SCM
Studies in Communication and Media

FULL PAPER

Wie zuverlässig ist das Peer-Review-Verfahren?
Eine Untersuchung der Interrater-Reliabilität von Gutachter*innen auf DGPuK-Tagungen

On the reliability of peer-reviewing
A study of inter-rater reliability between reviewers for DGPuK conference submissions

Thomas Koch & Stefan Geiß

Thomas Koch (Prof. Dr.), Department of Communication, Johannes Gutenberg-University Mainz, Jakob-Welder-Weg 12, 55128 Mainz, Germany, Contact: thomas.koch(at)uni-mainz.de

Stefan Geiß (Prof. Dr.), Department of Sociology and Political Science, Norwegian University of Science and Technology (NTNU), Dragvoll, 7055 Trondheim, Norway, Contact: stefan.geiss(at)ntnu.no

FULL PAPER

Wie zuverlässig ist das Peer-Review-Verfahren?
Eine Untersuchung der Interrater-Reliabilität von Gutachter*innen auf DGPuK-Tagungen

On the reliability of peer-reviewing
A study of inter-rater reliability between reviewers for DGPuK conference submissions

Thomas Koch & Stefan Geiß

Zusammenfassung: Bevor wissenschaftliche Beiträge in Fachzeitschriften publiziert oder auf Tagungen präsentiert werden, überprüfen Herausgeber*innen bzw. Organisator*innen die Qualität der Einreichungen. Dies geschieht zumeist im Peer-Review-Verfahren, bei dem unabhängige Kolleg*innen aus dem gleichen Forschungsgebiet die Einreichung begutachten. Die vorliegende Studie hinterfragt, wie zuverlässig das Review-Verfahren ist. Dazu untersuchen wir die Bewertungen der Einreichungen von DGPuK-Jahrestagungen und der Tagungen der fünf größten Fachgruppen über einen Zeitraum von fünf Jahren. Basierend auf 3537 Reviews von 23 Tagungen analysieren wir Interrater-Reliabilitäten (Krippendorffs α und Brennan und Predigers κ) und Spannweiten über verschiedene Einzelkriterien (Passung, Originalität, Relevanz, Theorie, Methode und Darstellung) und Gesamturteile; zudem fokussieren wir Ursachen von Dissens bzw. Konsens. Die Studie zeigt, dass unter Gutachter*innen durchaus Uneinigkeit besteht: Dies betrifft sowohl die Gesamtwertung als auch alle Einzelwertungskategorien. Die Bildung von Durchschnitten über mehrere Kriterien hinweg erhöht jedoch die Übereinstimmung der Urteile. Abschließend diskutieren wir Ideen, um Begutachtungsverfahren zukünftig zu verbessern.

Schlagwörter: Wissenschaftliche Qualitätssicherung; Peer-Review-Verfahren; Gutachter*innen; Übereinstimmung; Reliabilität; Reviewer*innen

Abstract: Journal editors and conference organizers frequently rely on peer-reviewing to assess the quality of submissions. Peer-reviewing is a technique in which independent colleagues with expertise in the same area of research rate the submission. The present study investigates the reliability of review ratings by different reviewers. To that end we studied the reviews made for the general conference of the German Communication Association (DGPuK) and the annual conferences of its five largest divisions (Fachgruppen) in the past five conferences. Based on 3537 reviews from 23 conferences, we analyze inter-rater reliability (Krippendorff's α and Brennan und Prediger's κ) and ranges, regarding both criteria-based scores (fit with conference theme, innovativeness, relevance, theory, method, clarity of presentation) and overall scores. The study shows that there is substantial disagreement between reviewers. This applies to overall scores as well as criteria-based scores. Calculating mean or sum scores across criteria leads to higher agreement between reviewers. We discuss potential modifications to optimize review procedures.

Keywords: Quality Assurance in Research; Peer Reviewing; Reviewer; Agreement; Reliability

1. Einleitung

Bei der Begutachtung eines wissenschaftlichen Abstracts, das für eine DGPuK-Fachgruppentagung 2016 eingereicht wurde, kamen drei Gutachter*innen zu völlig unterschiedlichen Urteilen: Reviewer 1 empfand den Beitrag als voll und ganz überzeugend und äußerte nur einen kleinen Kritikpunkt bei der Messung eines Konstrukts; auf der Skala von 0-5 vergab er über die vier bewerteten Kriterien hinweg 4,25 Punkte. Reviewerin 2 erkannte diverse Schwächen in dem Beitrag, empfand ihn jedoch als brauchbar und empfahl den Autor*innen, diese Schwächen noch vor der Präsentation zu beheben; sie vergab 2,5 Punkte im Schnitt. Reviewerin 3 beurteilte das Abstract als desaströs, kritisierte Theorie, Hypothesenableitung, Schreibstil, Verständlichkeit, mangelnde konzeptuelle Ausdifferenzierung, Methode und Datenauswertung sowie die daraus abgeleiteten Implikationen. Das Paper erhielt 0,5 Punkte. Der Autor des Abstracts fragte daraufhin die Organisatoren des Einreichungsverfahrens, ob solche Diskrepanzen zwischen Gutachter*innen häufiger vorkommen. Diese Frage konnte bislang nur aus dem Bauch heraus beantwortet werden – denn obwohl valide Daten dazu natürlich existieren (die Ergebnisse der Begutachtungsverfahren liegen ja vor), wurden diese weder systematisch gesammelt noch ausgewertet.

So war die Idee für die vorliegende Studie geboren. Sie widmet sich der Frage, wie reliabel die Begutachtung von Tagungseinreichungen in der deutschsprachigen Kommunikationswissenschaft ist: Stimmen Reviewer*innen im Fach zuverlässig überein, welche Beiträge angenommen bzw. abgelehnt werden sollten und wie häufig kommen konträre Bewertungen vor? Unterscheidet sich die Übereinstimmung zwischen verschiedenen Bewertungskriterien (z.B. Methode, Theorie, Darstellung, etc.)? Und lassen sich Fallstricke und Probleme identifizieren, die das Begutachtungsverfahren erschweren? Diesen Fragen werden wir im vorliegenden Beitrag nachgehen und analysieren, wie gut die wissenschaftliche Qualitätssicherung mittels Peer-Review-Verfahren bei kommunikationswissenschaftlichen Tagungen gelingt (Armstrong, 1997; Bailar, 1991; Campanario, 1998). Dies ist über die Kommunikationswissenschaft hinaus relevant: Die täglich betriebene Wissenschaftsevaluation entscheidet über die Themen von (Top-)Journals und Tagungen und bestimmt somit maßgeblich die Agenda der Wissenschaftscommunity; zudem entscheidet sie über die Reputation der Einreichenden und über die Verteilung von Stipendien, Auszeichnungen und Forschungsgeldern (Marsh, Jayasinghe, & Bond, 2008; Smith, 2006).

Der vorliegende Beitrag analysiert die Zuverlässigkeit des Peer-Review-Verfahrens anhand von Streuungsmaßen sowie der Interrater-Reliabilität von Reviewer*innen auf ausgewählten DGPuK-Tagungen. Dabei geben wir zunächst einen theoretischen Überblick über Peer-Review-Verfahren und die damit verbundenen Chancen und Risiken. Anschließend systematisieren wir Untersuchungen, die bereits die Zuverlässigkeit dieser Bewertungen erforscht haben und setzen uns mit Interrater-Reliabilitäten als möglichem Gütekriterium auseinander. Im empirischen Teil der Arbeit analysieren wir die Ergebnisse der jeweils letzten fünf Reviewverfahren der DGPuK- Jahrestagungen sowie die jeweils letzten fünf Verfahren der fünf größten Fachgruppen. Auf Basis dieser Analyse diskutieren wir schließlich Möglichkeiten zur Optimierung von Reviewverfahren.

2. Peer-Review und damit verbundene Probleme

Bevor wissenschaftliche Beiträge in Fachzeitschriften publiziert oder auf Tagungen präsentiert werden, überprüfen Herausgeber*innen bzw. Organisator*innen in der Regel die Qualität der Einreichungen. Dies soll gewährleisten, dass nur die besten Studien angenommen werden (Smith, 2006). Sehr weit verbreitet ist dabei das Peer-Review-Verfahren, also die Begutachtung eines Beitrags durch sogenannte Peers (Armstrong, 1997; Campanario, 1998; Cole, Cole, & Simon, 1981). Schon die Frage, wer eigentlich diese Peers sind, wird kontrovers diskutiert: Ist es einfach jemand aus dem gleichen Fach oder eine Person, die ähnliche Themen bearbeitet oder sollte die Person sogar den exakt gleichen Fragestellungen nachgehen (wodurch Gutachtende und Einreichende allerdings auch direkte Konkurrenz oder enge Kolleg*innen sein können; Smith, 2006)? Wie soll man mit dem Einfluss von Konkurrenz und Nähe umgehen, wer ist vom Reviewverfahren auszuschließen, wo zieht man die Grenze? Zudem ist diskussionswürdig, was genau ein Review beinhaltet: Geht es nur um ein „Daumen hoch – Daumen runter" oder um die akkurate Beurteilung von Relevanz, Aktualität und thematischer Passung, das Erkennen und Beheben von Schwächen in Theorie, Methodik, Auswertung und Schlussfolgerungen, um letztlich die Einreichung zu verbessern? Letzteres liegt nahe, weil aussichtsreichen Einreichungen oftmals die Chance zur Korrektur bestimmter Fehler oder Schwächen gegeben wird (Olson, 1990). Die Unschärfe dieser beiden Begriffe erschwert eine Definition, zumal das Verfahren in verschiedenen Disziplinen und Kontexten sehr unterschiedlich angewandt wird (Cicchetti, 1991; Wilson, 1978). Um dieser Heterogenität gerecht zu werden, definieren wir Peer-Review recht umfassend als einen Prozess, bei dem die Qualität von wissenschaftlichen Einreichungen begutachtet wird; diese Einschätzung sollte dabei von Personen vorgenommen werden, die über die Kompetenz verfügen, ein solches Urteil zu treffen (meist unabhängige Kolleg*innen aus dem gleichen Forschungsgebiet).

Zumeist folgt das Verfahren einer Doppel-Blind-Begutachtung: Die Gutachter*innen werden nicht darüber in Kenntnis gesetzt, wer die Autor*innen sind und letztere wissen auch nicht, wer den Beitrag begutachtet (Ceci & Peters, 1984). Die Anonymisierung soll verhindern, dass persönliche Beziehungen zwischen den Beteiligten oder die Bekanntheit eines Wissenschaftlers die Unabhängigkeit des Begutachtungsprozesses gefährden (McNutt, Evans, Fletcher, & Fletcher, 1990). Auch soll es unterbinden, dass Autor*innen die Gutachter*innen (oder umgekehrt) kontaktieren und zu beeinflussen versuchen. In den letzten Jahren gab es immer wieder Initiativen, Alternativen zur blinden Begutachtung zu schaffen, etwa indem das Verfahren als Open Review durchgeführt und z.B. die Namen der Gutachter*innen und/oder die Gutachten mit publiziert werden (van Rooyen, Godlee, Evans, Black, & Smith, 1999). Das Prinzip des Peer-Review-Verfahrens wird nicht nur bei wissenschaftlichen Zeitschriften und Tagungen genutzt, sondern findet u. a. auch bei Sammelbänden, Forschungsanträgen, Auszeichnungen oder Tenure-Entscheidungen in beinahe allen wissenschaftlichen Disziplinen statt (Neidhardt, 2010; Neumann et al., 2008).

Peer-Review entwickelte sich zunächst, weil die Herausgeber*innen von Zeitschriften im sich immer weiter ausdifferenzierenden Wissenschaftssystem

Fachleute im jeweiligen Gebiet hinzuziehen mussten, um die Qualität von Beiträgen prüfen zu können. Alternativen zum „klassischen" Blind-Peer-Review-Verfahren wurden und werden immer wieder getestet. Eine Möglichkeit ist dabei das oben angesprochene Open Peer-Review, bei dem Namen der Gutachter*innen (und teilweise auch der Autor*innen) transparent kommuniziert werden (Suls & Martin, 2009). Die Gutachten können auch mit dem überarbeiteten Aufsatz online gestellt werden, um den Überarbeitungsprozess transparent zu machen. Ein Vorteil bei der Aufhebung der Anonymisierung könnte darin liegen, dass Reviewer*innen seltener ethisch bedenklich urteilen und sich mehr Mühe bei den Gutachten geben. Einige der zahlreichen Nachteile haben wir im vorausgehenden Absatz geschildert. Andere Alternativen setzen auf Selbstselektion der Gutachter*innen: So werden eingereichte Aufsätze (meist nach kurzer Prüfung auf thematische Passung und grobe Fehler) direkt online gestellt und entweder einer spezifischen Community an Gutachter*innen oder potentiell allen Interessierten die Möglichkeit zur Kommentierung und Auseinandersetzung mit der Studie gegeben. Anschließend kann ein Urteil z.B. in Abhängigkeit vom geweckten Interesse oder basierend auf den Kommentaren fallen (Birukou et al., 2011; Suls & Martin, 2009). Eine weitere Alternative ist das sogenannte Adversary Model, bei dem Reviewer*innen gebeten werden, nicht differenziert über eine Einreichung zu urteilen, sondern das Paper explizit in Frage stellen und mit aller Kraft anzweifeln und widerlegen sollen. Gutachter*innen nehmen also die Rolle eines Anklägers ein, woraufhin die Autor*innen sich schriftlich verteidigen können und die Herausgeber*innen im Anschluss eine Entscheidung trifft (Bornstein, 1991). All diese Verfahren setzen bei bestimmten Mängeln des traditionellen Peer-Review an, können allerdings negative Folgewirkungen haben, z. B. wenn (1) Studien vorschnell der Öffentlichkeit präsentiert werden und falsche oder irreführende Schlussfolgerungen ziehen, (2) Studien nachträglich ständig verändert werden und eine definitive Finalversion, auf die man sich berufen kann, fehlt oder (3) die transparente Dokumentation des Reviews und der Überarbeitungen die Komplexität des Veröffentlichungsprozesses sowie das Rezipieren und Zitieren deutlich erhöht.

Derartige Alternativen haben sich in der Kommunikationswissenschaft allerdings noch nicht etabliert, weshalb das Peer-Review-Verfahren bei Fachtagungen und Zeitschriften als Standard der Qualitätssicherung gilt (Lauf, 2001; Meyen & Wiedemann, 2016). In Deutschland gibt es diesen Standard noch nicht allzu lange: Langenbucher (2016) beschreibt, dass die Zeitschrift „Publizistik" beispielsweise erst ab 2003 um anonymisierte Einreichungen zur Begutachtung bat und erst 2007 ankündigte, bei der Begutachtung regelmäßig externe Reviewer*innen einzubeziehen. Auch die Umstellung der DGPuK-Jahrestagung auf ein anonymisiertes Review-Verfahren erfolgte erst 2001 (Meyen & Wiedemann, 2016). Mittlerweile sind beispielsweise DGPuK-Jahrestagungen oder Fachgruppentagungen ohne ein Peer-Review-Verfahren schwer vorstellbar.

Spätestens seit den 1970er Jahren steht das Peer-Review-Verfahren selbst wissenschaftstheoretisch auf dem Prüfstand. Besonders in der Medizin und Psychologie wurde das Verfahren mehrfach untersucht und steht aus vielen Gründen in der Kritik (Hirschauer, 2004; Neff & Olden, 2006). Peer-Review gilt als langsam, teuer, subjektiv, anfällig für vorurteilsgeprägte Verzerrungen, leicht zu missbrau-

chen und fast nutzlos bei der Ermittlung von Betrug (Armstrong, 1997; Ceci & Peters, 1982; Marsh, Bond, & Jayasinghe, 2007; Zuckerman & Merton, 1971). Zudem lasse das Verfahren die Publikationswahrscheinlichkeit bestimmter Studien sinken: Dies betrifft z. B. Replikationen (Boster, 2002; Campanario, 1998; Neuliep & Crandall, 1990; Schmidt, 2009), nicht signifikante Befunde (Armstrong, 1997; Bornstein, 1991) oder unkonventionelle Beiträge, die den vorherrschenden paradigmatischen Ansichten entgegenlaufen (Epstein, 1990; Eysenck & Eysenck, 1992; Mahoney, 1987). Reviewer*innen nutzen zudem selten systematisches Vorgehen, sondern orientieren sich oftmals an leicht verfügbaren Hinweisreizen, die nicht unbedingt mit der Qualität einer Einreichung zusammenhängen: Typische Heuristiken sind dabei die (Nicht-)Signifikanz von Befunden, die Größe der Stichprobe, die Komplexität der Rechnungen und die Klarheit der Darstellung (Armstrong, 1997).

3. Reliabilität von Peer-Review-Verfahren

Neben den gerade angeführten Kritikpunkten wird auch über die Frage der Zuverlässigkeit bzw. Reliabilität des Peer-Review-Verfahrens häufig diskutiert (Marsh, Bond, & Jayasinghe, 2007). Ein geeignetes Instrument muss bei wiederholtem Einsatz zu dem gleichen Ergebnis kommen: Ein reliables Peer-Review-Verfahren sollte also sicherstellen, dass ein Manuskript bei derselben Zeitschrift oder Tagung ähnliche Gutachten bzw. Entscheidungen bekommt (Cicchetti, 1991). Dies wird dann gewährleistet, wenn (1) die Urteile verschiedener Gutachter*innen möglichst frei von Fehlern und Abweichungen sind, die z. B. durch subjektive Vorlieben, individuelle Fehler oder Vorurteile entstehen, und wenn (2) die beteiligten Gutachter*innen ähnliche Kriterien und Maßstäbe zur Evaluation nutzen. Um die Reliabilität der Urteile von Gutachter*innen zu überprüfen, gibt es zwei grundlegende Möglichkeiten: (1) Man testet die Intrarater-Reliabilität: Kommt der/die gleiche Gutachter/in bei nochmaliger Analyse des Beitrags zu dem gleichen Ergebnis? (2) Man prüft die Interrater-Reliabilität: Wie gut stimmen zwei oder mehr Gutachter*innen des gleichen wissenschaftlichen Dokuments in ihren Urteilen überein (Cicchetti, 1991; LeBreton & Senter, 2008)?

Weil wissenschaftliche Beiträge in der Regel von mehreren Gutachter*innen beurteilt werden, steht insbesondere die Frage nach der Interrater-Reliabilität im Vordergrund (Bornmann, Mutz, & Daniel, 2010). Die Frage, inwieweit Einschätzungen verschiedener Gutachter*innen bei denselben Beiträgen übereinstimmen, drängt sich hier geradezu auf. Bisherige Studien auf diesem Gebiet kommen zu unterschiedlichen Befunden, wenngleich sie in der methodischen Anlage recht ähnlich sind (vgl. Übersichten bei Bornmann, Mutz, & Daniel, 2010; Cicchetti, 1991; Goldman, 1994; Lindsey, 1988). Sie greifen meist auf bestehende Reviews von Tagungs- oder Zeitschrifteneinreichungen zurück und prüfen mittels verschiedener Koeffizienten, inwiefern die Gutachter*innen der gleichen Einreichung zu ähnlichen Einschätzungen kommen. Manchmal basieren die Studien auf binären Entscheidungen der Gutachter*innen (meist Annahme vs. Ablehnung), teilweise liegen ordinal skalierte Urteile zugrunde (z. B. Annahme – Minor Revision – Ma-

jor Revision – Ablehnung) und oftmals kommen auch quasi-metrische Entscheidungen zum Einsatz (z. B. Skalen von 1–5).

Nur selten reflektieren diese Studien, ob eine hohe Reliabilität tatsächlich Ziel eines Reviewverfahrens sein sollte. Hier lassen sich zwei konträre Positionen differenzieren: Einerseits kann man der von uns eingangs angeführten Idee folgen, dass ein geeignetes Instrument zum Erkennen qualitativ hochwertiger Studien auch zuverlässig sein muss, also bei wiederholtem Einsatz zu dem gleichen Ergebnis kommen sollte. Demnach sollte die Reliabilität des Peer-Review maximiert werden und Gutachter*innen bestenfalls zu den gleichen oder sehr ähnlichen Ergebnissen kommen (Cicchetti, 1991). Dies ist jedoch nur zu erwarten, wenn es eindeutige Kriterien gibt, die systematisch angewendet werden können.

Andererseits ließe sich argumentieren, dass Peer-Review-Verfahren einer anderen Logik folgen als klassische Inhaltsanalysen und zu viel Übereinstimmung gar nicht gewünscht wird (Bailar, 1991). Kontroverse Einschätzungen können nämlich hilfreiche Indikatoren sein, um Hinweise auf polarisierende Themen, Ansätze oder Methoden aufzuspüren. Solange keine einheitlichen und allgemein akzeptierten Kriterien existieren, ist Pluralismus zu erwarten. Hohe Übereinstimmung könnte eben auch aus einem Mangel an Perspektivenvielfalt resultieren. Hinzu kommt, dass bei extrem hoher Übereinstimmung auch Ressourcen verschwendet würden – überspitzt formuliert: Würden regelmäßig z. B. alle drei Gutachter*innen zu völlig identischen Ergebnissen kommen, wäre es überflüssig, drei Einschätzungen einzuholen. Ein einzelnes Gutachten würde ja zum selben Ergebnis führen. Erst durch die Auswahl sehr unterschiedlicher Gutachter*innen würde Perspektivenvielfalt ermöglicht und eben zwangsläufig auch zu Nichtübereinstimmung führen (Fiske & Fogg, 1990). Wir werden diese Diskussion im Fazit nochmal aufgreifen.

Die bislang durchgeführten Studien kommen zu recht unterschiedlichen Befunden, was die Reliabilität betrifft: So gibt es einerseits Untersuchungen, die zufriedenstellende Übereinstimmungswerte zwischen zwei oder mehr Gutachter*innen zeigen. Vecchio (2006) demonstriert beispielsweise anhand von 853 begutachteten Manuskripten, dass es in 38,9 Prozent der Fälle perfekte Übereinstimmung gab, in 82,5 Prozent der Fälle Abweichungen von einem Skalenpunkt oder weniger. Er schlussfolgert, dass es also nur in 17,5 Prozent der Fälle Dissens unter den Gutachter*innen gab. Auch eine der wenigen Studien im kommunikationswissenschaftlichen Bereich, die sich mit der Interrater-Reliabilität befasst hat, findet eine eher große Übereinstimmung zwischen Gutachter*innen: Neumann et al. (2008) werteten dafür Reviews des *Journal of Communication* aus und zeigen, dass in 69,4 Prozent der Fälle Einigkeit bezüglich der Entscheidung Annahme vs. Ablehnung besteht. Allerdings schränken die Autor*innen die Befunde auch dahingehend ein, dass bei einer Zeitschrift, die nur 16 Prozent der Einreichungen annimmt, Reviewer*innen generell stärker zur Ablehnung tendieren und dadurch schon eine entsprechende Verzerrung entsteht. Auch Cicchetti (1980) zeigt in seiner Studie eher hohe Übereinstimmungen zwischen den Urteilen von Gutachter*innen, wobei auch hier das Problem ist, dass häufiger Einreichungen abgelehnt werden und bei angenommenen Beiträgen oft Dissens zu finden ist. Und auch die Untersuchungen von Kemp (2005) oder Oxman et al. (1991) zeigen große Übereinstimmungen mit Reliabilitätswerten von über 0,5. Die angeführten Studien demonstrieren aber

auch, dass die Grenze zwischen einer hohen und einer niedrigen Reliabilität bei Reviews fließend ist und unterschiedlich interpretiert werden kann.

Die Mehrheit der Untersuchungen deckt eine eher geringe Übereinstimmung zwischen Gutachter*innen auf bzw. interpretiert die Daten so. Eine sehr häufig zitierte Studie ist das Experiment von Cole, Cole und Simon (1981): Sie ließen 150 bereits begutachtete Einreichungen für die *National Science Foundation* jeweils nochmal von einem neuen Set Reviewer*innen begutachten. Die Autor*innen zeigen, dass die hier getroffenen Entscheidungen zu 50 Prozent vom Zufall abhängen, also vom „luck of the reviewer draw", wie sie es nennen (S. 885). Peters und Ceci (1982) schickten zwölf Artikel, die kürzlich von einem psychologischen Top-Journal publiziert wurden, unter anderem Namen wieder an die gleiche Zeitschrift. Lediglich drei Aufsätze wurden als Wiedereinreichungen erkannt und zurückgewiesen. Von den übrigen neun Einreichungen wurden acht von den Herausgeber*innen abgelehnt, obwohl diese bereits unter anderem Namen in der eigenen Zeitschrift publiziert waren. 16 der 18 Gutachter*innen in der Studie plädierten für eine Ablehnung des Artikels. Viele Untersuchungen errechnen Korrelationen oder spezifische Reliabilitätskoeffizienten zwischen den Urteilen. So ermitteln Bornmann und Daniel (2008) beispielsweise Kappa Koeffizienten zwischen 0,10 und 0,21 für Gutachten bei der Zeitschrift *Angewandte Chemie*. Scott (1974) zeigt, dass verschiedene Dimensionen, die von Gutachter*innen der Zeitschrift *Journal of Personality and Social Psychology*, eingeschätzt wurden, nur im Bereich zwischen 0,07 und 0,37 korrelieren. Auch Bakanic, McPhail und Simon (1987), Callaham, Baxt, Waeckerie und Wears (1998) oder Blackburn und Hakel (2006) ermitteln in ihren Studien sehr geringe Interrater-Reliabilitäten. Übersichten über diese Studien finden sich beispielsweise bei Bornmann, Mutz und Daniel (2010) oder Lindsey (1988).

Es entstanden auch mehrere Metaanalysen zu Reviewer*innen-Reliabilitäten. Goldman (1994) bezieht in einer Metaanalyse 21 Datensätze aus 13 Studien ein und ermittelte ein Cohens Kappa von 0,31, was er als bedenklich einstuft. In einer weiteren Meta-Analyse beziehen Bornmann, Mutz und Daniel (2010) insgesamt 48 Studien mit 70 Koeffizienten ein und ermitteln noch geringere Korrelationen (r = 0,34) bzw. Übereinstimmungswerte (Cohens Kappa = 0,17). Ob sich derart niedrige Werte auch in kommunikationswissenschaftlichen Peer-Review-Verfahren zeigen, ist unklar. Außer der Studie von Neumann et al. (2008), die jedoch nur die prozentuale Übereinstimmung bei Annahme bzw. Ablehnung bei einer Zeitschrift untersuchen, gibt es keine Untersuchungen, die dies in Augenschein nehmen. Letztlich soll die Untersuchung Fallstricke und Probleme des Peer-Review-Verfahrens identifizieren, um zukünftige Begutachtungsverfahren verbessern zu können. Entsprechend verfolgen wir zwei forschungsleitende Fragen:

1. Stimmen Reviewer*innen im Fach zuverlässig überein, welche Tagungsbeiträge angenommen bzw. abgelehnt werden sollte und wie häufig kommen konträre Bewertungen vor?
2. Unterscheiden sich die Übereinstimmungsmaße zwischen verschiedenen Bewertungskriterien (z. B. Methode, Theorie, Darstellung, etc.)?

4. Maßzahlen zur Bestimmung der Reliabilität

Bisherige Studien greifen meist auf eine von drei Maßzahlen zurück: Sie berechnen Cohens Kappa, die intra-class correlation [ICC] und einfache Pearson Produkt-Moment-Korrelationen (*r*). Die *Produkt-Moment-Korrelation* ist als Übereinstimmungsmaß gänzlich ungeeignet (Krippendorff, 2004), da sie den eigentlichen Wert nicht berücksichtigt, sondern nur die Variation um den jeweiligen Erwartungswert: Wertet eine Reviewerin fünf Beiträge mit 3, 2, 2, 1 und 3 Punkten auf einer Fünferskala und ein anderer Gutachter vergibt 5, 4, 4, 3 und 5 Punkte, ist die Korrelation 1,0 – die Urteile stimmen aber in keinem Fall überein, sondern liegen sogar um jeweils 2 Skalenpunkte auseinander. *Cohens Kappa* ist als Reliabilitätskoeffizient sehr verbreitet, gerade in Medizin und Psychologie. Allerdings haben verschiedene Forscher*innen eine Reihe von Schwachpunkten aufgedeckt, die zu paradoxen Ergebnissen bei Cohens Kappa führen können. Der hauptsächliche Grund ist, dass Cohens Kappa die individuellen Randverteilungen einzelner Reviewer*innen (d. h. die individuellen Bewertungsmuster der einzelnen Reviewer*innen) verwendet, um erwartete Übereinstimmungen zu berechnen. Verschiebungen in den Randverteilungen verändern dann den Reliabilitätswert, ohne dass dies Verbesserungen oder Verschlechterungen der Reliabilität anzeigt. Es ist so sogar möglich, dass eine höhere Zahl von Übereinstimmungen den Wert senkt statt erhöht (Feinstein & Cicchetti, 1990; Krippendorff, 2004; Warrens, 2010). Das Problem wird behoben, indem eine verwandte Koeffizientenfamilie verwendet wird, die nicht individuellen Randverteilungen pro Reviewer*in, sondern eine gemeinsame Randverteilung für die Rater/Codierer einsetzt. Das beseitigt die Paradoxa ohne bekannte Nachteile zu verursachen. Die *Intra-Klassen-Korrelationen*, Scotts Pi, Fleiss' Kappa und Krippendorffs Alpha nutzen dieses Prinzip. Krippendorffs Alpha ist eine leicht angepasste Verallgemeinerung der anderen Koeffizienten, ist auf verschiedenen Datenniveaus verfügbar, sehr detailliert beschrieben, wird häufig eingesetzt und es ist relativ viel über seine Verteilung bekannt. Deshalb verwenden wir hier die ordinal-skalierte Variante von Krippendorffs Alpha (Krippendorff, 2004).

Allerdings ist auch Krippendorffs Alpha neuerdings häufiger Gegenstand von Kritik geworden. Sie entzündet sich vor allem an dem Phänomen, dass selbst extrem hohe prozentuale Übereinstimmung in sehr geringen oder sogar negativen Krippendorffs Alpha-Werten resultieren kann (z. B. Zhao, Liu, & Deng, 2012; Feng, 2013). Das gleiche trifft auf die oben genannten alternativen Koeffizienten wie ICC, Fleiss' Kappa, Scotts Pi und auch Cohens Kappa ebenfalls zu (Feinstein & Cicchetti, 1990). Es handelt sich hier allerdings nicht um ein Paradoxon (Krippendorff, 2012), sondern folgt aus extrem ungleichen Randverteilungen. Ein Beispiel: zwei Rater vergeben fast nur die Note 3 von 5. In einem Fall vergibt Rater 1 jedoch eine 5 und Rater 2 eine 4. Krippendorffs Alpha geht nun davon aus, dass ohnehin beide Rater „wissen", dass fast nur 3er vorkommen werden. Sie vergeben deshalb immer 3er; es könnten „echte" Übereinstimmungen, aber auch zufällige Übereinstimmungen sein. Konservativ geschätzt muss man annehmen, dass es beinahe alles zufällige Übereinstimmungen sind. Die wenigen abweichenden Fälle sind hingegen interessant, weil bei ihnen eine zufällige Übereinstimmung sehr un-

wahrscheinlich ist. Deshalb basiert der schlussendliche Koeffizient fast ausschließlich auf den vom Erwartbaren „abweichenden" Wertungen. Stimmen diese nicht überein, folgt ein Wert nahe bei null oder sogar im negativen Bereich, was heißt: schlechter als Zufall.

Hieran ist erneut die Berechnung der erwarteten Übereinstimmung anhand der empirischen Randverteilung schuld. Es wird als „worst case" davon ausgegangen (wenn auch nur als rechnerische Approximation, Krippendorff, 2012), dass beide Reviewer*innen die Häufigkeit der Ausprägungen, die sie selbst erst durch ihre Urteile erzeugen, bereits im Voraus kennen und diese heranziehen, um ein Urteil zu fällen, wenn sie die Merkmale, die sie bewerten sollen, nicht erkennen. Sie wissen viel über die Verteilung der Merkmale, und codieren so, wie ein Statistiker es ihnen raten würde, um eine möglichst hohe Übereinstimmung zu erzielen. Die Erkenntnis, welche Merkmale häufig und welche selten vorkommen, ist ja möglicherweise ein entscheidendes Ergebnis der Codierung, sie wird hier aber als Basislinie gewählt, um Zufallsübereinstimmungen quasi auszuschließen (Zhao, Liu, & Deng, 2012). Das heißt auch, um höhere Alpha-Werte zu erreichen, sollte das Material so heterogen wie möglich sein, damit alle Arten von Übereinstimmungen mit ähnlich hoher Gewichtung in das Endergebnis einfließen (Krippendorff, 2012).

Es gibt jedoch weitere Verfahren, Zufallsübereinstimmungen zu schätzen. Sehr einfach und intuitiv (und dadurch auch weniger fehleranfällig) ist die Überlegung, dass a priori alle Ausprägungen die gleiche Chance haben, gewählt zu werden: Ohne genauere Kenntnisse über einen Beitrag würde der/die Reviewer*in willkürlich die 1, die 2, die 3, die 4 oder die 5 ankreuzen. Dann wäre also die Wahrscheinlichkeit für eine Zufallsübereinstimmung bei allen Skalenpunkten gleich, nämlich 1/5. Diese Idee wurde mehrfach wohl unabhängig voneinander publiziert, so dass der gleiche Koeffizient unter mehreren verschiedenen Bezeichnungen kursiert, z. B. Bennett et al.'s S, Guilfords G, Maxwells RE und Brennan und Predigers Kappa (Krippendorff, 2004; Zhao, Liu, & Deng, 2012). Gwet hat eine Rechenvariante für Brennan und Predigers Kappa entwickelt, die es erlaubt, die Methode auch auf ordinal- und metrisch skalierte Daten anzuwenden (Gwet, 2014). Wir ergänzen also Krippendorffs Alpha mit Brennan und Predigers Kappa, jeweils nach Gwets (2014) Programmierung für ordinales Datenniveau.

5. Methodisches Vorgehen

Um zu analysieren, wie hoch die Interrater-Reliabilitäten auf DGPuK-Tagungen sind und welche Unterschiede es zwischen verschiedenen Kategorien gibt, analysieren wir die Bewertungen von Tagungseinreichungen. Dabei fokussieren wir DGPuK-Jahrestagungen und Tagungen der größten Fachgruppen. Die Fokussierung auf große Fachgruppen erfolgte aus forschungspragmatischen und statistischen Gründen: Die Vielzahl an Fach- und Ad-Hoc-Gruppen (zum Zeitpunkt der Studiendurchführung 19) erforderte eine Selektion, die Auswahl der großen Fachgruppen folgte dabei der Idee, so die für die Mitglieder relevantesten Gruppen abzudecken und zugleich eine große Anzahl an Einreichungen untersuchen zu können. Die größeren Fachgruppen ermöglichen auch einen direkten Vergleich der Fachgruppen untereinander, wohingegen bei kleineren Fachgruppen mit weni-

ger Einreichungen und Reviews die Fallzahlen pro Fachgruppe leicht zu klein werden können. Zum Zeitpunkt der Datenerhebung waren die fünf größten Fachgruppen die „Rezeptions- und Wirkungsforschung“ (407 Mitglieder), „Kommunikation und Politik“ (362 Mitglieder), „Methoden der Publizistik- und Kommunikationswissenschaft“ (ebenfalls 362 Mitglieder), „Journalistik/Journalismusforschung“ (326 Mitglieder) sowie „Digitale Kommunikation“ (314 Mitglieder). Die Fokussierung auf diese fünf Gruppen bot sich auch deshalb an, weil zwischen der fünft- und sechstgrößten Gruppe eine deutliche Größendifferenz von 100 Mitgliedern besteht.

Zwischen Ende November 2016 und Januar 2017 baten wir die Fachgruppensprecher und Tagungsorganisatoren, die Daten der Begutachtung anonymisiert zur Verfügung zu stellen. Dieser Schritt wurde im Vorfeld dem Vorstand der DGPuK angekündigt, welcher seine Zustimmung zum Vorgehen gab. Wir erhielten vollständig anonymisierte Datensätze ohne Beitragstitel sowie ohne Namen der Reviewer*innen- und Autor*innen. Wir erbaten dabei die Ergebnisse der jeweils letzten fünf Jahrestagungen, die von den entsprechenden Fachgruppen durchgeführt wurden.

Wir erhielten jeweils von allen Fachgruppen die Ergebnisse der Reviewverfahren der letzten fünf Tagungen. Lediglich die Daten einer Fachgruppentagung waren nicht mehr auffindbar. Im Datensatz ist zudem zu beachten, dass aufgrund einer gemeinsamen Tagung der Fachgruppe „Kommunikation und Politik“ und „Digitale Kommunikation“ ein Datensatz weniger existiert. Insgesamt gehen so 23 Fachgruppen-Reviewverfahren in unsere Auswertung ein. Bei den DGPuK-Jahrestagungen war einer der Datensätze nicht mehr gespeichert, einer der Ausrichter reagierte trotz mehrmaliger Nachfassaktion nicht auf unsere Anfrage, weshalb hier drei Datensätze in die Auswertung eingehen. Insgesamt lagen 21.995 einzelne Bewertungen vor, die sich in 8.486 Bewertungspaare/-tripel zu 1.409 Papern gliedern und aus 3.537 Reviews stammen (15,61 Bewertungen pro Paper, 6,22 Wertungen pro Review, 2,57 Reviews pro Paper). Von den 8.486 Bewertungspaarungen entfielen auf die DGPuK-Jahrestagung 4.131 und 4.355 auf die Fachgruppentagungen; Details sind Tabelle 1 zu entnehmen. Zu beachten ist, dass lediglich drei DGPuK-Jahrestagungen schon fast die Hälfte der Reviews ausmachen. 1.311 Bewertungspaare sind dabei lediglich die Durchschnittswertungen über verschiedene Kategorien hinweg, sie ergeben sich also aus den anderen Bewertungen.

Tabelle 1. Datengrundlage (Bewertungspaarungen)

	Jahr						gesamt
	2012	2013	2014	2015	2016	2017	
DGPuK	---	n. v.	n. v.	1120	1925	1086	4131
Fachgruppen, davon:	231	777	539	1062	832	914	4355
Methoden	---	94	133	60	114	216	617
RezFo	---	192	167	205	405	304	1273
KomPol	---	m. CvK	n. v.	307	204	252	763
Journ	---	115	92	450	82	142	881
DigiKom, CvK	231	376	147	40	27	---	821
Gesamt	231	777	539	2182	2757	2000	8486

Anmerkungen. n. v.: nicht verfügbar; ---: nicht angefragt; m. CvK: zusammen mit der Fachgruppe Computervermittelte Kommunikation ausgerichtet, welche das Reviewverfahren betreute.

Die unterschiedlichen Datenformate wurden vereinheitlicht, die Review-Kriterien wurden fachgruppenübergreifend verschlüsselt. Folgende Kriterien wurden unterschieden: (1) Darstellung/Verständlichkeit, (2) Methode/Vorgehensweise, (3) Originalität/Innovation, (4) Passung mit Tagungsthema, (5) Relevanz/Beitrag zum Forschungsfeld, (6) Theoretische Fundierung/Aufarbeitung des Forschungsstands, (7) Durchschnitt (errechnete Gesamtwertung) und (8) Gesamtwertung (unabhängige Bewertung). Die jeweilige Formulierung oder Beschreibung der Kriterien variiert aber je nach Fachgruppe bzw. ändert sich innerhalb einer Fachgruppe über die Zeit, was aber nicht berücksichtigt werden konnte. Auch die verschiedenen Bewertungsstufen, die entweder von 0–5, 1–5 oder 1–10 reichten, wurden auf die Skala von 1 bis 5 vereinheitlicht. Bei der 0-5-Skala wurde die 0 beinahe nie vergeben, so dass die 0 in 1 umcodiert wurde und die Skala dann als 1-5-Skala behandelt wurde.

Es wurden verschiedene Indikatoren für die Übereinstimmung der Reviewer*innen berechnet. Die prozentuale Übereinstimmung zeigt, welcher Anteil der paarweisen Bewertungen desselben Papers hinsichtlich desselben Kriteriums übereinstimmt. Vergaben drei Reviewer*innen beim selben Paper 4, 3 und 3 Punkte, gab es drei paarweise Vergleiche (Reviewer*in 1 und 2, Reviewer*in 1 und 3, Reviewer*in 2 und 3), wobei die Werte zweimal nicht übereinstimmen und einmal übereinstimmen; die Übereinstimmung wäre demnach 33%. *Krippendorffs* α zeigt an, wie stark die Codierungen übereinstimmen, gemessen an rein zufälliger Übereinstimmung, die bei der angefallenen Randverteilungen der Codierungen zu erwarten wäre. *Brennan und Predigers* α zeigt an, wie stark die Codierungen übereinstimmen, gemessen an zufälliger Übereinstimmung, wenn alle möglichen Skalenpunkte mit gleicher Häufigkeit gewählt werden würden. Bei beiden Maßen wurde eine ordinale Skalierung zugrunde gelegt und entsprechend die „Deutlichkeit" oder „Schwere" von Nichtübereinstimmungen berücksichtigt. Weiterhin wurde die Spannweite der Bewertungen in derselben Bewertungskategorie bei ein und demselben Paper berechnet; auch die Standardabweichung der Bewertungen des gleichen Papers in der gleichen Bewertungskategorie wurde ausgewertet.

6. Ergebnisse

6.1 Übereinstimmung und Spannweiten der Gesamtbewertung

Die Übereinstimmung zwischen Reviewer*innen ist – gemessen an den für Inhaltsanalysen empfohlenen Mindestwerten – gering. Krippendorffs α liegt insgesamt bei 0,276 – zur Erinnerung: 1 bedeutet: perfekte Übereinstimmung, 0 bedeutet zufällige Übereinstimmung und -1 bedeutet: perfekte Nichtübereinstimmung. Krippendorff empfiehlt für Inhaltsanalysen Werte größer 0,8 und rät davon ab, Codierungen mit $\alpha < 0{,}667$ zu verwenden. Bei vergleichbaren Maßen stufen Landis und Koch (1977) Werte von 0,8–1,0 als sehr gut, 0,6–0,8 als gut, 0,4–0,6 als befriedigend, 0,2–0,4 als ausreichend und 0–0,2 als schwach ein. Selbst bei diesen euphemistischen Bezeichnungen kommen die Reviewer*innen gerade einmal in den Bereich „ausreichend". Brennan und Predigers Kappa zeigt erwartungsgemäß einen günstigeren Wert an, nämlich 0,649. Das genügt immer noch nicht Krippendorffs Mindestanforderungen, wäre nach Landis und Koch hingegen „gut".

Die Spannweiten der Bewertungen illustrieren gut das Ausmaß der Einigkeit bzw. Uneinigkeit: Über alle Wertungen hinweg lag die Spannweite der Urteile bei 27 Prozent der Beiträge bei 0 (wir haben für diese Auswertung immer abgerundet, ansonsten die exakten Werte verwendet), d. h. die (üblicherweise 2 oder 3) Reviewer*innen waren sich absolut einig. Nachdem die zufällige Übereinstimmung allerdings schon bei 20 Prozent liegt, ist dieser Wert erstaunlich gering. 41 Prozent der Bewertungen zeigten eine Abweichung von einem Skalenpunkt. Abweichungen von 2 Skalenpunkten kamen in 22 Prozent der Bewertungen vor, 3 Skalenpunkte immerhin in 7 Prozent der Bewertungen. Die maximale Abweichung von 4 Skalenpunkten trat in etwa 3 von 100 Bewertungen auf. Insgesamt gibt es in 31 Prozent der Bewertungen also Abweichungen von 2 oder mehr Skalenpunkten auf einer Skala von 1 bis 5, was erhebliche Uneinigkeit zwischen den Gutachter*innen in beinahe einem Drittel der Bewertungspaare bedeutet. Auch der mittlere Abstand zwischen bester und schlechtester Bewertung (Spannweite) auf derselben Dimension von 1,25 ist relativ groß. Die Standardabweichung liegt bei 0,74. Die durchschnittliche Höchstwertung liegt bei 4,14, die durchschnittliche Niedrigstwertung bei 2,89 und der Durchschnitt aller Wertungen liegt bei 3,54 (Tabelle 1).

Tabelle 2. Verteilung der Spannweiten

Werte	Verteilung (n = 21.995) %	Minimum (n = 8.486) %	Maximum (n = 8.486) %	Intervall	Spannweiten (n = 8.458) %	Standardabweichung (n = 8.458) %
1	5	11	1	[0–1)	27	67
2	13	25	4	[1–2)	41	28
3	26	35	16	[2–3)	22	5
4	34	23	41	[3–4)	7	0
5	21	6	39	[4–5]	3	0
MW	3,54	2,89	4,14	*MW*	1,25	0,74

Anmerkungen. Die Fallzahl von Spannweite und Standardabweichungen weicht von der Gesamtzahl der Bewertungspaare/-tripel (n = 8.468) ab, weil bei einzelnen Kriterien eine Bewertung vorlag, etwa wenn ein/eine Reviewer*in keine Bewertung abgab. MW: Mittelwert; (X–Y) bedeutet die Wertespanne ab X bis Y, X und Y nicht eingeschlossen; [X–Y] ist die Wertespanne zwischen X und Y, X und Y eingeschlossen.

6.2 Übereinstimmung hinsichtlich verschiedener Reviewkategorien

Die Übereinstimmung zwischen den Reviewer*innen bleibt auch bei den einzelnen Kriterien gering. Besonders wenn man Krippendorffs Alpha heranzieht, sind die Werte durchgängig sehr niedrig, zwischen 0,165 und 0,293 (vgl. nachfolgend Tabelle 2). Am schwächsten fällt die Reliabilität der Urteile über Originalität (0,165) und Relevanz (0,173) aus. Letztere liegen offenbar besonders stark im Auge des Betrachters. Im Mittelfeld liegen Passung (0,207), Theorie (0,227), Methode (0,229), die (gesonderte) Gesamtbewertung (0,237) sowie die Darstellung (0,246). Am besten ist die Reliabilität der Durchschnittswerte (0,293) und wenn man alle Bewertungen zusammengenommen betrachtet (0,276). Hier deutet sich an, dass sich die Bildung eines Bewertungsmittelwerts oder eines Summenscores positiv auf die Reliabilität der Paperbewertungen auswirkt, wenn auch auf niedrigem Gesamtniveau.

Die Befunde fallen anders aus, wenn man Brennan und Predigers Kappa zugrunde legt, also die Zahl der Skalenpunkte statt der Randverteilung der beobachteten Kodierungen zur Grundlage der Schätzung macht, wie viele zufällige Übereinstimmungen zu erwarten wären. Hier schneidet von den Einzelkriterien die Passung am schlechtesten ab (0,463), die Relevanz weist die größte Reliabilität auf (0,581) und alle anderen bewegen sich zwischen 0,506 und 0,525. Die (gesonderte) Gesamtwertung liegt mit diesen etwa gleichauf (0,529). Mit Abstand am besten fällt die Reliabilität aus, wenn man alle Kriterien gemeinsam betrachtet (0,649) oder – noch besser – wenn man den Durchschnitt der Einzelbewertungen errechnet (0,757). Dies ist auch der einzige Wert, der zumindest die Minimalanforderung von 0,667 erfüllt (welche zumindest für Krippendorffs Alpha gelten soll).

Das Ausmaß der Nichtübereinstimmung lässt sich auch mit der durchschnittlichen Spannweite pro Beitrag und der durchschnittlichen Standardabweichung pro Beitrag beziffern. Die Spannweite R zwischen den Urteilen der Reviewer*innen über dieselbe Einreichung liegt bei allen Kriterien zwischen 1,20 und 1,41 Punkten

auseinander. Lediglich bei der Durchschnittswertung (R = 1,01; s = 0,58) und der Gesamtwertung (R = 1,04; s = 0,69) klaffen die Urteile weniger stark auseinander. Da aber Gesamtwertungen nur bei wenigen Tagungen verlangt wurden, ist dieser Wert mit Vorsicht zu interpretieren. Zwischen den Einzelkriterien Passung, Originalität, Relevanz, Theorie, Methode und Darstellung gibt es nur geringe Unterschiede im Ausmaß der Abweichungen.

Tabelle 3. Reviewer*innen-Übereinstimmung nach Review-Kriterien

	Passung	Originalität	Relevanz	Theorie	Methode	Darstellung	Durchschnitt	Gesamt	Alle Wertungen
Krippendorffs α (SE)	,207 (,028)	,165 (,028)	,173 (,019)	,227 (,019)	,229 (,018)	,246 (,019)	,293 (,020)	,237 (,055)	,276 (,009)
Brennan & Predigers κ (SE)	,463 (,024)	,524 (,020)	,581 (,011)	,525 (,013)	,506 (,014)	,522 (,013)	,757[a)] (,007)	,529[a)] (,034)	,649[a)] (,009)
Spannweite R	1,41	1,20	1,26	1,31	1,35	1,29	1,01	1,04	1,25
Standard-abweichung s	,84	,71	,75	,78	,79	,76	,58	,69	,74
Arithm, Mittel	3,50	3,52	3,49	3,50	3,50	3,50	3,49	3,48	3,50
Fallzahl	904	815	1342	1265	1306	1265	1311	278	8486

Anmerkungen. [a)] Ungerade Werte wurden echt auf Ganzzahlen gerundet, damit bei allen Brennan und Predigers Kappa-Berechnungen dieselbe Zufallsübereinstimmungsschätzung zugrunde liegt und keine Inflation von selten besetzten Ausprägungen das Ergebnis verfälscht (Krippendorff, 2004).

6.3 Ursachen von Konsens und Dissens

Um zu verstehen, wie Dissens zwischen den Urteilen zustande kommt, wurden lineare mixed models berechnet, in denen Tagungen und Paper (geschachtelt in Tagungen, weil jedes Paper genau einer Tagung zugeordnet werden kann) als random effect (Zufallsvariable) behandelt wurden. Wir versuchen, das Ausmaß des Dissens bei einzelnen Beiträgen – angezeigt durch die Spannweite oder die Standardabweichung der Urteile[1] – auf mögliche Erklärungsfaktoren zurückzuführen. Als Prädiktoren werden (1) die angewandten Reviewkriterien (auf die sich das Urteil bezieht), (2) die reviewenden Fachgruppen, (3) das Jahr (vielleicht sind die Urteile ja seit 2012 reliabler oder weniger reliabel geworden), (4) die Anzahl der eingesetzten Reviewer*innen sowie (5) die „Qualität" des Beitrags (gemessen als Durchschnittsbewertung der Reviewer*innen) in Betracht gezogen. Zusammen mit der Qualität wird auch der verbleibende Abstand zwischen Durchschnittsurteil und dem oberen (5) bzw. unteren Ende (1) der Skala als Prädiktor hinzuge-

1 Inter-Reviewer-Reliabilitätsindizes basieren auf der Codierung mehrerer Einheiten und sind deshalb nur für Aggregatdatenanalysen geeignet.

fügt, weil aufgrund der begrenzten Skalen zu erwarten ist, dass die Übereinstimmung bei sehr negativ und sehr positiv bewerteten Beiträgen größer ist als bei mäßig bewerteten Beiträgen – bei denen auf der Skala noch Platz nach unten und oben ist. In den Modellen zeigen positive Koeffizienten an, dass die Streuung größer wird, wenn der Prädiktor vorhanden ist bzw. in seinem Wert steigt. Negative Koeffizienten zeigen an, dass der Prädiktor den Konsens zwischen Urteilen steigert (bzw. den Dissens verringert).

Die Modelle berücksichtigen zuerst nur die Bewertungskriterien und nehmen dann nacheinander Fachgruppe, Jahr, Anzahl der Reviewer*innen und Durchschnittsurteil/Nähe des Durchschnittsurteils zum nächstgelegenen Skalenende hinzu. Die Analysen zeigen erst im finalen Modell ein vollständiges Bild.

Reviewkriterien. Konsistent zeigt sich, dass die Durchschnittsbildung über Reviewkriterien hinweg die Übereinstimmung zwischen Reviewer*innen erhöht, wohingegen die Passung mit dem Tagungsthema eine Quelle für vermehrten Dissens ist. Die Passung scheint stärker als andere Merkmale im Auge des Betrachters zu liegen. Insgesamt sind die Reviewkriterien ein wichtiger Prädiktor der Spannweite; das Modell wird signifikant besser und R^2_{marginal} zeigt an, dass die (durch fixed effects) erklärte Varianz um 1,4%-Punkte zunimmt.

Fachgruppen. Die Unterschiede zwischen den Fachgruppen schwanken je nachdem, welche Kontrollvariablen mit einbezogen werden. So ist der relativ hohe Konsens in den FG Kommunikation und Politik, Digitale Kommunikation und Journalistik/Journalismusforschung hingegen auf die relativ geringe Zahl von Reviewer*innen pro Beitrag zurückzuführen. Im finalen Modell ist der Konsens leicht verringert in der FG Journalistik/Journalismusforschung. Dass Konsens aber nicht unbedingt ein Gütekriterium sein muss, ist ein Thema für unsere Diskussion der Befunde. Die Fachgruppen tragen zwar zur 2,3% zur Erklärung der Spannweite bei, diese entpuppen sich aber bei genauerer Analyse als Effekt der Anzahl der Reviewer*innen pro Einreichung, die zwischen den Fachgruppen deutlich variiert: im Mittel sind es 3,00 bei Rezeptions- und Wirkungsforschung, 2,72 bei Methoden, 2,68 bei der DGPuK, 2,41 bei Kommunikation und Politik, 2,04 bei Digitale Kommunikation/CvK und 2,00 bei Journalistik/Journalismusforschung; im Durchschnitt waren es 2,57 Reviews pro Beitrag.

Zeit. Es gab keine lineare Entwicklung der Reviewer*innen-Übereinstimmung im Laufe der letzten fünf Jahre. In dem relativ kurzen betrachteten Zeitfenster wäre eine signifikante Zunahme des Konsenses auch überraschend gewesen.

Anzahl der Reviewer*innen. Die Anzahl der Reviewer*innen (entweder 2 [43% der Beiträge] oder 3 [57% der Beiträge] pro Beitrag) erhöht eindeutig den Dissens zwischen den Reviewer*innen. Die Spannweite steigt je nach Modell mit der/dem dritten Reviewer*in um 0,803 bzw. 0,757 Punkte. Spannweiten und Standardabweichungen reagieren allerdings sensibel auf die Fallzahl und diese Zusammenhänge sind möglicherweise für den beobachteten Effekt verantwortlich. Zur Erinnerung: Die einzelnen Spannweiten und Standardabweichungen pro Bewertungspaarung beruhen auf zwei bzw. auf drei Fällen. Wir testen dies, indem wir 100.000 2er- und 3er-Bootstrap-Samples aus der Menge aller Urteile zufällig ziehen und Standardabweichungen bzw. Spannweiten bei Samples mit Stichprobengröße 2 und 3 vergleichen. Tatsächlich gibt es erhebliche Unterschiede: Bei

den 3er-Stichproben ist die Spannweite um 0,611 Punkte, die Standardabweichung um 0,118 Punkte höher als in den Zweierstichproben. Die Standardfehler dieser Differenzen liegen aufgrund des riesigen Bootstrap-Samples quasi bei 0. Ziehen wir diese Werte von den Koeffizienten in Tabelle 4 (Spannweite) bzw. 5 (Standardabweichung) ab, verringert sich der Wert erheblich (gerade bei der Spannweite), es gibt aber immer noch signifikante Effekte der Anzahl der Reviewer*innen, die nicht aus den Maßzahlen resultieren.

Die Zahl der Reviewer*innen (als grober Indikator der Perspektivenvielfalt) ist der stärkste Prädiktor der Spannweite in dieser Studie. Die erklärte Varianz steigt um 9,3 Prozentpunkte auf 13,1 Prozent. Dass das conditional R^2 nicht in gleichem Maße steigt ist darauf zurückzuführen, dass die Unterschiede in der Zahl der Reviewer*innen vor allem von den Tagungen und den Einreichungen abhängen – so bestimmen z. B. die Fachgruppensprecher und/oder Konferenzorganisatoren maßgeblich mit, wie viele Reviewer*innen jeweils dasselbe Paper begutachten, so dass ein großer Teil der Varianz bereits durch random effects auf der Ebene der Tagungen erklärt wird. Noch präziser ist die Vorhersage, wenn man random intercepts (d. h. der als Zufallsverteilung aufgefassten individuellen Achsenabschnitte der Spannweite) der einzelnen Einreichungen und ihre jeweilige Zahl von Reviewer*innen betrachtet. Schließlich können Reviewausfälle dazu führen, dass einzelne Einreichungen von weniger Reviewer*innen begutachtet werden als angestrebt.

Qualität der Beiträge. Die Qualität der Beiträge beeinflusst, wie einig oder uneinig sich die Gutachter*innen bei einem Beitrag sind – allerdings ist der Zusammenhang kurvilinear. Rein linear betrachtet sinkt mit steigender Beitragsqualität der Dissens; der Dissens ist allerdings in der Mitte der Skala am größten (Skalenpunkt 2–4; dort ist noch Platz für Abweichungen nach oben und unten), wohingegen die Spannweite bei einer Durchschnittsbewertung über 4 deutlich sinkt (hier sind ja keine „6en" und „7en" möglich); Durchschnittsbewertungen von 2 und darunter zeigen ebenfalls deutlich größere Übereinstimmung; Durchschnittsbewertungen unter 1,5 kommen empirisch nicht vor. Die Abflachung der Spannweite an den Rändern kann zwar auf Decken- und Bodeneffekte (censoring) der Skalierung zurückzuführen sein; aber auch die Deutung, dass besonders gute und besonders schlechte Beiträge eher gemeinschaftlich als solche erkannt werden und daher einheitlichere Urteile resultieren, ist möglich. Daher wurde die Gefahr eines Decken- oder Bodeneffekts als eigener Prädiktor ins Modell eingefügt: Wie viele Skalenpunkte sind unterhalb bzw. oberhalb (der kleinere der beiden Werte) der mittleren Bewertung des Beitrags noch übrig? In diesem Modell zeigt sich, dass mit steigender mittlerer Wertung die Spannweite sinkt, aber gleichzeitig die größere Chance auf einen Decken- oder Bodeneffekt die Spannweite senkt. Beide Einflüsse sind statistisch signifikant. Das gilt sowohl, wenn man beide Werte als Rohwerte in das Modell einfügt (hier nicht ausgewiesen) als auch wenn man nur die Residuen der „übrigen Skalenpunkte" nach einer Regression auf die mittlere Bewertung einfügt (Tabelle 4). Letzteres dient dazu, die beiden rechnerisch aufeinander aufbauenden und stark korrelierten Prädiktoren voneinander unabhängig zu machen. Die Qualität der Beiträge – operationalisiert durch die Durchschnittsbewertung der betei-

ligten Reviewer*innen – ist ein wichtiger, aber nicht der wichtigste Prädiktor der Spannweite; das marginal R^2 steigt um 1,9 Prozentpunkte auf 15,0 Prozent.

Die Standardabweichung der Urteile wurde als alternatives Streuungsmaß ebenfalls untersucht und es zeigen sich wie erwartet die gleichen Ergebnisse: Reviewkriterien, Perspektivenvielfalt und Paper-Qualität sind starke Prädiktoren der Standardabweichung; Fachgruppen und Zeit haben nahezu keinen Einfluss. Die Abweichung der Urteile ist beim Kriterium „thematische Passung" besonders groß (zusätzlich auch beim Kriterium „Methoden") und nach Durchschnittsbildung über verschiedene Urteile hinweg am geringsten. Mehr Reviewer*innen (= mehr verschiedene Perspektiven) führen zu mehr Uneinigkeit, wohingegen Decken- und Bodeneffekte sowie hohe Paperqualität (positives Durchschnittsurteil) die Uneinigkeit verringern (Tabelle 5).

Tabelle 4. Einflussfaktoren auf die Spannweite der Urteile

	Abhängige Variable: Spannweite (n = 8277)					
	Leer *B* (*SE*)	Kriterien *B* (*SE*)	FG *B* (*SE*)	Zeit *B* (*SE*)	Perspektiven *B* (*SE*)	Qualität *B* (*SE*)
Konstante	1,232***	1,197***	1,266***	1,493***	-0,668**	0,201*
	(0,050)	(0,078)	(0,109)	(0,182)	(0,228)	(0,215)
Reviewkriterien (Vergleichskategorie: Gesamt)						
Darstellung		0,076	0,063	0,063	0,061	0,061
		(0,066)	(0,066)	(0,066)	(0,066)	(0,066)
Durchschnitt		-0,212**	-0,226***	-0,226***	-0,227***	-0,227***
		(0,066)	(0,066)	(0,066)	(0,066)	(0,066)
Methode		0,134*	0,120†	0,120†	0,119†	0,119
		(0,066)	(0,066)	(0,066)	(0,066)	(0,066)
Originalität		-0,024	-0,039	-0,040	-0,039	-0,040
		(0,070)	(0,070)	(0,070)	(0,070)	(0,070)
Passung		0,151*	0,140*	0,140*	0,146*	0,144*
		(0,067)	(0,067)	(0,067)	(0,067)	(0,067)
Relevanz		0,037	0,023	0,023	0,021	0,022
		(0,066)	(0,066)	(0,066)	(0,066)	(0,066)
Theorie		0,100	0,086	0,086	0,084	0,085
		(0,065)	(0,066)	(0,066)	(0,066)	(0,066)
Fachgruppe (Vergleichsgruppe: DGPuK)						
KommPol			-0,290*	-0,291*	-0,084	-0,095
			(0,136)	(0,130)	(0,116)	(0,094)
Journalistik			-0,198	-0,244†	0,295*	0,260*
			(0,124)	(0,122)	(0,115)	(0,095)
DigiKom			-0,281†	-0,420*	0,071	0,057
			(0,138)	(0,159)	(0,146)	(0,121)

	Abhängige Variable: Spannweite (*n* = 8277)					
	Leer *B* (*SE*)	Kriterien *B* (*SE*)	FG *B* (*SE*)	Zeit *B* (*SE*)	Perspek-tiven *B* (*SE*)	Qualität *B* (*SE*)
Methoden			0,121	0,068	-0,162	0,154
			(0,133)	(0,132)	(0,119)	(0,097)
RezFo			0,169	0,128	-0,128	-0,125
			(0,119)	(0,116)	(0,104)	(0,084)
Zeit				-0,045	-0,042	-0,034
(Jahre seit 2011)				(0,030)	(0,027)	(0,022)
Perspektivenvielfalt					0,803***	0,757***
(Anzahl Reviewer*innen)					(0,058)	(0,054)
Qualität						-0,224***
(Mittlere Wertung)						(0,022)
Decken- und Bodeneffekte						0,277***
(Punkte bis Skalenende)						(0,044)
σ^2 (Paper)	0,199	0,202	0,202	0,202	0,165	0,140
σ^2 (Tagung)	0,048	0,048	0,023	0,020	0,015	0,009
σ^2 (Residual)	0,718	0,702	0,702	0,702	0,701	0,701
Marginal R^2	,000	,014	,037	,038	,131	,150
Conditional R^2	,256	,273	,270	,269	,309	,299
Deviance	22114	21954	21933	21930	21750	21614
Deviance change	---	160 ***	21 ***	3 †	180 ***	136 ***
AIC	22122	21976	21965	21964	21786	21654
BIC	22150	22053	22073	22083	21913	21794

† $p < 0{,}10$; $^{*}p < 0{,}05$; $^{**}p < 0{,}01$; $^{***}p < 0{,}001$
Anmerkung. 8.568 Bewertungspaare/-tripel (Level 0) in 1.364 Beiträgen (Level 1) bei 25 Konferenzen (Level 2)

Tabelle 5. Einflussfaktoren auf die Standardabweichung der Urteile

	Abhängige Variable: Standardabweichung *(n* = 8277*)*					
	Leer *B* (*SE*)	Kriterien *B* (*SE*)	FG *B* (*SE*)	Zeit *B* (*SE*)	Persp. *B* (*SE*)	Qualität *B* (*SE*)
Konstante	0,738***	0,709***	0,728***	0,855***	0,068	0,562***
	(0,020)	(0,040)	(0,055)	(0,088)	(0,133)	(0,126)
Reviewkriterien						
Darstellung		0,054	0,043	0,043	0,042	0,042
		(0,039)	(0,039)	(0,039)	(0,039)	(0,039)
Durchschnitt		-0,131***	-0,143***	-0,143***	-0,143***	-0,143***
		(0,039)	(0,039)	(0,039)	(0,039)	(0,039)
Methode		0,086**	0,074*	0,074*	0,074*	0,074*
		(0,039)	(0,039)	(0,039)	(0,039)	(0,039)
Originalität		-0,003	-0,016	-0,016	-0,015	-0,016
		(0,041)	(0,041)	(0,041)	(0,041)	(0,041)

	Abhängige Variable: Standardabweichung *(n = 8277)*					
	Leer *B* (*SE*)	Kriterien *B* (*SE*)	FG *B* (*SE*)	Zeit *B* (*SE*)	Persp. *B* (*SE*)	Qualität *B* (*SE*)
Passung		0,119***	0,109***	0,109***	0,110***	0,110***
		(0,040)	(0,040)	(0,040)	(0,040)	(0,040)
Relevanz		0,036	0,024	0,024	0,023	0,023
		(0,039)	(0,039)	(0,039)	(0,039)	(0,039)
Theorie		0,068*	0,056	0,056	0,056	0,056
		(0,039)	(0,039)	(0,039)	(0,039)	(0,039)
Fachgruppen						
KommPol			-0,129**	-0,130**	-0,054	-0,060
			(0,064)	(0,059)	(0,067)	(0,054)
Journalistik			0,021	-0,006	0,194***	0,174***
			(0,059)	(0,056)	(0,067)	(0,055)
DigiKom			-0,057	-0,135*	0,047	0,040
			(0,065)	(0,074)	(0,085)	(0,070)
Methoden			0,017	-0,012	-0,091	-0,087
			(0,064)	(0,061)	(0,069)	(0,056)
RezFo			0,044	0,022	-0,070	-0,069
			(0,055)	(0,052)	(0,060)	(0,048)
Zeit				-0,025*	-0,023	-0,018
(Jahre seit 2011)				(0,014)	(0,015)	(0,013)
Perspektivenvielfalt					0,290***	0,265***
(Anzahl Reviewer*innen)					(0,034)	(0,032)
Qualität						-0,128***
(Mittlere Wertung)						(0,013)
Decken- und Bodeneffekte						0,162***
(Punkte bis Skalenende)						(0,026)
σ^2 (Paper)	0,061	0,062	0,062	0,062	0,057	0,048
σ^2 (Tagung)	0,006	0,006	0,005	0,004	0,005	0,003
σ^2 (Residual)	0,254	0,247	0,247	0,247	0,247	0,247
Marginal R^2	,000	,018	,024	,026	,065	,087
Conditional R^2	,208	,229	,231	,230	,252	,244
Deviance	13362	13170	13159	13155	13086	12956
Deviance change		192 ***	10 †	4 *	69 ***	130 ***
AIC	13370	13192	13191	13189	13122	12996
BIC	13398	13269	13304	13308	13249	13136

† $p < 0{,}10$; * $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$

Anmerkung. 8.568 Bewertungspaare/-tripel (Level 0) in 1.364 Beiträgen (Level 1) bei 25 Konferenzen (Level 2)

7. Fazit

7.1 Diskussion der Befunde

Die vorliegenden Daten zeigen, dass unter Gutachter*innen bei bedeutenden Fachtagungen der deutschsprachigen Kommunikationswissenschaft durchaus Uneinigkeit besteht. Dies betrifft sowohl die Gesamtwertung als auch alle Einzelwertungskategorien. Nochmals verstärkt tritt die Uneinigkeit bei der Passung zum Tagungsthema zutage, vermutlich, weil manche Gutachter*innen die Tagungsthemen enger auslegen als andere. Unsere Daten zeigen aber auch, dass die Bildung von Durchschnitten über mehrere Kriterien hinweg die Übereinstimmung der Urteile erhöht; eine durch die Gutachter*innen vergebene Gesamtwertung hat hingegen keine höhere Übereinstimmung als die verschiedenen Bewertungen von Einzelkriterien. Sieht man von der thematischen Passung ab, gibt es keine wesentlichen Unterschiede zwischen den Bewertungskriterien: Die Urteile sind also z. B. bei der Einschätzung der Methodik nicht reliabler oder weniger reliabel als bei der Beurteilung von Originalität, Relevanz, Theorie oder Darstellung. Dieser Befund erstaunt ein wenig, weil doch zu vermuten wäre, dass Originalität eher im Auge des Betrachters läge als methodische Qualität. Doch auch bei solch „härteren" Kriterien scheint es keinen generellen Konsens zu geben, welche Studie methodisch gut und welche eher schwach ist. Das ist möglicherweise auf den Status der Kommunikationswissenschaft als Integrationsfach zurückzuführen, das methodische Ansätze aus verschiedenen Disziplinen anwendet. Das führt dazu, dass sich die Präferenzen und Kompetenzen verschiedener Forscher*innen deutlich unterscheiden. Nur wenige dürften den gesamten Methodenkanon der Kommunikationswissenschaft überblicken und somit alle Arten von Studien kompetent beurteilen können. Das gleiche Argument trifft z. B. auch auf die Beurteilung der Theorie zu: Verschiedene Denkschulen und Zugänge unterscheiden sich in ihrem Theoriebegriff und in der Vorstellung, was die Theoriearbeit einer Einreichung leisten soll. Die Bemühungen der Programmmacher*innen, den Einreichungen kompetente Reviewer*innen zuzuteilen, sind deshalb umso wichtiger.

Zwischen den Fachgruppen gab es nur auf den ersten Blick scheinbare Unterschiede, die sich aber größtenteils als Folge unterschiedlicher typischer Reviewer*innenzahlen pro Paper entpuppten. Auch im Laufe der letzten Jahre hat sich das Maß der Übereinstimmung weder nach oben noch nach unten entwickelt. Die Zahl der Reviewer*innen muss rein stochastisch zu höherer Spannweite und Standardabweichung führen, doch auch über diesen Effekt hinaus führen mehr Reviewer*innen zu mehr Dissens bzw. geringerer Übereinstimmung in den Urteilen. Das ist sicher kein Grund, die Zahl der Reviews möglichst gering anzusetzen. Vielmehr wäre zu ergründen, wie sich sogar noch größere Zahlen von Reviews pro Beitrag auf die Reliabilität der Urteile auswirken. Das stößt zwar praktisch an enge Grenzen, wäre aber dem Verständnis dienlich, wie die Anzahl der Reviewer*innen und Reliabilität zusammenhängen.

Ein weiterer Befund ist, dass Paper umso uneinheitlicher beurteilt werden, je geringer die Gesamtwertung ausfällt. Nehmen wir dies als Näherung für die Qualität des Beitrags, dann kann man sich eher darauf einigen, welche Beiträge gut

sind, als welche Beiträge schlecht sind. Dies ist zum einen ein Indiz, dass sehr gute Beiträge sehr zuverlässig identifiziert werden und folglich über deren Annahme oder Nichtannahme keine Unsicherheit besteht. Man könnte sogar so weit gehen und – zum Zwecke der Qualitätssicherung – nur solche unzweifelhaft hochwertigen Beiträge annehmen und alle, bei denen es größere Divergenzen gibt, ablehnen. Allerdings wäre diese Vorgehensweise problematisch, weil dies gerade die innovativen und kontroversen Beiträge, die aus dem üblichen Rahmen fallen und einen ungewöhnlichen Zugang wählen, wohl ausschließen würde.

Der Einfluss der Qualität muss dabei vom Einfluss der Skalenränder (und damit einhergehenden Decken- und Bodeneffekten) getrennt werden. Der scheinbar kurvlineare Zusammenhang wird dadurch als linearer Zusammenhang erkennbar. Es muss zwar auch erlaubt sein und bleiben, ungewöhnliche Beiträge abzulehnen – auch diese können gut oder schlecht sein – aber der Konservatismus, der dem Peer-Review-Verfahren bisweilen zur Last gelegt wird, könnte sich in diesem Muster niederschlagen. Verderben also viele Reviewer*innen den Brei oder bringen sie vielfältige Perspektiven ein? Beides ist gleichermaßen der Fall. Neff und Olden (2006) plädieren dafür, mehr Reviewer*innen einzusetzen, um etwaige Fehler auszugleichen und zufällige Streuung zu reduzieren. Wie bereits ausgeführt ist die Reviewlast pro Forscher*in aber ohnehin schon sehr hoch und die Zahl der Reviews pro Beitrag lässt sich nicht beliebig steigern.

7.2 Empfehlungen

Auf dieser Grundlage gelangen wir zu fünf Empfehlungen, die zukünftig Begutachtungsverfahren verbessern könnten: Erstens sollte in die Selektion von Reviewer*innen mehr Sorgfalt investiert werden. Gerade bei den von uns untersuchten Tagungen erfolgt die Auswahl (bestenfalls) nur nach thematischen Schwerpunkten, weniger aber gemäß methodischer Expertise, Erfahrungen oder aus Diversitätsgedanken heraus. Diese Auswahl müsste freilich auch davon abhängen, ob Perspektivenvielfalt gewünscht wird oder nicht. Die Selektion homogener Gutachter*innen, die sich mit der jeweiligen Thematik und Methodik ähnlich gut auskennen, könnte allerdings die Reliabilitätswerte erhöhen.

Zweitens sollte man den Gutachter*innen genauere Anweisungen und/oder Beispiele an die Hand geben, die präzisieren, was beurteilt werden soll und was nicht. Das wäre insbesondere bei der Kategorie „Passung mit dem Tagungsthema“ ratsam. Hier scheint es besonders häufig vorzukommen, dass manche Reviewer*innen streng und andere lax urteilen.

Drittens ist es ratsam, die bisher übliche Praxis mit mehreren Urteilen auf verschiedenen Bewertungsdimensionen, die zu einem Summenwert oder Mittelwert zusammengefasst werden, beizubehalten. Eine explizite Gesamtwertung ist als Ersatz nicht geeignet, um eine höhere Übereinstimmung zwischen den Reviewer*innen zu erreichen. Dieses Ergebnis widerspricht auch dem bisweilen angewendeten Vorgehen, die oftmals separat erhobene Kategorie „Gesamtbewertung“ stärker zu gewichten. Allerdings kann man über eine andere Form der Gewichtung oder automatisierten Adjustierung von Reviewergebnissen nachdenken: So könnte mittels einer z-Standardisierung bei jeder/jedem einzelnen Reviewer*in

der Tatsache Rechnung getragen werden, dass manche Personen generell unter- oder überdurchschnittliche Werte verteilen. Dies ist allerdings nur bei einer ausreichenden Anzahl von begutachteten Beiträgen pro Person sinnvoll.

Viertens könnte man gerade Reviewverfahren von Tagungen einmal als „Experimentierfläche" für alternative Begutachtungsverfahren nutzen. Beim iterativen Peer Review könnten die Autor*innen ihre Ideen durch die Auseinandersetzung mit der interessierten Fachcommunity weiterentwickeln, bis hin zum fertigen Tagungsbeitrag. Auch ließe sich auch über eine Selbstselektion von Gutachter*innen nachdenken: Diese entscheiden sich anhand der Titel oder Kurzabstracts für Beiträge, die sie begutachten möchten. Dies erhöht die Motivation und sorgt zugleich dafür, dass Reviewer*innen thematisch und methodisch zu ihnen passende Beiträge begutachten. Bei Zeitschriften ließe sich zudem ein Open Peer-Review testen, bei dem die Autoren- und Gutachternamen, die Gutachten sowie die Antworten der Autor*innen (z. B. auf einer Homepage) publiziert werden. Wenngleich diese Alternativen auch mit spezifischen Nachteilen einhergehen, wäre es begrüßenswert, einmal ihre Vorteile kennenzulernen.

Fünftens könnte ein Training von Reviewer*innen die Qualität der Verfahren verbessern. Es ist fast abstrus, dass der wissenschaftliche Nachwuchs zwar Angebote zur Fortbildung bei Datenauswertung, Methodenkenntnissen oder dem Schreiben von Aufsätzen bekommt, jedoch kaum dazu, wie man Reviews verfasst. Dabei kann ein solches Training die Qualität der Gutachten signifikant erhöhen (Schroter et al., 2004). Es würde jedoch einen hinreichenden Konsens im Fach (oder zumindest in den jeweiligen Fachgruppen) voraussetzen, wie Beiträge beurteilt werden sollten. Das würde nicht nur eine intensive Diskussion über Qualität von Forschungsbeiträgen erfordern, sondern sich auch in konkreten Handreichungen niederschlagen, die Grundlage von Reviewschulungen oder zumindest von genaueren Beschreibungen der Bewertungskriterien sein könnten. Während dies in bestimmten Bereichen die Übereinstimmung zwischen den Reviewer*innen erhöhen dürfte, steht zu vermuten, dass die Beurteilung „weicherer" Kriterien wie „Originalität" sich nur in begrenztem Umfang standardisieren und schulen lässt. Trotz aller Anstrengungen, die Übereinstimmung so weit wie möglich zu erhöhen, wird man bis auf Weiteres mit erheblicher Diversität der Urteile leben müssen. Man sollte diese insofern als ein System von „checks and balances" begreifen, in dem die Vielfalt von Standpunkten zu großer Einseitigkeit vorbeugt.

7.3 Limitationen und Ausblick

Die Aussagekraft der vorliegenden Daten ist aus verschiedenen Gründen limitiert. Wir fokussieren uns insbesondere auf vier Aspekte: Erstens sollte der Nutzen und auch die Sinnhaftigkeit von Interrater-Reliabilitäten für die vorliegende Fragestellung diskutiert werden: Ein geringer Interrater-Reliabilitätswert muss nicht unbedingt als schlecht angesehen werden, da Wissenschaft auch aus Diskurs und unterschiedlichen Ansichten besteht und ein hoher Reliabilitätswert von zu viel Konsens zeugen kann. Bailar (1991, S. 138) spitzt diese Idee noch zu und meint: „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 red-

undant“ (Hervorhebung im Original). Aus dieser Perspektive ist Dissens kein Problem, sondern ein Ausweis dafür, dass verschiedene Perspektiven in das Gesamturteil einfließen. Konsens ist dann nicht erstrebenswert, weil er der Komplexität der Urteilssituation nicht gerecht wird. Dissens könnte ein Beleg dafür sein, dass eine Einreichung kontrovers beurteilt wird, was wiederum auf ein spannendes Thema hinweist. Dann müsste man allerdings noch weiterdenken: Sind überhaupt drei Reviewer*innen genug? Müsste man nicht viel mehr Stimmen zu Wort kommen lassen? Müsste man die Reviewer*innen nach Denkschulen quotiert auswählen und zuweisen? Und wann tritt eine Sättigung ein, dass zusätzliche Reviewer*innen keine oder nur noch geringe Steigerungen der Urteilsqualität erzeugen? Dies sind alles Fragen, mit denen sich Wissenschaftler*innen beschäftigen sollten. Wichtig ist auch zu betonen, dass geringe Reliabilität überhaupt nichts mit der Validität der Begutachtung zu tun hat.

Zweitens betrifft die Kritik auch die Auswahl unserer Stichprobe: Sie beschränkt sich auf Tagungseinreichungen großer deutschsprachiger Tagungen der letzten fünf Jahre. Die Auswahl vernachlässigt so erstens kleinere Fachgruppen sowie Einreichungen bei Zeitschriften. Auch gingen fast ausschließlich Abstracts in die Studie ein: Ob sich die Befunde bei Aufsätzen ähnlich zeigen würden, können wir (bislang) nicht klären. Hier wäre ein Vergleich kurzer Abstracts, Extended Abstracts und Full Paper spannend: So könnte man ermitteln, ob mit zunehmender Länge auch einheitlichere Bewertungen einhergehen oder ob der Dissens vielleicht sogar noch weiter zunimmt. Dabei sollte man nicht nur bedenken, dass Abstracts anders begutachtet werden könnten als Full Paper, sondern auch, dass Tagungseinreichungen anders begutachtet werden könnten als beispielsweise Aufsätze bei (Top) Journals: So wäre es möglich, dass Tagungseinreichungen schneller und oberflächlicher gelesen werden, weil man meist mehrere auf einmal begutachten muss und keine derart ausführlichen Kommentare wie bei Zeitschriften erwartet werden. Auch die Auswahl der Gutachter*innen dürfte hier oftmals einer anderen Logik folgen als bei Zeitschriften. Hinzu kommt, dass die Kommunikationswissenschaft als Fach eigene Spezifika aufweist, die ebenfalls Einfluss nehmen könnten: Als Integrationsfach nimmt es Anleihen aus verschiedensten Disziplinen und die Forschungsschwerpunkte und -methoden unterscheiden sich z.T. massiv zwischen einzelnen Standorten und Arbeitsbereichen. Unsere Befunde sollten entsprechend nicht vorschnell auf das Peer-Review-Verfahren generell übertragen werden.

Ein dritter Kritikpunkt sind die fehlenden Erklärungsvariablen: In unsere Modelle konnten wir nur jene Variablen aufnehmen, die bei den Reviews erhoben wurden bzw. die sich aus den uns zur Verfügung gestellten Datensätzen zweifelsfrei ergeben. Zukünftige Studien können die Ursachen für geringe Reliabilitätswerte tiefergehend erforschen: Welche Rolle spielen Erfahrung, zeitliche Ressourcen und Motivation der Gutachter*innen? Gibt es spezifische „Antworttendenzen“ und „Bandbreiten“ auf den Skalen der Reviewer*innen und „Lesarten“ der Bewertungskriterien? Was bedeutet es, wenn ein/eine Reviewer*in durchgängig eine „4“ (auf der Skala von 1–5) und nie weniger als „3“ Punkte vergibt? Bewertet diese/dieser Reviewer*in generell sehr wohlwollend und diese Tendenz sollte „herausgemittelt werden“? Hat die/der Reviewer*in das Abstract mit der „4“ gar nicht gelesen und einfach „durchgeklickt“ oder einen Gesamteindruck ausgedrückt, der sich über alle Bewertungs-

kriterien hinweg erstreckt? Was verstehen die verschiedenen Reviewer*innen unter den notwendigerweise sehr allgemeinen Bewertungskriterien und wie wenden sie sie auf spezifische Beiträge an? Hier bieten sich auch weiterführende qualitative Studien an. Darüber hinaus gehend wäre es spannend, weitere Perspektiven einzubeziehen, vor allem diejenige der Autor*innen der Beiträge und die der Organisatoren des Reviewverfahrens. Wie reagieren die Reviewer*innen auf die Urteile, halten sie sie für gerecht oder ungerecht, wie würden sie auf die Kritikpunkte antworten? Gerade bei Zeitschriften ist darüber hinaus interessant, wie der oder die Herausgeber*innen mit reliablen oder wenig reliablen Reviews umgehen und wann sie die Möglichkeit für eine Überarbeitung erwägen und wann nicht.

Viertens ist fraglich, ob die von Krippendorff vorgeschlagenen Cut-Off-Werte für Reliabilitätskoeffizienten, die ja für „klassische" Inhaltsanalysen entstanden, bei der Analyse von Reviewergebnissen ähnliche Gültigkeit beanspruchen; es zeigt aber eben gerade auch, dass die Reliabilität von Reviews an die von Codierungen nicht annähernd herankommt. Gerade, wenn Perspektivenvielfalt auch absichtlich angestrebt wird, wären Werte größer als 0,8 bzw. 0,667 weder zu erwarten noch wünschenswert. Die Übereinstimmung wäre demzufolge eine künstliche Übereinstimmung, die aus mangelnder Perspektivenvielfalt resultiert.

Über die Zukunft des Peer-Review-Verfahrens wird seit Jahren diskutiert (z. B. Gould, 2012). So schließt sich auch diese Studie am Ende dem Resümee von Smith (2006) an, wonach Peer-Review „a system full of problems" (S. 178) sei. Mit weiterer Forschung werden wir hoffentlich genauer beurteilen können, ob auch der zweite Teil seines Zitats zutrifft: Es sei „a system full of problems but the least worst we have" (S. 178).

Literaturverzeichnis

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics, 3*(1), 63–84. https://doi.org/10.1007/s11948-997-0017-3

Bailar, J. C. (1991). Reliability, fairness, objectivity, and other inappropriate goals in peer review. *Behavioral and Brain Sciences, 14*(1), 137–138. https://doi.org/10.1017/s0140525x00065705

Bakanic, V., McPhail, C., & Simon, R. J. (1987). The manuscript review and decision-making process. *American Sociological Review 52*(5), 631–642. https://doi.org/10.2307/2095599

Birukou, A, Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., & Wassef, A. (2011). Alternatives to peer review: novel approaches for research evaluation. Frontiers in Computational Neuroscience, 5, 1–12. https://doi.org/10.3389/fncom.2011.00056

Blackburn, J. L., & Hakel, M. D. (2006). An examination of sources of peer-review bias. *Psychological Science, 17*(5), 378–382. https://doi.org/10.1111/j.1467-9280.2006.01715.x

Bornmann, L., & Daniel, H.-D. (2008). The effectiveness of the peer review process: inter-referee agreement and predictive validity of manuscript refereeing at Angewandte Chemie. *Angewandte Chemie International Edition, 47*(38), 7173–7178. https://doi.org/10.1002/anie.200800513

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE, 5*(12), 1–10. https://doi.org/10.1371/journal.pone.0014331

Bornstein, R. F. (1991). Manuscript review in psychology: psychometrics, demand characteristics, and an alternative model. *The Journal of Mind and Behavior, 12*(4), 429–167.

Boster, F. J. (2002). On making progress in communication science. *Human Communication Research, 28*(4), 473–490. https://doi.org/10.1111/j.1468-2958.2002.tb00818.x

Campanario, J. M. (1998). Peer review for journals as it stands today – part I. *Science Communication, 19*(3), 181–211. https://doi.org/10.1177/1075547098019003002

Callaham, M. L., Baxt, W. G., Waeckerle, J. F., & Wears, R. L. (1998). Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *Journal of the American Medical Association, 280*(3), 229–231. https://doi.org/10.1001/jama.280.3.229

Ceci, S. J., & Peters, D. (1982). Peer review – a study of reliability. *Change, 14*(6), 44–48. https://doi.org/10.1080/00091383.1982.10569910

Ceci, S. J., & Peters, D. (1984). How blind is blind review? *American Psychologist, 39*(12), 1491–1494. https://doi.org/10.1037/0003-066X.39.12.1491

Cicchetti, D. V. (1980). Reliability of reviews for the American Psychologist: A biostatistical assessment of the data. American Psychologist, 35, 300-303.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*(1), 119–186. https://doi.org/10.1017/S0140525X00065675

Cole, S., Cole, J., & Simon, G. (1981). Chance and consensus in peer review. *Science, 214*(4523), 881–886. https://doi.org/10.1126/science.7302566

Epstein, W. M. (1990). Confirmational response bias among social work journals. *Science, Technology & Human Values, 15*(1), 9–38. https://doi.org/10.1177/016224399001500102

Eysenck, H. J., & Eysenck, S. B. (1992). Peer review: Advice to referees and contributors. *Personality and Individual Differences, 13*(4), 393–399. https://doi.org/10.1016/0191-8869(92)90066-X

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–49. https://doi.org/10.1016/0895-4356(90)90158-L

Feng, G. C. (2013). Factors affecting intercoder reliability: a Monte Carlo experiment. *Quality & Quantity, 47*(5), 2959–2982. https://doi.org/10.1007/s11135-012-9745-9.

Fiske, D. W., & Fogg, L. F. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. *American Psychologist*, *45*(5), 591–598. https://doi.org/10.1037/0003-066X.45.5.591

Goldman, R. L. (1994). The reliability of peer assessments – a meta-analysis. *Evaluation & the Health Professions, 17*(1), 3–21. https://doi.org/10.1177/016327879401700101

Gould, T. H. P. (2012). The future of peer review: Four possible options to nothingness. *Publishing Research Quarterly, 28*(4), 285–293. https://doi.org/10.1007/s12109-012-9297-9

Gwet, K. L. (2014). *R functions for calculating agreement coefficients. Advanced Analytics*. Retrieved from http://www.agreestat.com/r_functions.html.

Hirschauer, S. (2004). Peer Review Verfahren auf dem Prüfstand: Zum Soziologiedefizit der Wissenschaftsevaluation [Putting peer-reviewing to the test: On the sociology defi-

cit in the evaluation of science]. *Zeitschrift Für Soziologie, 33*(1), 62–83. https://doi.org/10.1515/zfsoz-2004-0104

Kemp, S. (2005). Editorial comment: agreement between reviewers of Journal of Economic Psychology submissions. *Journal of Economic Psychology 26*(5), 779–784. https://doi.org/10.1016/j.joep.2005.05.004

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Los Angeles, CA: SAGE.

Krippendorff, K. (2012). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. *Annals of the International Communication Association, 36*(1), 481–499. https://doi.org/10.1080/23808985.2013.11679143

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Langenbucher, W. R. (2016). Die "Philosophie" der Herausgeberzeitschrift und ihr (notwendiger?) Wandel [The editor-governed journal's "philosophy" and its (necessary) change]. *Publizistik, 61*(1), 7–15. https://doi.org/10.1007/s11616-016-0254-z

Lauf, E. (2001). Publish or perish? Deutsche Kommunikationsforschung in internationalen Fachzeitschriften [Publish or perish? German communication research in international scientific journals]. *Publizistik, 46*(4), 369–382. https://doi.org/10.1007/s11616-001-0119-x

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lindsey, D. (1988). Assessing precision in the manuscript review process: A little better than a dice roll. *Scientometrics, 14*(1–2), 75–82. https://doi.org/10.1007/BF02020243

Mahoney, M. J. (1987). Scientific publication and knowledge politics. *Journal of Social Behavior and Personality, 2*(2), 165–176.

Marsh, H. W., Bond, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist, 42*(1), 33–38. https://doi.org/10.1080/00050060600823275

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168. https://doi.org/10.1037/0003-066X.63.3.160

Meyen, M., & Wiedemann, T. (2016). Peer review revisited. Eine Untersuchung der SCM-Gutachten 2014/15. *Studies in Communication and Media, 5*(1), 6–30. https://doi.org/10.5771/2192-4007-2016-1-6

McNutt, R. A., Evans, A. T., Fletcher, R. H., & Fletcher, S. W. (1990). The effects of blinding on the quality of peer review. A randomized trial. *JAMA, 263*(10), 1371–1376. https://doi.org/10.1001/jama.1990.03440100079012

Neff, B. D., & Olden, J. D. (2006). Is peer review a game of chance? *BioScience, 56*(4), 333–340. https://academic.oup.com/bioscience/article/56/4/333/229033?searchresult=1

Neidhardt, F. (2010). Selbststeuerung der Wissenschaft: Peer Review [Self-governance of science: Peer review]. In D. Simon, A. Knie, & S. Hornbostel (Hrsg.), *Handbuch Wissenschaftspolitik* (S. 280–292). Wiesbaden: VS Verlag für Sozialwissenschaften.

Neumann, W. R., Davidson, R., Joo, S.-H., Park, Y. J., & Williams, A. E. (2008). The seven deadly sins of communication research. *Journal of Communication, 58*(2), 220–237. https://doi.org/10.1111/j.1460-2466.2008.00382.x

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5*(4), 85–90.

Olson, C. M. (1990): Peer review of the biomedical literature. *American Journal of Emergency Medicine, 8*(4), 356–358.

Oxman, A., Guyatt, G. H., Singer, J., Goldsmith, C. H., Hutchison, B. G., Milner, R. A., & Streiner, D. L. (1991). Agreement among reviewers of review articles. *Journal of Clinical Epidemiology, 44*(1), 91–98.

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5*(2), 187–195. https://doi.org/10.1017/S0140525X00011183

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100. https://doi.org/10.1037/a0015108

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., & Smith, R. (2004). Effects of training on quality of peer review: randomised controlled trial. *BMJ : British Medical Journal, 328*(7441), 673. https://doi.org/10.1136/bmj.38023.700775.AE

Scott, W. A. (1974). Interreferee agreement on some characteristics of manuscripts submitted to the Journal of Personality and Social Psychology. *American Psychologist, 29*, 698–702.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine, 99*(4), 178–182. https://doi.org/10.1258/jrsm.99.4.178

Suls, J., & Martin, R. (2009). The air we breathe: A critical look at practices and alternatives in the peer-review process. *Perspectives on Psychological Science, 4*(1), 40–50. https://doi.org/10.1111/j.1745-6924.2009.01105.x

van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ, 318*(1), 23–27. https://doi.org/10.1136/bmj.318.7175.23

Vecchio, R. P. (2006). Journal reviewer ratings: issues of particularistic bias, agreement, and predictive validity within the manuscript review process. *Bulletin of Science, Technology & Society, 26*(3), 228–242. https://doi.org/10.1177/0270467606288595

Warrens, M. J. (2010). A formal proof of a paradox associated with Cohen's Kappa. *Journal of Classification, 27*(3), 322–332. https://doi.org/10.1007/s00357-010-9060-x

Wilson J. D. (1978). Peer review and publication. Presidential address before the 70th annual meeting of the American Society for Clinical Investigation, San Francisco, California, 30 April 1978. *Journal of Clinical Investigation, 61*(6), 1697–1701.

Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind inter-coder reliability indices. In C. T. Salmon, *Communication Yearbook*, 36, 419–80. New York: Routledge.

Zuckerman, H., & Merton, R. (1971). Patterns of Evaluation in Science. *Minerva, 9*(1), 66–100.

EXTENDED ABSTRACT

On the reliability of peer-reviewing.
A study of inter-rater reliability between reviewers for DGPuK conference submissions

Thomas Koch & Stefan Geiß

Before academic studies are accepted for publication in journals or for presentations at conferences, journal editors or conference organizers typically check the quality of these submissions to ensure that only the best studies are accepted (Smith, 2006). This is usually done by peer review, an evaluation process in which independent colleagues with expertise in the same area of research rate the submission (Armstrong, 1997; Campanario, 1998). Thereby, reviewers have a significant voice in defining the topics of (top) journals and conferences and, in turn, the agenda of the science community. Peer review also decides about careers: Who will get scholarships, awards, (top) publications, and research funds (Marsh, Jayasinghe, & Bond, 2008; Smith, 2006)?

Since the 1970s, the peer review itself has been challenged. Especially in medicine and psychology, the peer review system itself came under scrutiny and has been criticized for many reasons (Hirschauer, 2004; Neff & Olden, 2006). It is criticized for being slow, expensive, subjective, prone to bias, easy to abuse, and almost useless in detecting cases of fraud (Armstrong, 1997; Ceci & Peters, 1982; Marsh, Bond, & Jayasinghe, 2007). Moreover, peer review is blamed to aggravate the publication of certain studies, e.g., replications (Boster, 2002; Campanario, 1998; Schmidt, 2009), non-significant findings (Armstrong, 1997; Bornstein, 1991), or unconventional contributions that contradict prevailing paradigmatic views (Epstein, 1990; Eysenck & Eysenck, 1992; Mahoney, 1987). In addition, reviewers do not always conduct systematic reviews, but are often guided by easily available heuristics that are not necessarily related to the quality of a submission, like e.g., (non-)significance of findings, sample size, or complexity of the calculations (Armstrong, 1997).

One major point of criticism concerns the reliability of peer review (Marsh, Bond, & Jayasinghe, 2007). A reliable instrument should yield the same or similar outcomes when being used at a later time. Optimally, a (fully) reliable peer review process should ensure that a manuscript receives similar reviews or decisions at the same journal or conference (Cicchetti, 1991), independent of the particular set of reviewers chosen. Since manuscripts are usually evaluated by several reviewers, the question of interrater reliability is particularly important (Bornmann, Mutz, & Daniel, 2010), and it can be checked empirically based on the reviews

that concern the same submission. Previous studies analyzing interrater reliability vary in their results and conclusions, although they are quite similar regarding their methods (overviews by Bornmann, Mutz, & Daniel, 2010; Cicchetti, 1991; Goldman, 1994; Lindsey, 1988). They usually rely on existing reviews of conference or journal submissions and examine the extent to which the reviewers of the same submission agree or disagree.

On the one hand, studies show substantial agreement between two or more reviewers (Kemp, 2005; Neumann et al., 2008; Vecchio, 2006). The majority of the studies, on the other hand, reveal a rather low level of agreement between reviewers (e.g., Bakanic, McPhail, & Simon, 1987; Blackburn & Hakel, 2006; Callaham, Baxt, Waeckerie, & Wears, 1998). Meta-analyses seem to confirm the latter findings. Goldman (1994) includes 21 data sets from 13 studies in a meta-analysis, yielding a Cohen's Kappa of .31, which he interprets as critical. A meta-analysis by Bornmann, Mutz, and Daniel (2010) including 48 studies shows even lower agreement values (Cohen's Kappa = .17) and correlations (r = .34). It is unclear whether peer review processes in communication science would yield similar findings. In addition, the factors influencing the extent of agreement or disagreement remain largely unknown.

The current study aims to address this research gap by analyzing the reliability of reviews in communication research. Do reviewers agree in their assessment of submissions and how prevalent are opposite reviews? Does the reliability between different evaluation criteria (e.g. method, theory, presentation, etc.) differ? Moreover, can we identify pitfalls and problems to improve reliability of the review process?

To that end, we studied the reviews made for the general conference of the German Communication Association (DGPuK) and the annual conferences of its five largest divisions in the past five conferences. Based on 3,537 reviews from 23 conferences, we analyze inter-rater reliability (Krippendorff's α und Brennan und Prediger's κ), ranges, and standard deviations, regarding both criteria-based scores (fit with conference theme, innovativeness, relevance, theory, method, clarity of presentation) and overall scores.

The study shows that there is substantial disagreement between reviewers, with a Krippendorffs α of 0.276 over all criteria. As expected, Brennan and Prediger's Kappa shows a more favorable value, namely 0.649. Across all scores, the range of reviewer judgments was 27% of contributions at 0 (we have always rounded down for this evaluation, otherwise used the exact values), i.e. the (usually 2 or 3) reviewers were absolutely in agreement. Since the random agreement is already at 20%, however, this value is surprisingly low. 41% of the ratings showed a deviation from one scale point. Deviations from 2 scale points occurred in 22% of the ratings, 3 scale points in 7% of the ratings. The maximum deviation of 4 scale points occurred in about 3 out of 100 evaluations. In total, there are deviations of 2 or more scale points on a scale of 1 to 5 in 31% of the evaluations, which means considerable disagreement between the experts in almost one third of the evaluation pairs.

The rather low level of agreement concerns both the overall evaluation and all criteria. However, the reliability regarding the criteria "fit with conference theme"

is significantly lower than the other criteria, presumably, because some reviewers interpret the conference topic more narrowly than others do. Yet, our data also show that calculating mean or sum scores across criteria leads to higher agreement between reviewers; an overall evaluation given explicitly by the reviewer, however, has no higher agreement than the different evaluations of individual criteria. Apart from the thematic fit, there are no significant differences between the evaluation criteria: Expert judgments, for example, are not more or less reliable in assessing the method than in assessing originality, relevance, theory or clarity of presentation.

We suggest five recommendations to improve future review processes: First, reviewers should be selected (probably more) carefully, based not only on their thematic focus, but also on other criteria, such as their methodological expertise, experience, or diversity considerations. Secondly, the reviewers should be given more precise instructions and examples. This could be especially helpful with respect to the category "fit with conference theme". Thirdly, we recommend to maintain the practice of calculating a mean or sum score across different evaluation criteria. An explicit overall rating of the reviewer is not an adequate alternative to a mean or sum score. Fourthly, future review processes (of conferences) could be used as an "experimental zone" for alternative review processes, like iterative approaches, self-selection of reviewers, or open peer review. Finally, training new reviewers is essential and could improve the quality of the process. It is almost absurd that young scholars are trained in data analysis, methods, or writing, but hardly learn how to review.

Literature

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics, 3*(1), 63–84. http://doi.org/10.1007/s11948-997-0017-3

Bakanic, V., McPhail, C., & Simon, R. J. (1987). The manuscript review and decision-making process. *American Sociological Review 52*(5), 631–642. https://doi.org/10.2307/2095599

Blackburn, J. L., & Hakel, M. D. (2006). An examination of sources of peer-review bias. *Psychological Science, 17*(5), 378–382. https://doi.org/10.1111/j.1467-9280.2006.01715.x

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE, 5*(12), 1–10. https://doi.org/10.1371/journal.pone.0014331

Boster, F. J. (2002). On making progress in communication science. *Human Communication Research, 28*(4), 473–490. https://doi.org/10.1111/j.1468-2958.2002.tb00818.x

Campanario, J. M. (1998). Peer review for journals as it stands today – part I. *Science Communication, 19*(3), 181–211. https://doi.org/10.1177/1075547098019003002

Callaham, M. L., Baxt, W. G., Waeckerle, J. F., & Wears, R. L. (1998). Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *Journal of the American Medical Association 280*(3), 229–231. https://doi.org/10.1001/jama.280.3.229

Ceci, S. J., & Peters, D. (1982). Peer review – a study of reliability. *Change, 14*(6), 44–48. https://doi.org/10.1080/00091383.1982.10569910

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*(1), 119–186. https://doi.org/10.1017/S0140525X00065675

Epstein, W. M. (1990). Confirmational response bias among social work journals. *Science, Technology & Human Values, 15*(1), 9–38. https://doi.org/10.1177/016224399001500102

Eysenck, H. J., & Eysenck, S. B. (1992). Peer review: Advice to referees and contributors. *Personality and Individual Differences, 13*(4), 393–399. https://doi.org/10.1016/0191-8869(92)90066-X

Goldman, R. L. (1994). The reliability of peer assessments – a meta-analysis. *Evaluation & the Health Professions, 17*(1), 3–21. https://doi.org/10.1177/016327879401700101

Hirschauer, S. (2004). Peer Review Verfahren auf dem Prüfstand: Zum Soziologiedefizit der Wissenschaftsevaluation. *Zeitschrift Für Soziologie, 33*(1), 62–83. https://doi.org/10.1515/zfsoz-2004-0104

Kemp, S. (2005). Editorial comment: agreement between reviewers of Journal of Economic Psychology submissions. *Journal of Economic Psychology 26*(5), 779–784. https://doi.org/10.1016/j.joep.2005.05.004

Lindsey, D. (1988). Assessing precision in the manuscript review process: A little better than a dice roll. *Scientometrics, 14*(1-2), 75-82. https://doi.org/10.1007/BF02020243

Mahoney, M. J. (1987). Scientific publication and knowledge politics. *Journal of Social Behavior and Personality, 2*(2), 165–176.

Marsh, H. W., Bond, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist, 42*(1), 33–38. https://doi.org/10.1080/00050060600823275

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168. https://doi.org/10.1037/0003-066X.63.3.160

Neff, B. D., & Olden, J. D. (2006). Is peer review a game of chance? *BioScience, 56*(4), 333–340. https://academic.oup.com/bioscience/article/56/4/333/229033?searchresult=1

Neumann, W. R., Davidson, R., Joo, S.-H., Park, Y. J., & Williams, A. E. (2008). The seven deadly sins of communication research. *Journal of Communication, 58*(2), 220–237. https://doi.org/10.1111/j.1460-2466.2008.00382.x

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100. https://doi.org/10.1037/a0015108

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine, 99*(4), 178–182. https://doi.org/10.1177/014107680609900414

Vecchio, R. P. (2006). Journal reviewer ratings: issues of particularistic bias, agreement, and predictive validity within the manuscript review process. *Bulletin of Science, Technology & Society, 26*(3), 228–242. http://doi.org/10.1177/0270467606288595