0.21. Mit anderen Worten: Es werden von den Gutachtern 10 bis 21 % mehr Manuskripte übereinstimmend beurteilt als rein zufällig übereinstimmend beurteilt werden würden. Obwohl unterschiedliche κ-Statistiken berechnet wurden und die Antworten einer unterschiedlich großen Anzahl an Gutachtern je Zuschrift in die Analyse eingingen, unterscheiden sich die ermittelten Werte kaum.[24] Gemessen an den Richtwerten für die Interpretation von κ[23] deutet die Höhe der Koeffizienten in Tabelle 1 alles in allem auf ein geringes Ausmaß an Übereinstimmung zwischen den Gutachterempfehlungen hin. Mit diesem Befund können wir das Ergebnis von Daniel[25] zum Peer-Review der *Angewandten Chemie* Mitte der 1980er Jahre

Tabelle 1: Übereinstimmung zwischen Gutachtern bei der Antwort auf die Frage: „Do you recommend acceptance of the Communication?“[a,b]

Anzahl Gutachter je Zuschrift	Anzahl gemeinsam begutachteter Zuschriften	Beobachtete Übereinstimmung [%]	Erwartete Übereinstimmung [%]	κ[c]	Konfidenzintervall[d]
Zwei Gutachter	952	41.8	31.8	0.15^{u}[e]	0.10–0.19
	952	69.1	61.2	0.21^{g}[e]	0.16–0.25
Drei Gutachter	535			0.10^{u}[e]	0.07–0.14
Zwischen zwei und fünf Gutachter	1507			0.12^{u}[f]	0.09–0.15

[a] Antwortvorgaben: „Yes, without alterations“, „Yes, after minor alterations“, „Yes, but only after major alterations“ und „No“. [b] In die statistische Analyse der Übereinstimmung zwischen den Gutachtern der *Angewandten Chemie* konnten 1507 von insgesamt 1899 Zuschriften einbezogen werden: Von insgesamt 4593 externen Gutachten liegen 3023 mit einem Evaluationsbogen vor, auf dem die Frage „Do you recommend acceptance of the Communication?“ vom externen Gutachter mit „Yes, without alterations“, „Yes, after minor alterations“, „Yes, but only after major alterations“ oder „No“ beantwortet wurde. Diese 3023 Gutachten beziehen sich auf 1840 Zuschriften. Da für die Prüfung der Gutachterübereinstimmung mindestens zwei Gutachten je Zuschrift vorliegen müssen, konnten 952 Zuschriften mit jeweils zwei Gutachten, 535 Zuschriften mit jeweils drei Gutachten, 19 Zuschriften mit jeweils vier Gutachten und eine Zuschrift mit fünf Gutachten in die Analyse einbezogen werden (zu 333 Zuschriften liegt jeweils nur ein Gutachten vor). Neben dem Kappa-Koeffizienten für Zuschriften mit jeweils zwei bis fünf Gutachten ($n = 1507$) sind Koeffizienten für Zuschriften mit jeweils zwei ($n = 952$) oder drei ($n = 535$) Gutachten berechnet worden. Für Zuschriften, bei denen jeweils vier oder fünf Gutachten vorliegen, sind diese Koeffizienten nicht separat berechnet worden, da mit 19 Zuschriften (vier Gutachten) und einer Zuschrift (fünf Gutachten) die Fallzahlen zu gering sind. In der Tabelle werden ungewichtete (^{u}) und gewichtete (^{g}) κ dargestellt. Beim gewichteten κ wird gegenüber dem ungewichteten κ zusätzlich berücksichtigt, dass eine nicht vorhandene Übereinstimmung bei den Antworten von zwei Gutachtern unterschiedlich deutlich ausgeprägt sein kann. So wurde in der Analyse eine Gewichtung von 0.6667 für jene Zuschriften verwendet, bei denen die Gutachter in einer „Zwei-Drittel-Übereinstimmung“ sind (d. h. die Gutachter haben bei der Beantwortung der Frage „benachbarte“ Antwortkategorien gewählt, also z. B. „Yes, without alterations“ und „Yes, after minor alterations“). Eine Gewichtung von 0.3333 wurde bei einer „Ein-Drittel-Übereinstimmung“ eingesetzt (z. B. „Yes, without alterations“ und „Yes, after major alterations“). Schließlich wurden in der Analyse eine Gewichtung von 0 (d. h. keine Gewichtung) bei vollständig konträren Antworten und eine Gewichtung von 1 bei vollständiger Übereinstimmung verwendet. [c] u = ungewichtet, g = gewichtet. [d] Basierend auf 1000 Stichprobentabellen. [e] $p < 0.05$. [f] Signifikanztest kann nicht berechnet werden.

Tabelle 2: Annahme und Ablehnung einer Zuschrift aufgeschlüsselt nach der Antwortkonfiguration von zwei Gutachtern auf die Frage: „Do you recommend acceptance of the Communication?“[a,b]

Antwortkonfiguration von zwei Gutachtern	Anzahl Zuschriften	Annahme ($n=302$)	Ablehnung ($n=416$)
„Yes, without alterations“ – „Yes, without alterations“	15	$^{+}$**100.0**	$^{-}$**.0**
„Yes, without alterations“ – „Yes, after minor alterations“	96	$^{+}$**100.0**	$^{-}$**.0**
„Yes, after minor alterations“ – „Yes, after minor alterations“	100	$^{+}$**98.0**	$^{-}$**2.0**
„Yes, after minor alterations“ – „Yes, but only after major alterations“	44	$^{+}$**97.7**	$^{-}$**2.3**
„Yes, without alterations“ – „Yes, but only after major alterations“	18	$^{+}$**94.4**	$^{-}$**5.6**
„Yes, but only after major alterations“ – „Yes, but only after major alterations“	9	44.4	55.6
„Yes, without alterations“ – „No“	24	37.5	62.5
„Yes, after minor alterations“ – „No“	73	$^{-}$**24.7**	75.3
„Yes, but only after major alterations“ – „No“	91	$^{-}$**2.2**	$^{+}$**97.8**
„No“ – „No“	248	$^{-}$**.0**	$^{+}$**100.0**
Gesamt	718	42.1	57.9

[a] Antwortvorgaben: „Yes, without alterations“, „Yes, after minor alterations“, „Yes, but only after major alterations“ und „No“ (Zeilenprozente, absteigend sortiert nach den Prozentwerten in der Spalte „Annahme“). [b] In diese Analyse konnten 718 von insgesamt 1899 Zuschriften einbezogen werden: Um in der statistischen Analyse zu prüfen, ob die Entscheidung eines Redakteurs auf den Empfehlungen der Gutachter beruht, ist es notwendig, dass bei *allen* Gutachten, die zu *einer* Zuschrift vorliegen, der Gutachter die Frage „Do you recommend acceptance of the Communication?“ beantwortet hat. Im Unterschied zu Tabelle 1 gingen in die vorliegende statistische Analyse keine Manuskripte ein, bei denen ein Obergutachter in die Manuskriptbegutachtung einbezogen wurde bzw. ein überarbeitetes Manuskript oder ein Einspruch, den ein Autor gegen die Ablehnung seines Manuskriptes eingelegt hatte, Gegenstand der gutachterlichen Stellungnahme war. Von den 1069 Zuschriften, bei denen diese Bedingungen erfüllt sind, liegen für 718 Zuschriften jeweils zwei Gutachten, für 241 Zuschriften jeweils drei und für sieben Zuschriften jeweils vier Gutachten vor (bei 103 Zuschriften gibt es jeweils nur ein Gutachten). Da eine statistische Analyse, in die eine Antwortkonfiguration mit drei ($n=241$) bzw. vier ($n=7$) Gutachten pro Zuschrift eingeht, aufgrund der geringen Fallzahlen und der großen Anzahl möglicher Konfigurationen nicht ratsam ist, wurde die Analyse mit 718 Zuschriften, für die jeweils zwei Gutachten vorliegen, durchgeführt. Bei dieser Teilgruppe fällt die Übereinstimmung zwischen den Empfehlungen der Gutachter mit Kappa-Koeffizienten von 0.27 (ungewichtetes κ) und 0.43 (gewichtetes κ) höher aus als in der Gesamtgruppe (siehe Tabelle 1). Im Hinblick auf die Antwortkonfigurationen ist der Unterschied zwischen den Entscheidungen zur Annahme oder Ablehnung statistisch signifikant; χ^2-Test: χ^2 (9, $n=718$) $=$ 606.2, $p<0.0001$. Zellen mit standardisierten Residuen größer 2 (bzw. kleiner -2) sind in der Tabelle fett gedruckt. Residuen sind ein Maß dafür, wie stark beobachtete und erwartete Häufigkeiten voneinander abweichen.

bestätigen. Auch die anderen Studien, die die Übereinstimmungsreliabilität gutachterlicher Empfehlungen für andere Fachzeitschriften untersucht haben, berichten niedrige Kappa-Koeffizienten.[8]

3.1.2. Übereinstimmung zwischen den Empfehlungen der Gutachter und der Entscheidung des Redakteurs

In der Manuskriptbegutachtung trifft der Redakteur seine Entscheidung über die Annahme oder Ablehnung einer Zuschrift auf der Grundlage der gutachterlichen Empfehlungen.[26] Falls die Empfehlungen der Gutachter über ein und dieselbe Zuschrift übereinstimmen, kann der Redakteur im Sinne beider Gutachter entscheiden. Doch wie entscheidet ein Redakteur, wenn von den Gutachtern unterschiedliche Empfehlungen zu einem Manuskript vorliegen (wie z. B. „Yes, after minor alterations“ und „No“)? Für die Beantwortung dieser Frage haben wir für jede Zuschrift aus den Antworten der Gutachter auf die Frage „Do you recommend acceptance of the Communication?“ eine Konfiguration gebildet (z. B. Gutachter 1: „Yes, without alterations“ und Gutachter 2: „Yes, after minor alterations“) und den Zusammenhang zwischen dieser Antwortkonfiguration und der Entscheidung des Redakteurs untersucht.

Wie die Ergebnisse in Tabelle 2 zeigen, korrespondieren Antwortkonfigurationen, bei denen die Gutachter die Annahme einer Zuschrift (mit kleineren oder größeren Überarbeitungsauflagen) empfehlen, mit der Entscheidung zur Annahme; bei den abgelehnten Zuschriften ist es genau umgekehrt. Entsprechend weist die Teststatistik hinsichtlich der Antwortkonfigurationen der Gutachter auf einen höchst signifikanten Unterschied zwischen den Annahme- und Ablehnungsentscheidungen hin (siehe Tabelle 2). Es kann demnach als statistisch abgesichert gelten, dass eine Zuschrift in der Regel dann vom Redakteur angenommen wird, wenn zuvor *beide* Gutachter (Gutachter 1 und Gutachter 2) eine Annahme empfohlen haben.

3.2. Vorhersagevalidität des Peer-Reviews

In den Hintergrundinformationen zu diesem Essay berichten wir ausführlich darüber, in welchen Fachzeitschriften die von der *Angewandten Chemie* abgelehnten Zuschriften publiziert wurden und wie groß die Veränderungen waren, die die Autoren an den bei der *Angewandten Chemie* abgelehnten Zuschriften für die Publikation andernorts vorgenommen haben. In der vorliegenden Arbeit werden wir uns im Folgenden anhand von mittleren Zitierhäufigkeiten für angenommene und abgelehnte, jedoch andernorts erschienenen Zuschriften sowie anhand von „Journal Impact Factors“ (JIFs) der Zeitschriften, in denen die abgelehnten Zuschriften erschienen sind, mit der Vorhersagevalidität des Peer-Reviews der *Angewandten Chemie* beschäftigen. JIFs werden von Thomson Reuters (Philadelphia, PA, USA) in den Journal Citation Reports (JCR) veröffentlicht und geben die „mittlere“ Resonanz an, die ein Artikel in einer Zeitschrift in der Fachwelt erhalten hat.[27] Ein JIF, der von Thomson Reuters für eine Zeitschrift jährlich veröffentlicht wird, ist

ein Quotient aus Zitierungen und zitierbaren Publikationen: Zur Bestimmung des JIFs werden die Artikel einer Zeitschrift jeweils aus einem Zeitfenster von zwei Jahren herangezogen und deren Zitierungen innerhalb des darauf folgenden Jahres ermittelt. Die Anzahl der Zitierungen wird dann durch die Anzahl der zitierbaren Publikationen (d.h. bei der *Angewandten Chemie* im Wesentlichen Zuschriften und wenige Übersichtsartikel) in den beiden vorangegangenen Jahren dividiert.[28]

Von den insgesamt 1021 bei der *Angewandten Chemie* abgelehnten Zuschriften sind nach den Ergebnissen einer Recherche in zwei Literaturdatenbanken 959 als Zeitschriftenbeitrag (93.9%), sieben als Patent und zwei als Beitrag in einem Sammelband erschienen. Zu 53 abgelehnten Zuschriften (5.2%) konnte keine Publikation ermittelt werden. Die 959 Zuschriften, die in einer anderen Zeitschrift erschienen sind, wurden in 136 Zeitschriften publiziert. Jeweils 50 oder mehr abgelehnte Zuschriften haben die Zeitschriften *Chemical Communications* ($n=119$), *Organic Letters* ($n=91$), *Journal of the American Chemical Society* ($n=70$), *Tetrahedron Letters* ($n=60$) und *Organometallics* ($n=50$) veröffentlicht. An etwa drei Viertel der abgelehnten Zuschriften sind keine oder nur geringfügige Veränderungen für die Publikation andernorts vorgenommen worden. Die abgelehnten Zuschriften sind in einer Zeitspanne von sieben Jahren (d.h. zwischen den Jahren 2000 und 2006) von anderen Zeitschriften publiziert worden.

Nach den Angaben in den JCR für das Jahr 2006 variieren die JIFs dieser Zeitschriften zwischen einem Wert kleiner als 1 (z.B. *Chimia* und *Chinese Chemical Letters*) und einem Wert von 9.96 (*Nano Letters*). Gemessen an den JIFs sind die Annahme- und Ablehnungsentscheidungen der Redakteure der *Angewandten Chemie* hoch valide. Keine der insgesamt 956 abgelehnten Zuschriften ist in einer Zeitschrift erschienen, die einen höheren JIF aufweist als die *Angewandte Chemie* (JIF 2006 = 10.23);[*] alle abgelehnten Zuschriften sind in einer Zeitschrift mit einem niedrigeren JIF publiziert worden (mit Ausnahme von 16 abgelehnten Zuschriften, die nach einer erneuten Einreichung in der *Angewandten Chemie* erschienen sind). Dasselbe Ergebnis zeigte sich bereits bei Daniel[14] zu den im Jahr 1984 bei der *Angewandten Chemie* abgelehnten Zuschriften. Diese Befunde bestätigen die allgemeine Beobachtung von Cronin und McKenzie, dass Manuskripte, die von angesehenen Zeitschriften mit einem hohen JIF abgelehnt werden, in der Regel bei Zeitschriften mit einem niedrigeren JIF eingereicht (und publiziert) werden: „*It is widely recognised that there is an informal journal pecking order in almost every discipline, and that a manuscript rejected by a high-ranking journal will often be re-submitted to one of lesser repute*“.[29]

Der JIF stellt ein nur sehr grobes Maß für die Bestimmung der Vorhersagevalidität dar, weil alle Beiträge in einer Zeitschrift durch einen Durchschnittswert charakterisiert werden.[30] Sehr häufig zitierte Beiträge werden dadurch unterbewertet, nicht oder kaum zitierte dagegen überbewertet.[14] Deshalb haben wir über die JIFs hinaus ermittelt, wie häufig die angenommenen und abgelehnten, jedoch andernorts erschienenen Zuschriften nach ihrer Veröffentlichung bis Ende 2006 zitiert wurden. Die Zitierhäufigkeiten für die einzelnen Zuschriften wurden in der Literaturdatenbank Scopus recherchiert.[31] Scopus ist eine neue, fachübergreifende Datenbank der Firma Elsevier (Amsterdam, Niederlande), die über 33 Millionen Abstracts und Quellenangaben von etwa 15000 Zeitschriftentiteln enthält, die von mehr als 4000 Verlagen herausgegeben werden. Sie gilt als derzeit größte multidisziplinäre Datenbank für wissenschaftliche Literatur.[32] Nach den Ergebnissen von Norris und Oppenheim[33] kann Scopus als eine gleichwertige Alternative zu Web of Science (WoS) von Thomson Reuters angesehen werden.

Wie die Ergebnisse für die Publikationsjahre 2000 und 2001 in Abbildung 1 zeigen, sind die angenommenen Zuschriften deutlich häufiger zitiert worden als die abgelehnten (ein ähnliches Ergebnis zeigte sich bei der Analyse von Zitierungen, die im Science Citation Index (SCI, Thomson Reuters) und in der Datenbank Chemical Abstracts (Chemical Abstracts Services, Columbus, OH, USA) recherchiert wurden).[34] Für die Publikationsjahre 2002 bis 2006 konnte dieser Vergleich nicht durchgeführt werden, da nach 2001 nur abgelehnte und andernorts publizierte, jedoch *keine* angenommenen Zuschriften (die im Jahr 2000 begutachtet wurden) in der *Angewandten Chemie* erschienen sind (die Publikationsfristen bei der *Angewandten Chemie* sind vergleichsweise kurz). Die Mittelwertsunterschiede zwischen den Zitierhäufigkeiten sind statistisch signifikant (siehe Abbildung 1). Dieses Ergebnis korrespondiert mit dem Befund, dass die abgelehnten Manuskripte in Zeitschriften mit einem niedrigeren JIF erschienen sind (siehe oben). Da sich – wie das Ergebnis der Zitationsanalyse verdeutlicht – die Zitierhäufigkeiten der angenommenen und abgelehnten Zuschriften statistisch signifikant unterscheiden, bestätigen die Befunde der Zitationsanalyse die Vorhersagevalidität des Peer-Review-Verfahrens der *Angewandten Chemie*.

4. Diskussion

Vor dem Hintergrund der Kritik am Peer-Review-Verfahren ist in der vorliegenden Studie das Peer-Review der *Angewandten Chemie* nach den Kriterien der Übereinstimmungsreliabilität und der Vorhersagevalidität untersucht worden. Im Hinblick auf die Reliabilität des Peer-Reviews deutet die Höhe der Kappa-Koeffizienten (zwischen 0.10 und 0.21) auf ein geringes Ausmaß der Übereinstimmung zwischen den Empfehlungen der Gutachter zu einer Zuschrift hin. Über ein ähnliches Ergebnis zum Peer-Review der *Angewandten Chemie* berichtete bereits Daniel.[25]

[*] Legen wir die JIFs zugrunde, die von Thomson Reuters in den JCR für die Jahre 2000 und 2001 veröffentlicht wurden (d.h. für die Jahre, in denen die meisten der abgelehnten Manuskripte andernorts publiziert wurden), ist insgesamt nur ein abgelehntes Manuskript in einer Zeitschrift mit einem höheren JIF erschienen, nämlich in *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) (JIF 2000 = 10.79; JIF 2001 = 10.90). Die damaligen JIFs der *Angewandten Chemie* betrugen 8.55 für das Jahr 2000 und 8.26 für das Jahr 2001. In den JCR für das Jahr 2006 hat die *Angewandte Chemie* (JIF 2006 = 10.23) einen höheren JIF als PNAS (JIF 2006 = 9.64).

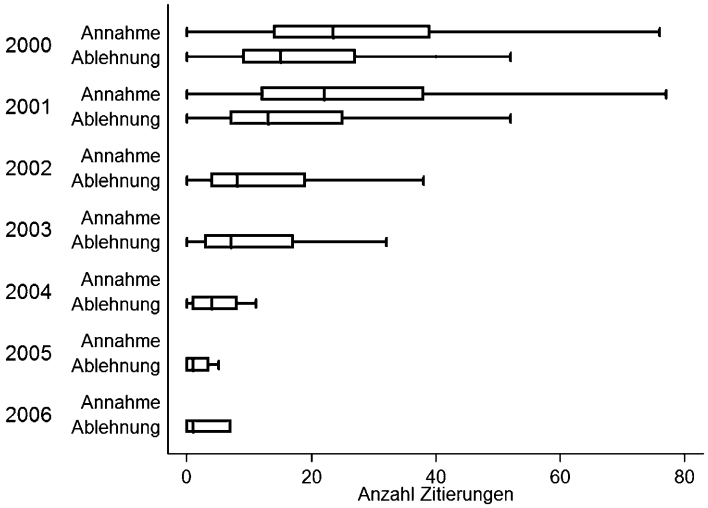

Abbildung 1. Box-Plots: Mittlere Zitierhäufigkeiten (die vertikale Linie in der Mitte der Box gibt den Medianwert an) aufgeschlüsselt nach dem Publikationsjahr der angenommenen und abgelehnten, jedoch andernorts publizierten Zuschriften und der Entscheidung des Redakteurs (Ausreißer sind in der Abbildung nicht dargestellt). In die Analyse konnten 1827 von insgesamt 1899 Zuschriften einbezogen werden: 62 der bei der *Angewandten Chemie* abgelehnten Zuschriften sind – nach unserer Recherche – nicht andernorts publiziert worden. Zu 10 abgelehnten Zuschriften, die andernorts erschienen sind, konnten keine Zitierungen in Scopus recherchiert werden. Der Unterschied zwischen den mittleren Zitierhäufigkeiten für angenommene (Median = 23.5) und abgelehnte (Median = 15) Manuskripte, die im Jahr 2000 publiziert wurden, ist statistisch signifikant; U-Test nach Mann und Whitney: $Z(n=778) = -6.29$, $p < 0.0001$. Der Unterschied zwischen den mittleren Zitierhäufigkeiten für angenommene (Median = 22) und abgelehnte (Median = 13) Manuskripte, die im Jahr 2001 publiziert wurden, ist statistisch ebenfalls signifikant; U-Test nach Mann und Whitney: $Z(n=885) = -7.45$, $p < 0.0001$. Nach 2001 wurden nur abgelehnte Zuschriften andernorts veröffentlicht.

Obwohl eine hohe Übereinstimmung bei Bewertungen generell als wünschenswert angesehen wird, wird sie im Hinblick auf das Peer-Review von einigen Wissenschaftlern, wie z. B. Bailar, als nachteilig für den Begutachtungsprozess beurteilt: „*Too much agreement is in fact a sign that the review process is not working well, that reviewers are not properly selected for diversity, and that some are redundant*“.[35] Durch die Auswahl von Gutachtern nach dem Komplementaritätsprinzip (z. B. Auswahl eines Generalisten und eines Spezialisten) werde zwar die Übereinstimmungsreliabilität verringert; davon könne aber nach Langfeldt die Validität des Verfahrens profitieren: „*Low inter-reviewer agreement on a peer panel is no indication of low validity or low legitimacy of the assessments. In fact, it may indicate that the panel is highly competent because it represents a wide sample of the various views on what is good and valuable research*“.[36] Unterschiedliche Empfehlungen bei der Begutachtung seien nicht unbedingt ein Zeichen für Uneinigkeit, sondern würden sich durch unterschiedliche paradigmatische Positionen („Schulen“), Herangehensweisen und Mentalitäten der Gutachter erklären lassen.[37] Auch können Gutachter eher streng oder mild urteilen;[38] sie richten – nach Eckberg – ihre Aufmerksamkeit auf „*different points, and may draw different conclusions about ‚worth‘*“.[39]

Die Recherche nach dem Schicksal der bei der *Angewandten Chemie* abgelehnten Zuschriften hat gezeigt, dass bis auf wenige Ausnahmen alle Zuschriften andernorts publiziert wurden. Die andernorts veröffentlichten Zuschriften erschienen immer in Zeitschriften, die einen niedrigeren JIF als die *Angewandte Chemie* aufweisen. Die Zitationsanalyse hat gezeigt, dass angenommene Zuschriften im Durchschnitt statistisch signifikant häufiger zitiert werden als abgelehnte, jedoch andernorts publizierte Zuschriften. Beide Befunde bestätigen die Vorhersagevalidität des Peer-Reviews der *Angewandten Chemie*. Über ein ähnliches Ergebnis haben bislang nicht nur Daniel[14] für die *Angewandte Chemie*, sondern auch Lock,[40] Wilson[41] sowie Opthof, Furstner, van Geer und Coronel[42] für das *British Medical Journal*, das *Journal of Clinical Investigation* und für die Zeitschrift *Cardiovascular Research* berichtet.

Obwohl das Peer-Review der *Angewandten Chemie* seit Mitte der 1980er Jahre (d. h. seit der Studie von Daniel)[14] unter deutlich veränderten Rahmenbedingungen stattfindet (wie z. B. einer drastischen Zunahme der eingereichten Zuschriften und einem deutlichen Rückgang der Annahmequote), belegen die Ergebnisse der vorliegenden Studie zur Übereinstimmungsreliabilität und Vorhersagevalidität die nach wie vor hohe Güte des Peer-Reviews der *Angewandten Chemie*.

Wir möchten uns bei Dr. Christophe Weymuth (vormals am organisch-chemischen Institut der Universität Zürich, derzeitig bei der BIOSYNTH AG in der Schweiz tätig) für die Recherche nach den bei der Angewandten Chemie abgelehnten und andernorts publizierten Manuskripten bedanken. Ferner danken wir Dr. Andras Schubert (Hungarian Academy of Sciences) für die Bereitstellung von Journal Impact Factors, die von Thomson Reuters in den Journal Citation Reports für die Jahre 2000 und 2001 veröffentlicht wurden, für die Angewandte Chemie und für die Zeitschriften, in denen abgelehnte, jedoch andernorts publizierte Manuskripte erschienen sind. Wir danken Dr. Peter Gölitz, Chefredakteur der Angewandten Chemie, dem Kuratorium der Angewandten Chemie und der Gesellschaft Deutscher Chemi-

ker (GDCh) für die Erlaubnis, das Gutachterwesen der Zeitschrift einer Verfahrensevaluation unterziehen zu dürfen, sowie den Mitarbeitern der Redaktion für die großzügig gewährte Unterstützung während der Durchführung der Studie. Die Gesamtstudie, in der auch die Qualitätssicherung bei Open-Access-Zeitschriften untersucht wird, wird von der Max-Planck-Gesellschaft unterstützt. Die Autoren möchten sich bei drei anonymen Gutachtern für ihre wertvollen Kommentare bedanken.

Eingegangen am 31. Januar 2008,
veränderte Fassung am 27. März 2008
Online veröffentlicht am 25. Juli 2008

[1] J. Ziman, *Real Science. What it is, and what it means*, Cambridge University Press, Cambridge, **2000**.
[2] M. Eisenhart, *Res. Sci. Educ.* **2002**, *32*, 241.
[3] P. F. Ross, *The Sciences' Self-management: Manuscript Refereeing, Peer Review, and Goals in Sience*, The Ross Company, Todd Pond, **1980**.
[4] H. J. Eysenck, S. B. G. Eysenck, *Pers. Indiv. Differ.* **1992**, *13*, 393.
[5] J. M. Campanario, *Sci. Commun.* **1998**, *19*, 277.
[6] J. M. Campanario, *Sci. Commun.* **1998**, *19*, 181.
[7] J. Overbeke, E. Wager in *Peer Review in Health Sciences*, 2nd ed. (Hrsg.: F. Godlee, T. Jefferson), BMJ, London, **2003**, S. 45.
[8] A. C. Weller, *Editorial Peer Review: Its Strengths and Weaknesses*, Information Today, Medford, **2002**.
[9] V. Demicheli, C. Pietrantonj in *The Cochrane Library*, *Issue 1*, Wiley, Chichester, **2004**.
[10] S. Wessely, *Lancet* **1998**, *352*, 301.
[11] L. Bornmann, H.-D. Daniel in *Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung* (Hrsg.: S. Schwarz, U. Teichler), Campus, Frankfurt, **2003**, S. 211.
[12] *Peer Review in Health Sciences*, 2nd ed. (Hrsg.: F. Godlee, J. Jefferson), BMJ, London, **2003**.
[13] M. F. Stieg Dalton, *ARIST* **1995**, *30*, 213.
[14] H.-D. Daniel, *Guardians of Science. Fairness and Reliability of Peer Review*, Wiley-VCH, Weinheim, **1993/2004**. Online veröffentlicht am 16. Juli 2004, Wiley Interscience, DOI: 10.1002/3527602208.
[15] D. V. Cicchetti, *Behav. Brain Sci.* **1991**, *14*, 119.
[16] J. Ziman, *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*, Cambridge University Press, Cambridge, **1991**.
[17] H.-D. Daniel, *Learn. Publ.* **2005**, *18*, 143.
[18] A. F. J. van Raan, *Scientometrics* **1996**, *36*, 397.
[19] D. Lindsey, *Scientometrics* **1989**, *15*, 189.
[20] Anonymous, *Chem. Eng. News Peer Review* **2008**, *86*(6), 48.
[21] A. M. Coghill, L. R. Garson, *The ACS Style Guide*, 3rd ed., American Chemical Society, Washington, **2006**.
[22] H. P. F. Peters, A. F. J. van Raan, *J. Am. Soc. Inf. Sci.* **1994**, *45*, 39.
[23] A. von Eye, E. Y. Mun, *Analyzing Rater Agreement. Manifest Variable Methods*, Lawrence Erlbaum, Mahwah, **2005**.
[24] U. W. Jayasinghe, H. W. Marsh, N. Bond, *Educ. Eval. Policy. Anal.* **2001**, *23*, 343.
[25] H.-D. Daniel, *Angew. Chem.* **1993**, *105*, 247; *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 234.
[26] V. Bakanic, C. McPhail, R. J. Simon, *Am. Sociol. Rev.* **1987**, *52*, 631.
[27] L. Bornmann, L. Leydesdorff, W. Marx, *CHIMIA* **2007**, *61*, 104.
[28] W. Marx, H. Schier, *Phys. Bl.* **2001**, *57*, 25.
[29] B. Cronin, G. McKenzie, *J. Doc.* **1992**, *48*, 310.
[30] T. Braun, I. Dióspatonyi, S. Zsindely, E. Zádor, *Scientometrics* **2007**, *71*, 541.
[31] J. Bar-Ilan, M. Levene, A. Lin, *J. Informetrics* **2007**, *1*, 26.
[32] F. de Moya-Anegón, Z. Chinchilla-Rodríguez, B. Vargas-Quesada, E. Corera-Álvarez, F. Muñoz-Fernández, A. González-Molina, V. Herrero-Solana, *Scientometrics* **2007**, *73*, 53.
[33] M. Norris, C. Oppenheim, *J. Informetrics* **2007**, *1*, 161.
[34] L. Bornmann, H.-D. Daniel, *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 830.
[35] J. C. Bailar, *Behav. Brain Sci.* **1991**, *14*, 137.
[36] L. Langfeldt, *Soc. Stud. Sci.* **2001**, *31*, 820.
[37] R. N. Kostoff, *Scientometrics* **1995**, *34*, 163.
[38] S. S. Siegelman, *Radiology* **1991**, *178*, 637.
[39] D. L. Eckberg, *Behav. Brain Sci.* **1991**, *14*, 145.
[40] S. Lock, *A Difficult Balance: Editorial Peer Review in Medicine*, ISI, Philadelphia, **1985**.
[41] J. D. Wilson, *J. Clin. Invest.* **1978**, *61*, 1697.
[42] T. Opthof, F. Furstner, M. van Geer, R. Coronel, *Cardiovasc. Res.* **2000**, *45*, 255.

