

Diskussionspapier des
Instituts für Organisationsökonomik

2/2019

Begutachtungsverfahren
nach Zahl, Gewichtung und Fehlern der Gutachten

Alexander Dilger

Discussion Paper of the
Institute for Organisational Economics

**Diskussionspapier des
Instituts für Organisationsökonomik
2/2019**

Februar 2019

ISSN 2191-2475

**Begutachtungsverfahren
nach Zahl, Gewichtung und Fehlern der Gutachten**

Alexander Dilger

Zusammenfassung

Begutachtungsverfahren werden analysiert anhand der Zahl der Gutachten, der Entscheidungsregel bei abweichenden Gutachtenempfehlungen und den Annahme- und Ablehnungsfehlern der Gutachten. Begutachtungsverfahren können grundsätzlich strenger oder lockerer sein, was entsprechend zu weniger fehlerhaften Annahmen oder Ablehnungen führt.

JEL-Codes: C44, D71, D81, I23

Review Processes by the Number, Weights and Mistakes of Reviews

Abstract

Review processes are analysed by the number of reviews, the decision rule in case of contradictory recommendations and the mistakes of the reviews concerning acceptance and rejection. Review processes can be stricter or laxer in general with accordingly less incorrect acceptances or rejections.

Im Internet unter:

http://www.wiwi.uni-muenster.de/io/forschen/downloads/DP-IO_02_2019

Westfälische Wilhelms-Universität Münster
Institut für Organisationsökonomik
Scharnhorststraße 100
D-48151 Münster

Tel: +49-251/83-24303 (Sekretariat)
E-Mail: io@uni-muenster.de
Internet: www.wiwi.uni-muenster.de/io

Begutachtungsverfahren nach Zahl, Gewichtung und Fehlern der Gutachten*

1. Einleitung

Formale Begutachtungsverfahren (*Peer Review*) haben eine hohe und noch zunehmende Bedeutung.¹ Das betrifft nicht nur wissenschaftliche Zeitschriften, sondern z. B. auch Konferenzen, Drittmittel und die Evaluation von Hochschulen und Forschungseinrichtungen. Früher haben die Zeitschriftenherausgeber, Konferenzveranstalter, Drittmittelgeber oder Träger von Hochschulen und Forschungseinrichtungen nach eigenem Ermessen und vielleicht noch unterstützt durch informellen Rat entschieden. Inzwischen haben formelle Begutachtungsverfahren stark an Bedeutung gewonnen und werden sogar als Kriterium für (hohe) wissenschaftliche Qualität angesehen.²

In diesem Beitrag wird abstrakt analysiert, wie die Zahl, Gewichtung (der annehmenden und ablehnenden Empfehlungen, falls sich diese widersprechen) und Fehler der Gutachten in solch einem Verfahren die Annahmewahrscheinlichkeit beeinflussen. Zur Vereinfachung wird angenommen, dass die Gutachten nur die Annahme oder Ablehnung empfehlen³ und die nachfolgende Entscheidung allein auf diese Empfehlungen gestützt wird. In der Realität haben die Herausgeber, Veranstalter oder Drittmittelgeber weiterhin das Letztentscheidungsrecht, wobei sie sich auch auf die Inhalte der Gutachten und zusätzliche (schlimmstenfalls außer- und unwissenschaftliche) Überlegungen stützen können.⁴ Trotzdem kommt es häufig vor, dass allein den Gutachterempfehlungen gefolgt wird. Einen weiteren großen Einfluss haben die Herausgeber etc. bei der Auswahl der Gutachter. Die Berechnungen in diesem Beitrag sind kein Selbstzweck, sondern strukturieren die inhaltliche Diskussion und erlauben die Herleitung von Handlungsempfehlungen für die Begutachtungspraxis.

* Dieser Beitrag wurde beim 21. Workshop Hochschulmanagement der wissenschaftlichen Kommission Hochschulmanagement des Verbandes der Hochschullehrer für Betriebswirtschaft e. V. (VHB) am 14. Februar 2019 in Wien präsentiert. Ich danke den Veranstaltern und Teilnehmern. Für die ausgedrückten Ansichten und mögliche Fehler bin natürlich ich allein verantwortlich.

¹ Zur historischen Entwicklung siehe Burnham (1990) und Spier (2002).

² Vgl. Ware (2008).

³ Tatsächlich können Gutachter nicht nur Empfehlungen dazwischen abgeben, sondern empfehlen in der Praxis insbesondere von wissenschaftlichen Zeitschriften meistens auch Verbesserungen oder zumindest Veränderungen der Beiträge, was selbst wieder kritisiert wird (siehe Tsang/Frey 2007), den Prozess deutlich komplizierter macht und hier nicht näher analysiert wird.

⁴ Siehe Schwartz/Zamboanga (2009) mit Empfehlungen für Herausgeber, um den Prozess zu verbessern.

Im nachfolgenden zweiten Kapitel geht es um die Zahl der Gutachten und im dritten Kapitel um ihre Gewichtung. Damit ist nicht gemeint, dass ein Gutachten oder auch Gutachter per se mehr zählt als ein anderes bzw. anderer, sondern wie bei abweichenden Gutachterempfehlungen verfahren wird, ob also z. B. bei zwei unterschiedlichen Gutachten die Annahme- oder Ablehnungsempfehlung den Ausschlag gibt. Im vierten Kapitel geht es um Fehler der Gutachten relativ zu einem objektiven, intersubjektiven oder auch subjektiven Maßstab und wie sich diese Fehler auf die eigentliche Annahme- oder Ablehnungsentscheidung auswirken. Der Beitrag schließt mit einem kurzen Fazit und Ausblick.

2. Zahl der Gutachten

Es gibt eine natürliche, meist kleine (einstellige) Zahl an Gutachten. Hier wird zur Vereinfachung angenommen, dass die Gutachten gleichzeitig und unabhängig voneinander erstellt werden (mit einer im nächsten Kapitel betrachteten Ausnahme). Die Annahme- oder Ablehnungsentscheidung hinsichtlich eines Beitrags oder Antrags wird allein auf diese Gutachten und deren Empfehlungen gestützt.

Wenn es nur ein einziges Gutachten gibt, wird dementsprechend dessen Empfehlung gefolgt. Empfiehlt der Gutachter mit Wahrscheinlichkeit⁵ x die Annahme, ist x folglich (unter den getroffenen Annahmen) die Annahmewahrscheinlichkeit. Bei zwei und mehr Gutachten kommt es, wenn diese nicht alle dasselbe empfehlen, auf die Gewichtung der Empfehlungen an, die Gegenstand des nächsten Kapitels ist.

3. Gewichtung der Gutachten

Gewichtung meint hier nicht, dass Gutachten bzw. Gutachter verschieden wichtig sind und unterschiedlich behandelt werden, obgleich das in der Praxis durchaus vorkommt und ebenfalls modelliert werden könnte. In diesem Kapitel geht es darum, wie formal gleichrangige Voten zur Annahme- oder Ablehnungsentscheidung aggregiert werden.

Wenn alle Gutachter dasselbe empfehlen, spielt die Gewichtung keine Rolle und wird einfach dieser Empfehlung gefolgt, wie sich direkt aus der entsprechenden Annahme im ersten Kapitel ergibt.

⁵ Es gilt $0 < x < 1$, da in diesem Beitrag von den Randfällen $x = 0$ und $x = 1$ abgesehen wird.

Bei definitiv zwei Gutachten kann bei gegenläufigen Voten ein Beitrag entweder (1) angenommen, (2) abgelehnt oder (3) gelöst werden. Die Annahmewahrscheinlichkeit ist bei unkorrelierten Gutachten

$$(1) \quad 1-(1-x)^2 = 2x-x^2,$$

$$(2) \quad x^2,$$

$$(3) \quad x^2+(1-x)x = x.$$

Die Ablehnungswahrscheinlichkeit ist jeweils die Gegenwahrscheinlichkeit. Diese und die nachfolgenden Werte ergeben sich einfach durch Berechnung aller möglichen Fällen (einschließlich gleichlautender Voten). Es gilt $(1) > (3) > (2)$, weil die entsprechenden Entscheidungsregeln zunehmend strenger werden. Bei (3) ist die Annahmewahrscheinlichkeit gleich dem Fall von nur einem Gutachten und der Informationsgehalt ist durch das zufällige Lösen bei gegenläufigen Voten auch nicht höher. Dementsprechend ist ein Gutachten effizienter als zwei Gutachten mit möglichem Losentscheid.

Falls als vierte Möglichkeit bei zwei gegenläufigen Gutachten noch ein drittes Schiedsgutachten eingeholt wird und entscheidet, beträgt die Annahmewahrscheinlichkeit insgesamt

$$(4) \quad 3x^2-2x^3.$$

Es gilt $(1) > (4) > (2)$. Bei $x = 0,5$ ist $(3) = (4)$. Bei $x < 0,5$ gilt $(3) > (4)$ und bei $x > 0,5$ umgekehrt $(3) < (4)$. In jedem Fall ist der Informationsgehalt durch das Schiedsgutachten höher als durch einen Losentscheid.

Wenn bei von vornherein drei Gutachten die Mehrheit entscheidet, ist die Annahmewahrscheinlichkeit ebenfalls $(4) 3x^2-2x^3$. Demnach ist in $1-2x+2x^2$ der Fälle (wenn die ersten beiden Gutachten zu derselben Empfehlung kommen) das 3. Gutachten redundant und unnötig. Erst zwei Gutachten und nur bei widersprechenden Voten noch ein drittes ist also effizienter als gleich drei Gutachten. Das gilt nicht bei knapper Zeit, falls diese für ein späteres Einholen eines dritten Gutachtens nicht ausreicht, was gerade vor Konferenzen der Fall sein könnte.⁶ Bei hinreichend viel Zeit bestünde jedoch ein weiterer Vorteil in einem späteren Gutachten in der hier nicht modellierten Möglichkeit, dass dieses auf die inhaltlichen Argumente der ersten beiden Gutachten Bezug nimmt, die häufig verschieden stark sind. In der Praxis kann auch

⁶ Für (bessere) Zeitschriften könnte hingegen ein längerer Begutachtungsprozess vorteilhaft sein, um Einreichungen zu geringer Qualität abzuschrecken, vgl. Azar (2005).

der Herausgeber oder Veranstalter als Schiedsgutachter fungieren. Teilweise stützt er sich dann nur auf die Argumente der vorliegenden Gutachten, ohne den Beitrag eigenständig von sich aus zu würdigen.

Wenn bei drei Gutachten ein positives Gutachten reicht, beträgt die Annahmewahrscheinlichkeit

$$(5) \quad 1-(1-x)^3 = 3x-3x^2+x^3.$$

Wenn alle von drei Gutachten positiv sein müssen, sinkt die Annahmewahrscheinlichkeit auf

$$(6) \quad x^3.$$

Es gilt $(5) > (4) > (6)$. Zugleich gilt $(5) > (1)$ und $(6) < (2)$. Laxe Begutachtungsverfahren werden also mit zunehmender Gutachtenanzahl immer lockerer und strenge Verfahren immer strenger.

Wenn von vier Gutachten mindestens zwei positiv sein müssen, ist die Annahmewahrscheinlichkeit

$$(7) \quad 6x^2-8x^3+3x^4.$$

Wenn mindestens drei von vier Gutachten positiv sein müssen, beträgt diese Wahrscheinlichkeit

$$(8) \quad 4x^3-3x^4.$$

Wenn bei Gleichstand (2 zu 2) gelöst wird, liegt die Annahmewahrscheinlichkeit bei

$$(9) \quad 3x^2-2x^3.$$

$(9) = (4)$ bedeutet, dass vier Gutachten mit Losen nicht mehr bringen als drei Gutachten oder sogar anfangs nur zwei Gutachten mit einem dritten Gutachten bei entgegengesetzten Entscheidungen. Losen dürfte deshalb insgesamt keine gute Idee in Begutachtungsverfahren sein, da es keine zusätzlichen Informationen liefert und dasselbe Ergebnis mit weniger Gutachten und dadurch weniger belastete Gutachter erreicht werden kann.

Wenn bei vier Gutachten ein positives Gutachten reicht, beträgt die Annahmewahrscheinlichkeit

$$(10) \quad 1-(1-x)^4 = 4x-6x^2+4x^3-x^4.$$

Wenn alle vier Gutachten positiv sein müssen, sinkt die Annahmewahrscheinlichkeit auf

$$(11) \quad x^4.$$

(10) > (5) und (11) < (6) unterstreichen die bereits bei drei Gutachten beobachteten Trends. Für fünf und mehr Gutachten gilt prinzipiell dasselbe, nur die Berechnungen werden immer komplizierter.

4. Fehler der Gutachten

Gutachter geben nicht einfach zufällig irgendwelche Gutachten ab. Gute Gutachten beziehen sich auf die wissenschaftliche Qualität eines Beitrags oder Antrags. Dabei gibt es α -Fehler der Gutachten, die zu einer falsche Annahmeempfehlung führen, und β -Fehler mit einer falschen Ablehnungsempfehlung.⁷ Die Fehler können je nach Disziplin und Beitrag objektiv, intersubjektiv oder zumindest subjektiv aus Herausgebersicht bestehen. Es sei y die richtige Annahmewahrscheinlichkeit oder -quote je nach Maßstab. Ein konkreter Beitrag oder Antrag wird richtigerweise definitiv angenommen oder abgelehnt, aber für eine beliebige Einreichung ist die korrekte Annahmewahrscheinlichkeit vor der näheren Betrachtung y .

Dann ist die Wahrscheinlichkeit einer korrekten Ablehnungsempfehlung

$$(12) \quad (1-y)(1-\alpha).$$

Die Wahrscheinlichkeit einer richtigen Annahmeempfehlung liegt bei

$$(13) \quad y(1-\beta).$$

Zu einer falschen Annahmeempfehlung kommt es mit der Wahrscheinlichkeit

$$(14) \quad (1-y)\alpha.$$

Schließlich ist die Wahrscheinlichkeit einer falschen Ablehnungsempfehlung

$$(15) \quad y\beta.$$

Für die bislang betrachtete Wahrscheinlichkeit einer Annahmeempfehlung gilt

$$(16) \quad x = (1-y)\alpha + y(1-\beta).$$

⁷ Vgl. Dilger (2017), insbesondere S. 8 f.

Im Folgenden soll der Fall von zwei Gutachten genauer betrachtet werden. Dies lässt sich einfach auf drei und mehr Gutachten erweitern. Wenn von zwei Gutachten eines positiv sein muss, ist die Wahrscheinlichkeit für eine richtige Ablehnung (sei es objektiv, intersubjektiv oder subjektiv):

$$(17) \quad (1-y)(1-\alpha_1)(1-\alpha_2).$$

Die Wahrscheinlichkeit für eine richtige Annahme liegt bei

$$(18) \quad y(1-\beta_1)+y\beta_1(1-\beta_2).$$

Die Wahrscheinlichkeit für eine falsche Annahme beträgt

$$(19) \quad (1-y)\alpha_1+(1-y)(1-\alpha_1)\alpha_2.$$

Schließlich ist die Wahrscheinlichkeit für eine falsche Ablehnung

$$(20) \quad y\beta_1\beta_2.$$

Wenn beide Gutachten positiv sein müssen, steigt die Wahrscheinlichkeit für eine richtige Ablehnung auf

$$(21) \quad (1-y)(1-\alpha_1)+(1-y)\alpha_1(1-\alpha_2).$$

Dafür sinkt die Wahrscheinlichkeit einer korrekten Annahme auf

$$(22) \quad y(1-\beta_1)(1-\beta_2).$$

Die Wahrscheinlichkeit einer falschen Annahme sinkt auf

$$(23) \quad (1-y)\alpha_1\alpha_2.$$

Dafür steigt die Wahrscheinlichkeit einer falschen Ablehnung auf

$$(24) \quad y\beta_1+y(1-\beta_1)\beta_2.$$

Strengere Begutachtungsverfahren führen entsprechend zu mehr Ablehnungen, sowohl berechtigten als auch fälschlichen, während die Annahme von ungeeigneten wie auch geeigneten Einreichungen zurückgeht. Bessere Gutachter sind übrigen solche, deren α -Fehler und β -Fehler zugleich geringer sind. Dabei dürfte es auch hier einen Trade-off geben, also strengere Gutachter tendenziell einen geringeren α -Fehler mit einem höheren β -Fehler verbinden, während es bei großzügigen Gutachtern umgekehrt ist.

5. Fazit und Ausblick

Zeitschriften (Konferenzen, Drittmittelgeber etc.) können verschieden streng bei ihren Begutachtungsverfahren sein allein schon durch Variation der Gutachtenzahl und -gewichtung. Daneben können natürlich auch unterschiedlich strenge Gutachter gesucht und den Gutachtern unterschiedlich strenge Vorgaben gemacht werden.

Strengere Verfahren sind dann sinnvoll, wenn entweder der α -Fehler besonders schlimm ist (wenn etwa bei klinischen Studien die falsche Behandlung tödlich sein kann) oder richtige Annahmeentscheidungen relativ selten sind (weil y klein ist). Letzteres kann endogen sein, z. B. wenn es bei einer Zeitschrift viele Einreichungen gibt, von denen nur relativ wenige angenommen werden können, so dass die Qualitätsanforderungen erhöht werden. Die Gefahr dabei ist, dass dann auch eher besonders gute und innovative Beiträge abgelehnt werden können,⁸ insbesondere wenn verschiedene Qualitäten unterschiedlich streng beurteilt werden.⁹

Um Gutachter nicht zu sehr mit vielen, z. T. unnötigen Gutachten zu belasten, sollte ein Schiedsgutachten nur in Zweifelsfällen eingeholt werden. Das Schiedsgutachten könnte dann zusätzlich noch die Argumente der anderen Gutachten abwägen, was hier nicht modelliert wurde, was auch auf den zusätzlichen Zeitbedarf für ein nachträgliches Gutachten zutrifft.

Es sind diese und weitere Modellvariationen möglich wie z. B. die ungleiche Gewichtung von Gutachten, wenn die Gutachter als verschieden gut bekannt sind, oder die Berechnung von Korrelationen zwischen Gutachten oder auch Begutachtungsfehlern. So ist zu erwarten, dass reale Gutachten entgegen dem dritten Kapitel korreliert sind, weil sie denselben Inhalt mit ähnlichen Kriterien beurteilen.¹⁰ Diese Korrelation ist grundsätzlich ein Vorteil und führt zu zuverlässigeren Ergebnissen, während eine Korrelation der Begutachtungsfehler im vierten Kapitel die Ergebnisse des Begutachtungsprozesses eher verschlechtert.

⁸ Vgl. Campanario (1996).

⁹ Das könnte Rigor gegenüber Relevanz bevorzugen, vgl. Dilger (2012), bzw. allgemein Nebenaspekte gegenüber den eigentlichen Ideen, vgl. Ellison (2002). Auch das ließe sich durch eine Erweiterung des hier vorgestellten Grundmodells modellieren, wenn es dafür z. B. unterschiedliche α - und β -Fehler gibt.

¹⁰ Für Evidenz dazu siehe z. B. Oxman et al. (1991), während Starbuck (2003) nur eine Korrelation von 0,12 zwischen Gutachternoten findet.

Literatur

- Azar, Ofer H. (2005): „The Review Process in Economics: Is It Too Fast?“, *Southern Economic Journal* 72 (2), S. 484-491.
- Burnham, John C. (1990): „The Evolution of Editorial Peer Review“, *Journal of the American Medical Association* 263 (10), S. 1323-1329.
- Campanario, Juan Miguel (1996): „Have Referees Rejected Some of the Most-cited Articles of All Times?“, *Journal of the American Society for Information Science* 47 (4), S. 302-310.
- Dilger, Alexander (2012): „Rigor, wissenschaftliche und praktische Relevanz“, Diskussionspapier des Instituts für Organisationsökonomik 3/2012, Münster.
- Dilger, Alexander (2017): „Zur Stärkung wissenschaftlicher Kritik“, Diskussionspapier des Instituts für Organisationsökonomik 8/2017, Münster.
- Ellison, Glenn (2002): „Evolving Standards for Academic Publishing: A q - r Theory“, *Journal of Political Economy* 110 (5), S. 994-1034.
- Oxman, Andrew D./Guyatt, Gordon H./Singer, Joel/Goldsmith, Charlie H./Hutchison, Brian G./Milner, Ruth A./Streiner, David L. (1991): „Agreement Among Reviewers of Review Articles“, *Journal of Clinical Epidemiology* 44 (1), S. 91-98.
- Schwartz, Seth J./Zamboanga, Byron L. (2009): „The Peer-Review and Editorial System: Ways to Fix Something That Might Be Broken“, *Perspectives on Psychological Science* 4 (1), S. 52-61.
- Spier, Ray (2002): „The History of the Peer-review Process“, *Trends in Biotechnology* 20 (8), S. 357-358.
- Starbuck, William H. (2003): „Turning Lemons into Lemonade: Where Is the Value in Peer Reviews?“, *Journal of Management Inquiry* 12 (4), S. 344-351.
- Tsang, Eric W. K./Frey, Bruno S. (2007): „The As-Is Journal Review Process: Let Authors Own Their Ideas“, *Academy of Management Learning & Education* 6 (1), S. 128-136.
- Ware, Mark (2008): „Peer Review in Scholarly Journals: Perspective of the Scholarly Community – Results from an International Study“, *Information Services & Use* 28 (2), S. 109-112.

Diskussionspapiere des Instituts für Organisationsökonomik

Seit Institutsgründung im Oktober 2010 erscheint monatlich ein Diskussionspapier. Im Folgenden werden die letzten zwölf aufgeführt. Eine vollständige Liste mit Downloadmöglichkeit findet sich unter <http://www.wiwi.uni-muenster.de/io/de/forschen/diskussionspapiere>.

- DP-IO 2/2019** Begutachtungsverfahren nach Zahl, Gewichtung und Fehlern der Gutachten
Alexander Dilger
Februar 2019
- DP-IO 1/2019** 100 Diskussionspapiere des Instituts für Organisationsökonomik
Eine deskriptive Übersicht
Alexander Dilger/Michael Hickfang/Milan F. Klus
Januar 2019
- DP-IO 12/2018** The Impact of Stock Options on Risk-Taking
Founder-CEOs and Innovation
Michael Hickfang/Ulrike Holder
Dezember 2018
- DP-IO 11/2018** Identifying Leadership Skills Required in the Digital Age
Milan Frederik Klus/Julia Müller
November 2018
- DP-IO 10/2018** 8. Jahresbericht des Instituts für Organisationsökonomik
Alexander Dilger/Milan Frederik Klus
Oktober 2018
- DP-IO 9/2018** Konzeption einer direktdemokratischen Plattformpartei
Alexander Dilger
September 2018
- DP-IO 8/2018** Erfahrungen aus der Programmkommission für die VHB-Pfingsttagung 2018
Alexander Dilger
August 2018
- DP-IO 7/2018** 20 Jahre Workshop Hochschulmanagement
Ein deskriptiver Überblick
Alexander Dilger/Joachim Prinz/Daniel Weimar
Juli 2018
- DP-IO 6/2018** Strategic Alliances between Banks and Fintechs for Digital Innovation
Motives to Collaborate and Types of Interaction
Milan F. Klus/Todor S. Lohwasser/Friedrich Holotiuik/Jürgen Moormann
Juni 2018
- DP-IO 5/2018** Sieben Szenarien zum Euroausstieg
Alexander Dilger
Mai 2018
- DP-IO 4/2018** Die Verantwortung von Wirtschaftswissenschaftlern für Wirtschaftskrisen und die
Wirtschaft allgemein
Alexander Dilger
April 2018
- DP-IO 3/2018** Effects of the Three-Point Rule in German Amateur Football
Alexander Dilger/Gerrit Froböse
März 2018



Herausgeber:
Prof. Dr. Alexander Dilger
Westfälische Wilhelms-Universität Münster
Institut für Organisationsökonomik
Scharnhorststr. 100
D-48151 Münster

Tel: +49-251/83-24303

Fax: +49-251/83-28429

www.wiwi.uni-muenster.de/io