Programa de Doctorado en Tecnologías de la Información, Comunicaciones y Computación

Tesis Doctoral

Métodos para la caracterización del sistema de revisión por pares de artículos científicos

Daniel García Costa

Dirigido por Francisco Grimaldo Moreno y Emilia López Iñesta

Octubre 2022

A aquellos que ya no están, pero que pese a no estar, están.

Resumen:

El proceso de revisión por pares constituye una pieza clave de los engranajes que componen el sistema de publicación de la ciencia. Este vela por la calidad, integridad, reproducibilidad y robustez de los trabajos que se envían a las revistas científicas y es objeto continuo de estudio y discusión en la comunidad científica, involucrando a autores, editores y revisores en una tarea conjunta plagada de procesos sociales que se manifiestan a modo de negociación entre las diferentes partes, siendo el editor el árbitro del juego. A su vez, el proceso de revisión por pares es la insignia de calidad de las propias revistas. Mantener unos altos estándares y hacer del sistema de revisión un elemento constructivo para con los artículos que se envían es un objetivo primordial de las revistas para mantener su estatus y reputación.

Esta tesis doctoral propone el uso de la minería de datos y de textos, junto con la aplicación de técnicas de procesado de lenguaje natural para la caracterización del sistema de revisión por pares y de los textos de revisión que en él se generan, a partir de unos conjuntos de datos únicos conseguidos a través de diferentes acuerdos de compartición de datos con grandes editoriales científicas.

Por un lado, una de las novedades presentadas en este trabajo son las diferentes caracterizaciones lingüísticas sobre los textos de revisión, en los que se analiza el tipo de lenguaje empleado y se comparan su uso según la recomendación o el género del revisor. Por otro lado, se presentan diferentes trabajos sobre la caracterización del valor constructivo del proceso de revisión, analizando el tipo y la cantidad de cambios que sufren los artículos debido a los comentarios de los revisores y el efecto que estos tienen sobre la probabilidad de ser citados. Por último, se propone una métrica para medir el valor constructivo y la completitud de una revisión y se comparan diferentes grupos poblacionales según género, edad o país, así como el área o el factor de impacto de la revista.

Abstract:

The peer review process is a key part of the gears that make up the science publication system. It ensures the quality, integrity, reproducibility and robustness of the papers submitted to scientific journals and is a continuous object of study and discussion in the scientific community, involving authors, editors and reviewers in a joint task plagued by social processes that manifest themselves as a negotiation between the different parties, with the editor being the arbiter of the game. In turn, the peer review process is the badge of quality of the journals themselves. Maintaining high standards and making the review system a constructive element for the articles submitted is a primary objective of the journals in order to maintain their status and reputation.

This doctoral thesis proposes the use of data and text mining, together with the application of natural language processing techniques for the characterization of the peer review system and the review texts generated in it, based on unique data sets obtained through different data sharing agreements with large scientific publishers.

On the one hand, one of the novelties presented in this work are the different linguistic characterizations of the review texts, in which the type of language used is analyzed and its use is compared according to the recommendation or gender of the reviewer. On the other hand, different works on the characterization of the constructive value of the review process are presented, analyzing the type and amount of changes that articles undergo due to reviewers'comments and the effect that these have on the probability of being cited. Finally, a metric is proposed to measure the constructive value and completeness of a review and different population groups are compared according to gender, age or country, as well as the area or impact factor of the journal.

Resum:

El procés de revisió per parells constitueix una peça clau dels engranatges que componen el sistema de publicació de la ciència. Aquest procés vetlla per la qualitat, integritat, reproducibilitat i robustesa dels treballs que sénvien a les revistes científiques i és objecte continu déstudi i discussió en la comunitat científica. Involucra a autors, editors i revisors en una tasca conjunta plagada de processos socials que es manifesten com a una negociació entre les diferents parts, sent léditor o editora lárbitre del joc. Al seu torn, el procés de revisió per parells és la insígnia de qualitat de les pròpies revistes. Mantindre uns alts estàndards i fer del sistema de revisió un element constructiu envers els articles que sénvien és un objectiu primordial de les revistes per a mantenir el seu estatus i reputació.
Aquesta tesi doctoral proposa l'ús de la mineria de dades i de textos, juntament amb láplicació de tècniques de processament de llenguatge natural per a la caracterització del sistema de revisió per parells i dels textos de revisió que en ell es generen, a partir dúns conjunts de dades úniques aconseguides a través de diferents acords de compartició de dades amb grans editorials científiques. Dúna banda, es presenten diferents caracteritzacions lingüístiques sobre els textos de revisió, en els quals sánalitza el tipus de llenguatge emprat i es comparen el seu ús segons la recomanació o el gènere del revisor. Dáltra banda, es presenten diferents treballs sobre la caracterització del valor constructiu del procés de revisió, analitzant el tipus i la quantitat de canvis que pateixen els articles a causa dels comentaris dels revisors i léfecte que aquests tenen sobre la probabilitat de ser citats. Finalment, es presenta una mètrica per a mesurar el valor constructiu i la completitud dúna revisió i es comparen diferents grups poblacionals segons gènere, edat o país, així com l'àrea o el factor dímpacte de la revista.

Agradecimientos:

Quisiera empezar agradeciendo el apoyo incondicional de mi director y mi directora de tesis, Fran y Emi. Desde que empecé a trabajar con vosotros, allá por el 2015, mientras cursaba el grado, me habéis acogido de una manera increíble, no solo a nivel académico, sino también a nivel personal. Gracias a vosotros descubrí el mundo de la investigación y encontré en él un lugar donde enfrentarme a nuevos retos día a día y seguir llenando ese hueco que deja la sensación de no saber nada. Me habéis guiado, aconsejado y acompañado como nunca hubiera imaginado a lo largo de todos estos años. No hay palabras que puedan agradecer todo lo que hacéis por mi, no hay palabras que puedan describir lo que siento por vosotros. Gracias. Siempre hallaréis en mi un compañero, pero sobre todo, un amigo.

A Elena, "mini yo", por tu inestimable ayuda en todo este trabajo. Por arrimar el hombro siempre que se te pide y por esa actitud de luchadora nata que tienes, aunque a veces seas una dramitas... Por que aunque creas que siempre soy yo el que te enseña a ti, yo aprendo de ti cada día. Y sobre todo, por que ahora es tu turno, pequeña saltamontes.

A mi familia, por estar siempre ahí.

A mis compañeros de batallas, Adri, Pedro, Rafa y Toni. Nuestro sindicato del mal siempre estará ahí haciendo de las suyas para acometer nuestros objetivos.

A mis amigos y amigas de esa maravillosa tierra que alberga mis raíces y a la que siempre volveré, Raque, Lau, Dani, David, Sarita, Martín e Ian, por todos esos fríos inviernos y esos preciosos e inolvidables veranos, con algunos incluso, desde que tengo uso de razón.

A los hermanos Mariano, Jonathan y Alejandro, por ser mi familia adoptiva en esta otra tierra a la que, poco a poco, le voy cogiendo cariño.

Y al resto de personas que, del modo que sea, ocupáis un espacio en mi vida.

A todos vosotros, gracias.

Índice general

Capítulo 1

Introducción

1.1. Introducción

La ciencia, tal y como la conocemos hoy en día, constituye un campo que se encuentra en plena fase de expansión. Según datos del portal Dimensions.ai[1] (figura 1.1), en los últimos 12 años, el número de publicaciones científicas se ha incrementado en un 100 %, pasando de aproximadamente 3 millones de publicaciones en 2010 a 6 millones en 2021. Esto, unido a la completa digitalización de las editoriales científicas, ha propiciado la aparición de un área conocida como *Science of Science*, que tiene como principal objetivo estudiar y entender cómo funcionan los procesos de publicación de la ciencia a través de la explotación de los datos que se producen (Fortunato et al., 2018).

Aun con el auge ocasionado por este nuevo campo, en el que se han ido involucrando investigadores e investigadoras de diferentes áreas de estudio de la ciencia, el acceso a los datos no es una cuestión sencilla. La hermeticidad de las editoriales científicas en lo referente a sus procesos de publicación ha hecho que el acceso a los datos generados por estos procesos no sea una tarea trivial (Squazzoni et al., 2017b). A raíz de este problema, surgen diferentes iniciativas de compartición de datos para tratar de facilitar el acceso a los mismos desde la comunidad científica (Squazzoni et al., 2020). Por ejemplo, la COST Action, New Frontiers of Peer Review (PEERE)[2], llevada a cabo entre 2014 y 2018 y que involucró a investigadores e investigadoras de más de 30 países, logró un acuerdo de compartición de datos con editoriales científicas como Elsevier[3], Wiley[4], Springer-Nature[5] o Royal Society[6] con el fin de realizar estudios del proceso de revisión por pares de artículos científicos, o más comúnmente conocido por su término en inglés, *peer review*.

En particular, el proceso de revisión por pares constituye la piedra

[1] https://app.dimensions.ai/discover/publication
[2] https://www.peere.org/
[3] https://www.elsevier.com/es-es
[4] https://www.wiley.com/en-us
[5] https://www.springernature.com/
[6] https://royalsociety.org/

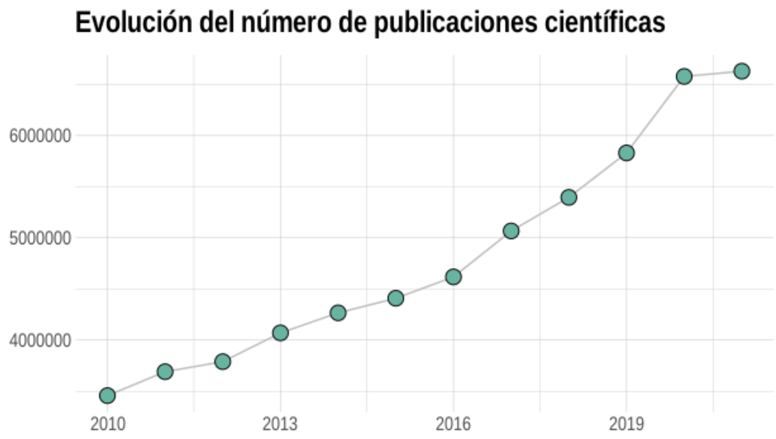

Figura 1.1: Evolución del número de publicaciones científicas, según datos obtenidos de Dimensions.ai

angular del sistema de publicaciones científicas, siendo el encargado de velar por la veracidad, completitud, reproducibilidad y, en general, por la calidad de los trabajos que reciben las revistas (Kassirer and Campion, 1994) (Bornmann, 2011). En él recae la responsabilidad de revisar los manuscritos que llegan a las revistas y de proponer las mejoras oportunas a la vez que se recomienda al editor, qué acción debe tomar. Se trata, por la cantidad de actores humanos que interactúan en él (autores, revisores y equipos editoriales) y por su propia naturaleza, de un proceso social en el que está muy presente la habilidad negociadora de cada uno y la experiencia del editor o editora en saber tomar una decisión dada toda la información generada. Por todo esto, es un proceso de gran interés en ser estudiado, ya que genera información de gran valor para entender cómo se comportan las personas involucradas a lo largo del proceso y comprender el valor que aporta al sistema de publicación de la ciencia (Lee and Moher, 2017).

En este contexto, el interés por estudiar el funcionamiento del proceso de revisión por pares es compartido por los diferentes grupos de personas que en él intervienen. Por un lado, puede ayudar a los equipos editoriales de las revistas científicas a mejorar sus procesos y profundizar en el funcionamiento de los mismos, por ejemplo, en la selección de revisores. Por otro lado, ayuda a las autoras y autores a entender correctamente el proceso de escrutinio al que se somete su obra (Rigby et al., 2018). Además, ayuda en general, la comunidad científica, a medir el valor del proceso de revisión y a proponer mejoras y/o alternativas al mismo. Por último, el estudio de estos procesos sociales (Piel, 1986) y su caracterización también puede resultar de utilidad a organismos y entidades con capacidad de toma de decisiones a la hora de identificar problemas, deficiencias, sesgos, etc (Kharasch et al., 2021).

Estudiar el funcionamiento de este proceso no es tarea sencilla, es necesario buscar y desarrollar métricas que permitan definir y caracterizar el proceso en sí mismo para medir su funcionamiento en diferentes grupos, contextos y áreas (Cowley, 2015). La cantidad de personas involucradas, asumiendo diferentes roles, y las diferentes áreas de conocimiento de la ciencia, cada una con sus propias particularidades, añaden complejidad al problema, ya que incrementa en gran medida la casuística y dificulta la generalización. Por esta razón es necesario obtener y recolectar información de calidad, siendo la mejor manera de obtener esta información directamente a través de quienes la generan, es decir, las editoriales científicas.

Además de obtener la información, es igual o más importante tratarla correctamente. Cada fuente de datos tiene sus propias particularidades y nunca se proveen listos para ser utilizados. La falta de homogeneidad en las fuentes de datos dificulta en buena parte disponer de un procedimiento estándar y requiere estudiar, para cada caso, la manera correcta de proceder, teniendo en cuenta no solo las particularidades del conjunto de datos, sino también los posibles riesgos de privacidad, su tamaño, el objetivo con el que van a ser utilizados, etc. Es necesario, por tanto, apoyarse en técnicas de minería de datos para limpiar, estandarizar y estructurar los datos y convertirlos en información rica y explotable, así como técnicas de minería de textos para extraer de estos toda la información pertinente y necesaria para poder trabajar con ellos.

Haciéndose eco de esta necesidad, a mediados de septiembre de 2022 el International Center for the Study of Research (ICSR Lab)[7] de Elsevier sacó a la luz el nuevo Peer Review Workbench[8]. Una iniciativa que pretende poner a disposición de la comunidad científica un conjunto de datos de artículos enviados a las revistas de Elsevier entre 2018 y 2021 con información de más de 5 millones de autores, revisores y editores. Esta iniciativa, derivada en parte de las metodologías y flujos de trabajo empleados en las diferentes contribuciones de esta tesis doctoral, pone a disposición de la comunidad el mayor conjunto de datos del sistema de publicación científico compartido a hasta la fecha y marca un antes y un después en el acceso a este tipo de datos.

Concretamente, esta tesis presenta diferentes trabajos de caracterización del sistema de revisión por pares de artículos científicos, así como el flujo de trabajo y la aproximación metodológica de estos, empleando la minería de datos y de textos para el tratamiento de la información. Centrándose sobre todo en los textos de revisión, extrayendo características y métricas a nivel lingüístico, pero teniendo en cuenta también, las diferentes dimensiones y variables referentes a las personas involucradas, es decir, teniendo en cuenta los procesos sociales que intervienen.

1.2. Objetivos

El principal objetivo de esta tesis doctoral es estudiar el funcionamiento del sistema de revisión de artículos científicos a través de los datos que se

[7] https://www.elsevier.com/icsr/icsrlab

[8] https://lab.icsr.net/icsr_lab/workbenches.html

generan en las revistas a lo largo de este proceso. Se pretende estudiar, a través de la información que genera el propio proceso y de los textos de revisión que escriben los revisores y revisoras, diferentes métodos que permitan caracterizar los aspectos más importantes del proceso. A través de estas caracterizaciones se pretende generar métricas que permitan medir el funcionamiento, la calidad o la completitud de las revisiones de manera cuantitativa.

Se quiere caracterizar las revisiones desde diferentes puntos de vista, tanto desde las propiedades más formales del proceso, es decir, parámetros más puramente editoriales, como desde el punto de vista lingüístico, caracterizando el tipo de lenguaje empleado o de los puntos tratados en las revisiones. Pero también desde el punto de vista humano, teniendo en cuenta diferentes factores socio-demográficos. Para alcanzar este objetivo general, se proponen los siguientes objetivos específicos:

- Tratar, adecuar y procesar los datos disponibles y generar las estructuras de información adecuadas para su posterior explotación.
- Analizar las características lingüísticas de los textos de revisión generados durante el proceso de revisión por pares.
- Proponer métricas que permitan medir la calidad y/o completitud de las revisiones.
- Generar metodologías que permitan la extracción de características de las personas involucradas.
- Analizar las características propias del proceso de revisión.
- Estudiar los aspectos socio-demográficos en los que incurre el proceso a través de sus actores humanos y posibles sesgos.

Además, debido a la situación pandémica originada por la COVID-19 durante la realización de esta tesis doctoral, se añadió como objetivo estudiar el efecto de la misma sobre el sistema de publicaciones científico, su proceso de revisión y los posibles de género derivados de esta.

Estos objetivos se acometerán haciendo uso de técnicas de minería de datos para tratar toda la información necesaria y, concretamente, de técnicas de minería de textos para trabajar con la información textual extraída de los textos de revisión. Además, empleando los métodos de modelado estadístico pertinentes para realizar los análisis necesarios para cada uno de los estudios.

1.3. Organización de la tesis

La presente tesis se desarrolla en modalidad de compendio de artículos. Bajo esta premisa, este documento se ha estructurado en 3 partes.

- La primera corresponde con la parte introductoria y de contextualización del problema (capítulos 1 y 2), donde se introduce y exponen los motivos de la realización de este trabajo, así como el contexto en el que se desarrolla y los conceptos necesarios para abordar el problema.
- La segunda parte presenta, a modo de resumen, las contribuciones derivadas de esta tesis doctoral, sus conclusiones y líneas de trabajo futuro (capítulos 3 y 4).
- Por último, en los Anexos A, B, C, D y E se encuentran las versiones completas de los trabajos publicados en revistas.

En la redacción de esta tesis doctoral se ha intentado tratar de manera igualitaria y equilibrada a ambos sexos, haciendo uso, siempre que ha sido posible y que la concordancia y facilidad de lecturta del texto lo permitía, de las estrategias propuestas en las guías de la Universitat de València[9], la Universitat Autònoma de Barcelona[10] y la Xarxa Lluis Vives[11].

[9] https://www.uv.es/igualtat/GUIA/GUIA_CAS.pdf
[10] https://www.uab.cat/doc/llenguatge
[11] http://diposit.ub.edu/dspace/bitstream/2445/127832/4/criteris_multil.pdf

Capítulo 2

Contexto y conceptos previos

2.1. Contexto

El proceso de revisión por pares de artículos científicos es considerado una de las partes más importantes del sistema de publicación de la ciencia (Squazzoni et al., 2017a). Este es el principal encargado mejorar el contenido de los artículos enviados a publicación y velar por su calidad, así como ayudar a los equipos editoriales a tomar decisiones y filtrar aquellos artículos no aptos para su publicación, garantizando así, unos estándares de calidad en el contenido que publican las revistas científicas y aportando prestigio a las mismas (Bornmann, 2011).

Aunque esta tesis doctoral se centra en el estudio de la revisión por pares de artículos científicos, este proceso se utiliza también en muchos otros escenarios de la ciencia, como en la revisión de contribuciones a congresos, la valoración de ayudas de investigación pre o postdoctorales o la evaluación de propuestas de proyectos de investigación. Es, por tanto, un proceso ampliamente utilizado en el marco de trabajo de la ciencia y que ayuda a la toma de decisiones en muchos de sus ámbitos.

Existen diferentes modelos de revisión por pares, aunque generalmente, uno de los aspectos que los diferencia es en que forma se revela la identidad de sus participantes. Desde este punto de vista, existen 3 tipos de revisión por pares. En la revisión por pares de anonimización simple (*single-blind peer review*), los nombres de los revisores o revisoras no son conocidos, en cambio, si se revelan los nombres de los autoras o autores del artículo. En la anonimización doble (*double-blind peer review*), ninguno de los dos colectivos conoce el nombre del otro (Martín, 2016). En cambio, en la revisión por pares abierta (*open peer review*) no se aplica ningún tipo de anonimización.

Cuando un autor, o grupo de autores, envía un artículo a una revista para su publicación, este llega a manos de algún miembro de su equipo editorial. Estos tienen la labor de tomar una primera decisión, si el artículo no es de interés para el ámbito de la revista o no presenta la suficiente calidad para sus estándares, se rechazará directamente el artículo. Muy rara vez, el artículo se aceptará directamente si este cumple con todos los requisitos establecidos por el equipo editorial. En cualquier otro caso, el artículo se enviará a revisión.

En este proceso, el editor selecciona, generalmente 3 personas, a quienes envía el artículo para que lo revisen. Los revisores, emitirán una recomendación que generalmente puede ser; aceptar, requiere cambios menores, requiere cambios mayores o rechazar. Junto con esta recomendación, el se emiten también una serie de comentarios dirigidos a los autores, en los que valora diferentes aspectos del artículo y recomienda los cambios oportunos para mejorar la calidad del mismo. El editor recopila las recomendaciones y los comentarios de los revisores y emite una decisión. Se trata de un proceso iterativo, que se repite hasta alcanzar una unanimidad en la recomendación por parte de los revisores o hasta que el editor tiene información suficiente para tomar una decisión.

Como ya se ha comentado en el capítulo anterior, un problema fundamental a la hora de estudiar el proceso de revisión es el acceso a los datos. Al tratarse de un proceso editorial, son datos que se generan en el seno de la revista científica, por tanto, el acceso a los mismos depende de estas. Existen varias iniciativas de open peer review que pretenden hacer accesibles estos datos y dejar patente la traza de revisión (Ross-Hellauer, 2017), donde tanto los textos de revisión, como todas recomendaciones y decisiones tomadas sobre el artículo, se reflejen de manera pública. Aunque se trata de iniciativas activas en estos momentos, son las revistas quienes tienen que decidir finalmente adoptarlas y suelen requerir ciertas adaptaciones a nivel técnico para la publicación de estos datos junto con la versión final publicada del artículo. Con esto se busca favorecer la transparencia del propio proceso de revisión por pares. En la práctica, son pocas las revistas que adoptan políticas de open peer review, de hecho, de las 3700 revistas registradas en la plataforma Publons, se estima que solo el 3.5 % de estas permite a los revisores firmar sus revisiones y solamente el 2.3 % permite hacer públicas las revisiones (Wilkinson, 2017). Por lo tanto, las bases de datos de open peer review disponibles son escasas y con poca cantidad de registros. Además, existe cierta controversia en la comunidad científica acerca de los beneficios del open peer review sobre el procedimiento tradicional no abierto (Groves, 2010)(Khan, 2010). Esto hace que, habitualmente, los estudios que se publican sobre este tema se basen en conjuntos de datos con cientos o pocos miles de registros.

Bajo esta tesitura, algunos investigadores e investigadoras tratan de cerrar acuerdos de compartición de datos con revistas y/o editoriales científicas que les faciliten el acceso a sus datos para realizar estudios. Es el caso, por ejemplo, de la COST Action New frontiers on Peer Review, ya comentada en el capítulo anterior. Concretamente los datos utilizados para los estudios realizados en el marco de esta tesis doctoral provienen de la citada COST Action y de diversos acuerdos de compartición de datos firmados con la casa editorial Elsevier, quienes además, facilitaron el acceso a la base de datos de Scopus para realizar búsquedas de datos a nivel individual sobre los autores y revisores implicados. Todo esto da lugar a un conjunto de datos único que no ha sido explotado ni estudiado por nadie hasta la fecha.

En este proceso no solo intervienen aspectos científicos. Al tratarse de actores humanos, intervienen multitud de interacciones sociales que se manifiestan en una negociación entre autores y revisores. Estas interacciones, conocidos como los procesos sociales de la ciencia (Edmonds et al., 2011), son de vital importancia para entender el comportamiento de los diferentes grupos poblacionales a nivel geográfico, cultural, biológico y social. Estudiar estos procesos sociales requiere ser capaz de caracterizar correctamente los actores que en ellos intervienen, obteniendo datos como su género, su país, su edad, su estatus académico o científico, etc. Para ello, es necesario obtener datos de calidad y lo más completos posible, incluyendo datos personales, como nombres o correos electrónicos, que permitan enriquecerlos mediante bases de datos externas.

No obstante, no se trata solo de obtener datos, estos tienen que ser tratados de la manera correcta. Los datos recopilados para la realización de esta tesis doctoral son datos en bruto, en la mayoría de casos, volcados directamente de los sistemas de gestión editorial y no han sido tratados ni trabajados con anterioridad. Esto supone un reto en sí mismo, que es necesario abordar desde un punto de vista de la minería de datos, limpiando, estandarizando, estructurando y enriqueciendo los conjuntos de datos originales, para convertirlos en estructuras de información de las que poder extraer las características necesarias para la caracterización del sistema de revisión y generar métricas.

2.2. Metodología

En la presente sección se nombran y explican diferentes técnicas y metodologías empleadas en los trabajos de esta tesis doctoral. La mayoría de ellas, englobadas dentro de la minería de datos, que constituye un proceso y conjunto de técnicas para la extracción, transformación y estructuración de conjuntos de datos, pero sobre todo, que provee de los mecanismos para el descubrimiento de información, es decir, la capacidad de realizar las transformaciones adecuadas para descubrir información que podría parecer que no estaba inicialmente (Maimon and Rokach, 2010). Cuando estas técnicas se aplican sobre textos, comúnmente pasan a recibir el nombre de minería de textos, que es el conjunto de técnicas englobadas en el área de la minería de datos enfocadas a trabajar con textos en vez de con datos estructurados (Jo, 2019). Normalmente se aplican, además de los métodos necesarios para pre-procesar, limpiar y estandarizar el texto para adecuarlo al uso que se le vaya a dar, técnicas de procesado de lenguaje natural que permiten extraer información sobre los textos que sirve a su vez para caracterizarlos.

2.2.1. Recolección, limpieza y estandarización

Una de las tareas más transversales en cualquier trabajo o proyecto con datos, sea cual sea su índole u objetivo, es la limpieza y estandarización. Generalmente los datos en bruto suelen presentar multitud de factores que los hacen que dificultan su uso o los hacen directamente inservibles en su estado original (datos no estructurados o en bruto, no homogeneizados, sin categorizar, etc.) (Luengo et al., 2006). Esta tarea, según el último informe de la empresa Anaconda[1] (Anaconda, 2022), consume un 38 % del tiempo que un científico de datos dedica a cualquier proyecto, convirtiéndola con creces, en la tarea que más tiempo consume del marco de trabajo de la ciencia de datos.

Durante este proceso, es común encontrarse con datos en diferentes soportes, formatos o plataformas, que es necesario extraer y procesar para obtener su información. Esta parte se acentúa considerablemente cuando se trabaja con textos. Los textos pueden presentarse en multitud de formas, no es lo mismo procesar archivos de texto plano, que archivos PDF, DOC/DOCX, archivos Latex o incluso imágenes, por ejemplo en el caso de textos escaneados. Así pues, es habitual enfrentarse a una falta de homogeneidad de formatos, que requiere un tratamiento específico en función a estos.

Otro paso importante directamente relacionado es cómo estructurar y almacenar la información. En función de que técnicas, análisis o aplicaciones vayan a tratar los datos será conveniente almacenarlos de una u otra manera. En este caso, lo más común es transformarlos a un formato tabular, en el que poder almacenar tanto los textos como todos sus datos asociados. Otra manera obvia de almacenarlos sería por medio de algún motor de bases de datos, ya sea relacional o no relacional. No obstante, sea cual sea el formato elegido, siempre estará guiado por su aplicación posterior (García et al., 2006).

Cada aplicación es tiene sus propias características y complejidades que dificultan la definición de un método de limpieza universal. Elegir si es necesario mantener signos de puntuación, mayúsculas, diacríticos, mantener las palabras completas o eliminar sufijos (*word stemming*) o quedarse con su lema principal (*lemmatization*), todos estos elementos serán necesarios o no en función del tipo de técnicas que se vayan a aplicar. Por ejemplo, no es lo mismo trabajar con textos procedentes de tweets, donde es importante tratar el significado de los emojis o los hashtags, que trabajar con textos científicos, donde no existen este tipo de elementos. No obstante es recomendable eliminar marcas indeseadas, muchas veces producidas por conversiones de formato, caracteres u otros elementos que no aporten información textual, espacios o saltos de línea repetidos, etc.

Por último, para todas aquellas variables asociadas al texto, es deseable estandarizar su contenido. sobre todo en el caso de variables categóricas, para no tener diferentes valores que respondan a un mismo significado o un mismo grupo. Este proceso suele ser mayormente manual, aunque es habitual ayudarse de herramientas como las expresiones regulares, que permiten buscar y convertir patrones a un mismo valor.

En el caso concreto de esta tesis doctoral, para cada conjunto de datos utilizado y en función del estudio a realizar, se emplearon unos u otros mecanismos de limpieza y estandarización. Por ejemplo, en el caso de los datos procedentes de Royal Society, los textos de artículos y de revisión se

[1] https://www.anaconda.com/

encontraban en ficheros de textos en múltiples formatos, lo que requirió un procesado muy metódico, basado en detectar cada uno de los formatos y ejecutar, en cada caso, el procedimiento más adecuado para trabajar con él. Siempre con el objetivo de estandarizar todos los textos a un mismo formato y generar una única estructura de datos con la que trabajar. En otros casos, como en los conjuntos de datos procedentes de Elsevier o de Open Research Central, la información se extrajo de diferentes ficheros JSON y/o XML.

Independientemente de la procedencia y del formato de los datos, el objetivo en esta línea fue generar estructuras de datos lo más homogéneas posibles, estandarizando la información siempre de la misma manera y con la misma nomenclatura para generar estructuras lo más claras e interpretables posible que facilitaran su manipulación e incluso su integración.

2.2.2. Enriquecimiento de datos

El enriquecimiento de datos es un proceso bastante habitual a la hora de trabajar con datos. Suele ser común necesitar añadir información adicional, que complemente o mejore la ya existente, aportando así mayor riqueza al conjunto. En estos casos es necesario identificar qué información puede ser de utilidad para el trabajo concreto a realizar, buscar la fuente correcta de la que extraer esa información y analizar la manera óptima de integrarla con los datos existentes.

La alta disponibilidad hoy en día de APIs de todo tipo, que permiten realizar consultas y extraer información resulta extremadamente útil a la hora de realizar esta tarea y facilita muchísimo el acceso a la información. Por ejemplo, en el caso de los estudios realizados en el marco de esta tesis doctoral, se añadieron datos como el área de conocimiento, el factor de impacto y el cuartil JCR de las revistas, a través de las aplicaciones web del Journal Citation Reports de Clarivate[2].

Otra fuente muy utilizada en esta tesis doctoral fue la base de datos de Scopus [3]. A través del ICSR Lab, se obtuvo acceso a sus bases de datos y se realizaron consultas para extraer información de millones de científicos y científicas. De ahí se extrajeron, por ejemplo, la fecha de su primera publicación para estimar su antigüedad académica, su número de citas o su H-Index.

Otro tema muy interesante en este contexto es la inferencia de información a partir de otra existente, es el caso, por ejemplo, del género de las personas. Existen multitud de librerías y APIs que permiten la inferencia del género a partir del nombre, apellidos y el país de procedencia de la persona. En este caso concreto, se empleó una combinación de la librería para Python, Gender-Gesser[4], que tiene un error menor del 3 % (Santamaría

[2] https://jcr.clarivate.com/
[3] https://www.scopus.com/
[4] https://pypi.org/project/gender-guesser/

and Mihaljević, 2018) y de GenderAPI[5] que tiene un error de clasificación menor del 5 % (Santamaría and Mihaljević, 2018). Así pues, si la primera es capaz de inferir un género inequívocamente (sin determinarlo como mayormente hombre, mayormente mujer o indeterminado), se da preferencia a esta y en cualquier otro caso, se pasa esa combinación de nombre y país por GenderAPI. Este procedimiento permite añadir información de género a un conjunto de datos que inicialmente no disponía de ella, con un margen de error muy aceptable.

2.2.3. Anonimización y minimización

La anonimización es imprescindible cuando se trabaja con datos personales, confidenciales o con ciertas limitaciones establecidas por el propietario de los datos. No existe una manera universal y concreta para anonimizar datos, no obstante si existen ciertas premisas y técnicas que se suelen aplicar, pero teniendo siempre en cuenta que cada conjunto de datos es único y requiere un tratamiento concreto y específico ajustado a la naturaleza de los mismos (Murthy et al., 2019).

Un elemento comúnmente utilizado con este fin son los *tokens*. Estos se usan para substituir ciertos aspectos sensibles, como por ejemplo nombres de personas, de compañías o de países, URLs o direcciones de correo electrónico. Se pueden emplear técnicas de etiquetado gramatical (*Part of Speech Tagging o POS Tagging*) o puede realizarse de manera manual, dependiendo de lo que se pretenda substituir.

Por otro lado, es habitual aplicar secreto estadístico sobre el conjunto de datos. Aunque tampoco existe una fórmula mágica para su aplicación y es necesario establecer ciertos criterios manualmente. En este caso es necesario detectar grupos minoritarios, que puedan ser identificables en un contexto concreto. Por ejemplo, si en la muestra poblacional con la que se está trabajando solo existen dos mujeres y estas pertenecen a un colectivo concreto fácilmente identificable, será necesario excluirlas del estudio para evitar su identificación.

Relacionada con la anonimización esta también la minimización de los datos. A diferencia de la anonimización, esta se centra en identificar exactamente que parte del conjunto de datos son necesarias y que partes no lo son. Minimizar un conjunto de datos consiste en generar un nuevo conjunto de datos que contenga únicamente aquella información relevante y estrictamente necesaria para llevar a cabo de reproducir un estudio concreto garantizando en todo momento que no se expone información innecesaria (Goldsteen et al., 2011) (Biega et al., 2020).

Para esta tesis doctoral se han seguido estrictas políticas de anonimización y minimización de datos, garantizando en todo momento que ninguna entrada de datos fuera identificable, ya sea por aspectos humanos, como el género, la afiliación o el rol de las personas, por aspectos propios de los artículos o por aspectos editoriales propios de las revistas. Los datos compartidos para la reproducibilidad de cada uno de los estudios

[5] https://gender-api.com

se minimizó a la información estrictamente necesaria, garantizando así no solo el anonimato, sino también claridad y sencillez en los conjuntos de datos.

2.2.4. Extracción de características de textos

Históricamente se han utilizado diccionarios de términos para la búsqueda de palabras específicas sobre los textos para extraer información de los mismos. A día de hoy, con la proliferación de la inteligencia artificial y, concretamente, los modelos de lenguaje computacionales y las técnicas de procesado de lenguaje natural, se han extendido una gran cantidad de métodos automatizados, basados en aprendizaje máquina, para la extracción de características e identificación de propiedades del texto (Jones, 1994).

Pese a poder parecer más arcaicas por su antigüedad, las técnicas basadas en diccionarios son aún ampliamente utilizadas en áreas como la sociología o la psicología, ya que existen diccionarios validados y con muy buena reputación. En este contexto, también existen multitud de librerías, por ejemplo, de análisis de sentimiento, que basan su funcionamiento en diccionarios de términos (Cook and Jensen, 2019). No obstante, el problema radica en encontrar un diccionario válido para la tarea en cuestión que se quiera realizar. Es probable no encontrar ninguno que satisfaga las necesidades o que no esté validado o lo suficientemente probado como para garantizar su buen funcionamiento.

Existen diferentes maneras de generar un diccionario para una tarea concreta. La primera y más obvia es recurrir a expertas y expertos en la materia para confeccionar, de manera totalmente manual, una lista de términos relacionados a aquello que se quiere medir. Esta aproximación requiere un alto conocimiento de la materia y, en muchos casos, la necesidad de trabajar conjuntamente con lingüistas para garantizar la máxima especificidad posible.

Otra aproximación es mediante la automatización del proceso. Haciendo uso de los propios textos que se van a analizar, es posible crear un modelo de lenguaje computacional sobre el que extraer los términos, seleccionando términos similares o cercanos en el espacio vectorial del modelo que tienen significado similar a la palabra en cuestión. Esta aproximación requiere poco conocimiento sobre la materia ya que los términos se seleccionan basándose en el modelo generado, partiendo de una pequeña lista de palabras inicial creada manualmente y estableciendo umbrales de tolerancia para decidir con que términos quedarse. No obstante, una aproximación totalmente automática puede dar lugar a una selección poco precisa o incorrecta, por lo tanto siempre es necesario y recomendable revisar los términos seleccionados por el algoritmo. Los diccionarios creados a partir de estas técnicas automatizadas han reportado niveles de robustez y resultados similares a los diccionarios confeccionados manualmente (Muresan and Klavans, 2002) (Godbole et al., 2010) (Deng et al., 2017) (Mpouli et al., 2020).

Teniendo esto en cuenta, una solución intermedia es la generación semiautomática. En este caso, la selección se realiza sobre un modelo de lenguaje, al igual que en la generación automática, pero de manera iterativa. De modo que, en cada iteración, un conjunto de humanos valida los términos seleccionados por el algoritmo, eliminando aquellos incorrectos, poco claros o poco específicos. La validación puede apoyarse también en técnicas como el *Key Word in Context*, para verificar el contexto en el que se utiliza una determinada palabra en un corpus de textos concreto. La salida de cada iteración es utilizada como entrada en las iteraciones posteriores y se repite el proceso hasta obtener tolerancias fuera del umbral definido. En la figura 2.1 se puede observar un ejemplo del flujo de este procedimiento, utilizado en varias de las contribuciones de esta tesis doctoral. Los puntos 1 y 2 corresponden a las fases de estandarización y limpieza de los datos, sobre el conjunto resultante de este proceso se crea un modelo de *Word Embeddings* (punto 3) que posteriormente, en el punto 4, se utiliza para la extracción de términos similares partiendo de la lista inicial. En el punto 5, se validan los términos extraídos de manera manual y se vuelve a iterar sobre el punto 4 hasta obtener precisiones fuera del umbral establecido.

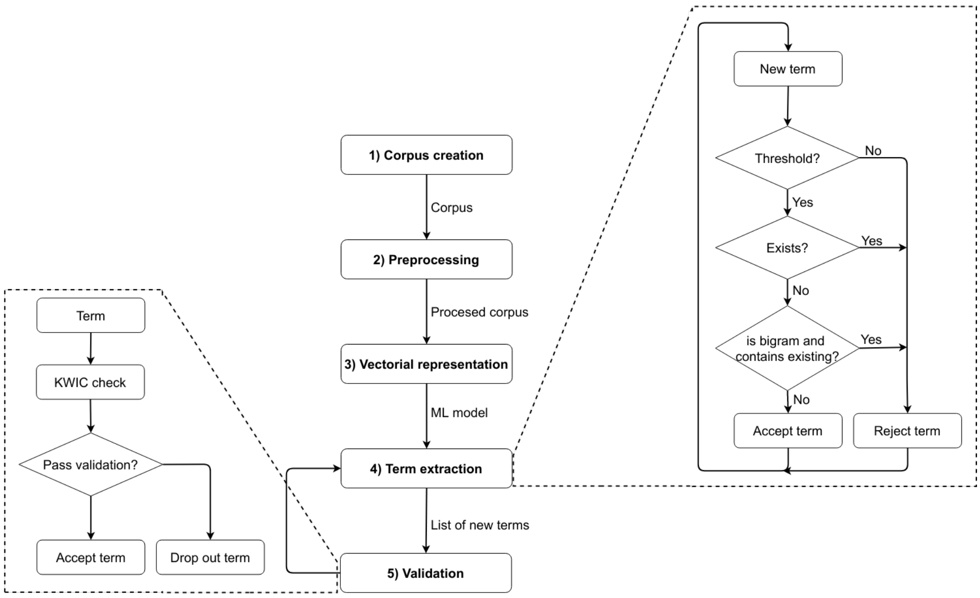

Figura 2.1: Diagrama de flujo del proceso de creación de un diccionario de manera semiautomática

En el otro extremo se encuentran las técnicas de caracterización basadas en redes neuronales. Estas técnicas se basan en extraer vectores de características (*embeddings*), que generalmente pueden ser de diferentes tipos, en función de en que características se centren para representar el texto (Taher Pilehvar and Camacho-Collados, 2021). Estos vectores de características se pueden utilizar con infinidad de objetivos, por ejemplo, para clasificar, agrupar o identificar aquellos que presentan una de terminada serie de propiedades.

Aunque estos modelos se pueden entrenar desde cero con un corpus de textos propios, en la actualidad existen modelos lingüísticos computacionales creados con billones de parámetros que han sido entrenados con millones de textos como el tradicional Word2Vec (Mikolov et al., 2013) o Sentence BERT (Reimers and Gurevych, 2019), centrados en la representación del texto como vectores numéricos o modelos más generales, pero que también pueden utilizarse para caracterización como GPT3 (Brown et al., 2020) de Open AI[6] o el recientemente liberado por BigScience[7], BLOOM[8].

No obstante, estos grandes modelos de lenguaje, conocidos como LLMs (*Large Language Models*), están entrenados de manera no supervisada con ingentes cantidades de datos y pese a que son muy útiles y precisos para algunas tareas pueden resultar poco útiles para según qué objetivos. Es por esto que habitualmente se recurre a realizar un ajuste fino de los pesos de estos modelos, entrenando sobre el conjunto de textos con el que se va a trabajar. De este modo, se aprovechan las características ya aprendidas y se ajustan sus pesos a un conjunto específico, es decir, se ajusta el modelo para funcionar mejor con los datos de los que se dispone (Ziegler et al., 2019). Esto es común sobre todo en aplicaciones muy específicas que usan a su vez lenguaje muy concreto, como por ejemplo, el lenguaje científico, que nada tiene que ver con el lenguaje literario o con el que se usa en la prensa escrita.

Otra aplicación importante de este tipo de técnicas son las centradas en el etiquetado de lenguaje, conocido en inglés por el acrónimo POS (Part of Speech). Estos modelos son entrenados para clasificar cada una de las palabras de un texto dado y etiquetarlas según su tipología sintáctica o gramatical. Son extremadamente útiles, por ejemplo, a la hora de identificar nombres propios, de empresas, países, para etiquetar ciertos elementos que se encuentran en el texto o, incluso, para identificar elementos sintácticos como verbos, adjetivos, etc.

Algunos de estos modelos de *embeddings* se han utilizado en varios trabajos de esta tesis doctoral, en algunos casos, para la propia caracterización del texto y en otros, para realizar las búsquedas de términos necesarias para la generación semiautomática de diccionarios.

A continuación se detallan las contribuciones y aportaciones realizadas en el marco de esta tesis doctoral, que emplean la metodología y las técnicas detalladas anteriormente.

[6] https://openai.com/blog/openai-api/
[7] https://bigscience.huggingface.co/
[8] https://bigscience.huggingface.co/blog/bloom

Capítulo 3

Contribuciones y resultados derivados de la tesis

El presente capítulo detalla los diferentes trabajos realizados durante el desarrollo de esta tesis doctoral, desde publicaciones en revistas y aportaciones a congresos hasta registros de propiedad intelectual o contratos de transferencia con empresas.

3.1. Publicaciones en revistas

- Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Flaminio Squazzoni, Ana Marušić
  Meta-Research: Large-scale language analysis of peer review reports.
  Año de publicación: 2020
  Revista: eLife, 9:e53249
  Factor de impacto JCR: 8.713 (Q1)
  DOI: 10.7554/eLife.53249
  (Apéndice A)

- Flaminio Squazzoni, Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa, Mike Farjam, Bahar Mehmani
  Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals
  Año de publicación: 2021
  Revista: PLoS One, 16(10): e0257919
  Factor de impacto JCR: 3.752 (Q2)
  DOI: 10.1371/journal.pone.0257919
  (Apéndice E)

- Daniel Garcia-Costa, Flaminio Squazzoni, Bahar Mehmani, Francisco Grimaldo
  Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals
  Año de publicación: 2022
  Revista: PeerJ, 10:e13539

Factor de impacto JCR: 3.061 (Q2)
DOI: 10.7717/peerj.13539
(Apéndice D)

- Daniel Garcia-Costa, Anabel Forte, Emilia López-Iñesta, Flaminio Squazzoni, Francisco Grimaldo
  Does peer review improve the statistical content of manuscripts? A study on 27,467 submissions to four journals
  Año de publicación: 2022
  Revista: Royal Society Open Science, Volume 9, Number 9, 210681
  Factor de impacto JCR: 3.653 (Q2)
  DOI: 10.1098/rsos.210681
  (Apéndice C)

- Federico Bianchi, Daniel García-Costa, Francisco Grimaldo, Flaminio Squazzoni
  Measuring the effect of reviewers on manuscript change: A study on a sample of submissions to Royal Society journals (2006–2017)
  Año de publicación: 2022
  Revista: Journal of Informetrics, Volume 16, Issue 3, 101316
  Factor de impacto JCR: 4.373 (Q2)
  DOI: 10.1016/j.joi.2022.101316
  (Apéndice B)

3.2. Contribuciones a congresos

- Ivan Buljan, Daniel Garcia-Costa, Flamninio Squazzoni, Francisco Grimaldo, Ana Marušić
  Are reviewers too subjective? A large scale language analysis of peer review reports
  REWARD|EQUATOR Conference, Berlín (Alemania), 2020
  Libro de actas del congreso

- Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Flaminio Squazzoni, Ana Marušić
  Large-scale language analysis of peer review reports
  PEERE International Conference on Peer Review, Valencia (Spain), 2020
  DOI: 10.48448/8ym3-tp63

- Federico Bianchi, Daniel Garcia-Costa, Francisco Grimaldo, Flaminio Squazzoni
  Peer Review Improves Manuscripts of Moderate Initial Quality. A Study on Ten Journals from the Royal Society (2006-2017)
  PEERE International Conference on Peer Review, Valencia (Spain), 2020
  DOI: 10.48448/krz2-cn12

- Anabel Forte, Daniel Garcia-Costa, Emilia López-Iñesta, Phil Hurst, Flaminio Squazzoni, Francisco Grimaldo

Change in statistical terms in peer-reviewed journals
PEERE International Conference on Peer Review, Valencia (Spain), 2020
DOI: 10.48448/tk7r-h687

- Andrijana Perković Paloš, Antonija Mijatović, Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Ana Marušić
  Linguistic and semantic characteristics of articles and peer review reports in social and medical sciences: analysis of articles published in Open Research Central
  9th Conference on Scholarly Communication in the Context of Open Science (PUBMET), 2022, Zadar (Croatia), 2022

- Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Richard Klein, Marjan Bakker, Ana Marušić
  Development of a List to Detect Statistical and Methodological Terms in Peer Reviews
  International Congress on Peer Review and Scientific Publication, Chicago, (United States), 2022

- Mario Malički, Taym Alsalti, Daniel García-Costa, Francisco Grimaldo, Elena Álvarez-García, Ana Jerončić, Steven M. Goodman, Flaminio Squazzoni, Bahar Mehmani
  Unprofessional Comments in Peer Review Reports Across Scholarly Disciplines
  International Congress on Peer Review and Scientific Publication, Chicago, (United States), 2022

- Daniel Garcia-Costa, Francisco Grimalo, Emilio Soria-Olivas, Rafael Magdalena, Joan Vila
  NLP challenges and solutions in Science of Science
  International Conference of the Catalan Association for Artificial Intelligence (CCIA), Lleida (Spain), 2021

3.3. Estructuras de información

Uno de los objetivos de esta tesis doctoral es el tratamiento, adecuación y procesado de las diferentes fuentes de datos para generar las estructuras de información adecuadas y permitir su explotación para el estudio del sistema de revisión por pares.

Los estudios realizados parten, principalmente, de 4 conjuntos de datos.

- Royal Society. Procedente de la compartición de datos realizada siguiendo el protocolo de compartición de datos PEERE (Squazzoni et al., 2017b), contiene información de revisiones y artículos enviados a las revistas de Royal Society desde 2006 hasta 2017. Incluye información editorial sobre los envíos, textos de revisión, textos de los artículos, e información personal sobre las personas involucradas.

- Elsevier. Procedente de la compartición de datos realizada siguiendo el protocolo de compartición de datos PEERE (Squazzoni et al., 2017b), contiene información sobre revisiones realizadas en más de 60 revistas de Elsevier, incluyendo información editorial sobre los envíos, textos de revisión e información personal sobre las personas.
- Elsevier. Procedente de compartición de datos detallada en el punto Contratos de transferencia. Este conjunto contiene datos sobre el proceso editorial, textos de revisión, e información personal de más de 11 millones de autores y autoras y 7 millones de artículos enviados a más de 2300 revistas.
- Open Research Central. Descargados directamente desde su repositorio[1]. Al tratarse de un proceso de revisión abierto (open peer review), se puede descargar toda la traza de cada uno de los artículos, incluyendo texto de los mismos, textos de revisión y nombres de las personas involucradas.

Estos conjuntos de datos en bruto se presentan almacenados en diferentes formatos, desde archivos estructurados como XML o JSON hasta archivos de texto con formato enriquecido como DOCX u otros archivos de almacenamiento de texto como PDF (primer nivel en la figura 3.1). Se trata, en total, de más de 500GB de archivos de datos que fueron procesados siguiendo la metodología expuesta en el capítulo anterior.

Como se puede observar en la figura 3.1, primero se procedió a extraer toda la información de los archivos de datos en bruto, extrayendo el texto de los textos de revisión, la información referente a tablas y figuras y construyendo las diferentes variables que almacenan nombres, apellidos, afiliaciones, etc. Acto seguido, se limpiaron y estandarizaron todos estos datos. En este paso se preprocesó el texto, se estandarizaron las diferentes variables, se enlazaron las personas para poder identificarlas a lo largo de todo el conjunto de datos y se propagó su información para rellenar aquellos campos que en algunos casos aparecieran vacíos. Una vez extraída y preprocesada la información, se enriqueció con fuentes externas para obtener el género de las personas, su edad o estatus académico y diferente información académica e investigadora como el H-Index, el número de publicaciones, etc.

Después de realizar todas estas transformaciones se crearon y almacenaron las estructuras de datos que contienen toda la información extraída y procesada, con las que posteriormente, se construyen y anonimizan los diferentes conjuntos de datos empleados para cada uno de los estudios, en función de las variables necesarias para llevar a cabo los análisis pertinentes. Todos los conjuntos de datos generados para los estudios de esta tesis doctoral se han publicado en diferentes repositorios de datos y son accesibles para la comunidad investigadora, estos son:

- Replication Data for: Does peer review improve the statistical content of manuscripts? A study on 27,467 manuscripts submitted to four journals. DOI: 10.7910/DVN/MOKJED

[1] https://openresearchcentral.org/browse/articles

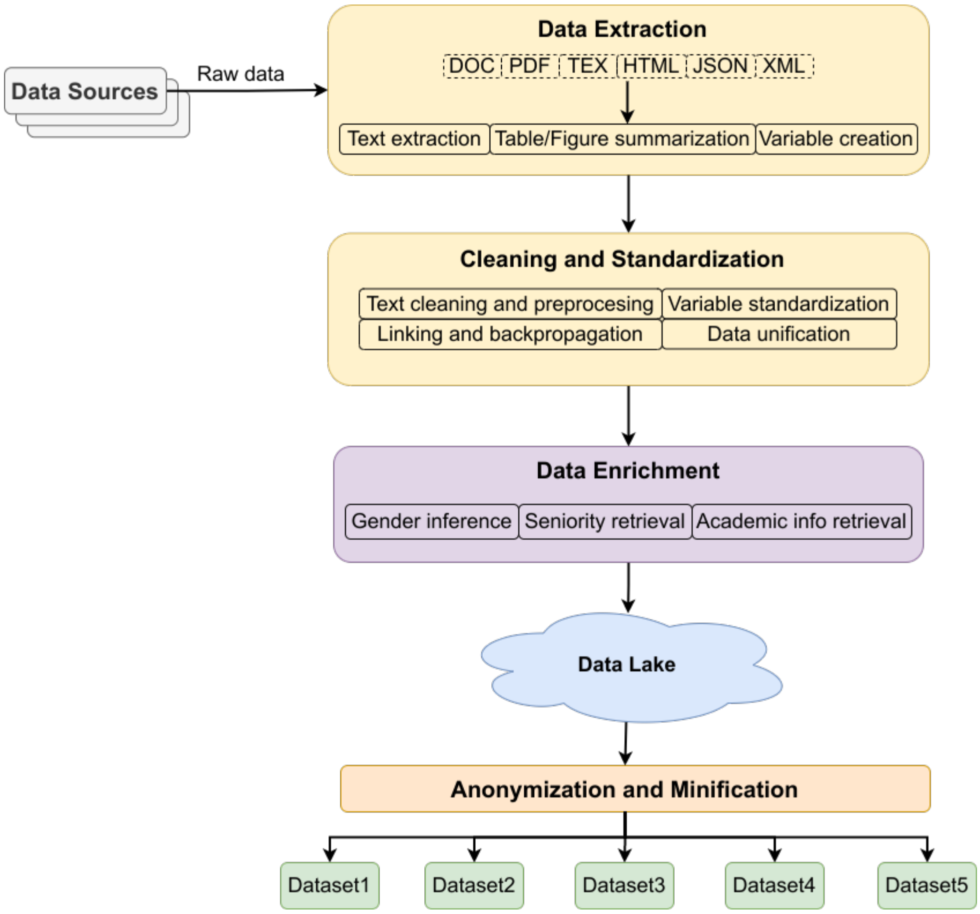

Figura 3.1: Flujo de trabajo y metodología empleada para el tratamiento de los datos

- Replication Data for: Measuring the effect of reviewers on manuscript change: A study on a sample of submissions to Royal Society journals (2006–2017). DOI: 10.7910/DVN/WHKULA

- Replication data for: Measuring the developmental function of peer review: A multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. DOI: 10.7910/DVN/D96G2I

- Replication data for: No tickets for women in the COVID-19 race? A study on manuscript submissions and reviews in 2329 Elsevier journals during the pandemic. 10.7910/DVN/S0T7Z5

Todo el conocimiento adquirido trabajando con esta información también se ve reflejado en la creación, por parte de Elsevier del Peer Review Workbench[2], quienes ponen a disposición de la comunidad investigadora un conjunto de datos de millones de registros, tratados y estructurados siguiendo la metodología definida en esta tesis doctoral.

[2] https://lab.icsr.net/icsr_lab/workbenches.html

3.4. Contratos de transferencia

- **Elsevier data sharing agreement**. Contrato de compartición de datos firmado con la editorial científica Elsevier, en el que se compartieron datos de más de 10 millones de artículos y 1.7 millones de textos de revisión. Gracias a este acuerdo de compartición de datos se pudieron llevar a cabo algunos de los trabajos presentados en esta tesis doctoral. Este contrato fue desarrollado entre el 30/05/2020 y el 1/12/2022, actuando el autor de esta tesis doctoral como miembro del equipo investigador.

- **ReviewerCredits**. Contrato firmado con la empresa ReviewerCredits [3] con el fin de realizar la implementación del RCI (*Reviewer Contribution Index*)[4], como una integración del índice F3Index (Bianchi et al., 2019) en su plataforma. Este contrato se desarrolló entre el 06/10/2021 y el 19/11/2021, por un importe de 4344.90€, actuando el autor de esta tesis doctoral como investigador principal.

3.5. Registros de propiedad intelectual

- **Review Metrics**. A raíz de la publicación *Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals* en la revista PeerJ, se desarrolló la aplicación Review Metrics, que permite medir la completitud de un texto de revisión a través de una interfaz web donde, introduciendo el texto de revisión, y mediante el diccionario generado, muestra indicadores en las 10 dimensiones medidas en el artículo y la puntuación global obtenida, así como una comparativa respecto al resto de textos de revisión disponibles en la plataforma.

3.6. Contribuciones

Las publicaciones derivadas de este trabajo pueden agruparse en cuatro contribuciones principales, como se muestra en esquema de la figura 3.2. En él se pueden observar los principales conceptos abordados (bloques morados), algunas de las técnicas o métodos utilizados (bloques amarillos) y cómo se relacionan o interactúan entre ellas las diferentes partes involucradas. De igual modo, las publicaciones en revista se muestran de color azul oscuro, mientras que las contribuciones a congreso están representadas en color azul claro.

[3] https://www.reviewercredits.com/

[4] https://www.reviewercredits.com/reviewer-contribution-f3-index/

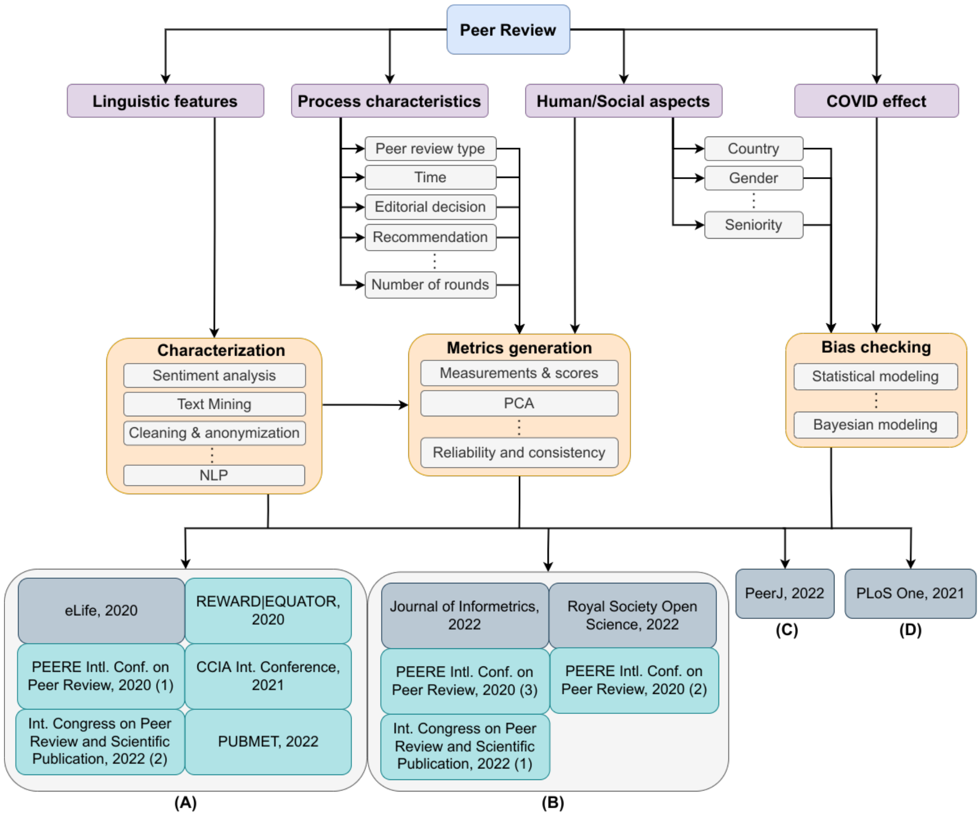

Figura 3.2: Resumen conceptual de los trabajos abordados en la tesis doctoral

3.6.1. Contribución A

El primer bloque de contribuciones (bloque A en la figura 3.2) corresponden a aquellas más centradas en el estudio de las características lingüísticas de los textos de revisión. Este grupo de contribuciones está formado por los siguientes artículos y contribuciones a congresos.

- Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Flaminio Squazzoni, Ana Marušić. Meta-Research: Large-scale language analysis of peer review reports. eLife, 2020, 9:e53249. (Apéndice A)

- Ivan Buljan, Daniel Garcia-Costa, Flamninio Squazzoni, Francisco Grimaldo, Ana Marušić. Are reviewers too subjective? A large scale language analysis of peer review reports. REWARD|EQUATOR Conference, Berlín (Alemania), 2020.

- Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Flaminio Squazzoni, Ana Marušić. Large-scale language analysis of peer review reports. PEERE International Conference on Peer Review, Valencia (Spain), 2020.

- Andrijana Perković Paloš, Antonija Mijatović, Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Ana Marušić. Linguistic and semantic characteristics of articles and peer review reports in social and medical sciences: analysis of articles published in Open Research

Central. 9th Conference on Scholarly Communication in the Context of Open Science (PUBMET), 2022, Zadar (Croatia), 2022

- Mario Malički, Taym Alsalti, Daniel García-Costa, Francisco Grimaldo, Elena Álvarez-García, Ana Jerončić, Steven M. Goodman, Flaminio Squazzoni, Bahar M. Unprofessional Comments in Peer Review Reports Across Scholarly Disciplines. International Congress on Peer Review and Scientific Publication, Chicago, (United States), 2022

- Daniel Garcia-Costa, Francisco Grimalo, Emilio Soria-Olivas, Rafael Magdalena, Joan Vila. NLP challenges and solutions in Science of Science. International Conference of the Catalan Association for Artificial Intelligence (CCIA), Lleida (Spain), 2021.

En estos trabajos se plantean diferentes métodos para la caracterización de los textos de revisión, centrados en la extracción y análisis de sus características lingüísticas. Dicha extracción se realiza mediante el uso de técnicas de procesamiento de lenguaje natural, como el análisis de sentimientos, análisis de emociones, detección de odio o la aplicación de diferentes diccionarios para detectar términos concretos e identificar el tipo de lenguaje empleado.

Concretamente, la publicación *Meta-Research: Large-scale language analysis of peer review reports* publicada en la revista *eLife* y su trabajo previo presentado en el *PEERE Intl. Conference on Peer Review*, se centran en la extracción de características referentes al tono analítico, la autenticidad, la influencia, el sentimiento y los aspectos morales de los textos de revisión. Estas características se utilizan para estudiar que particularidades tienen los textos de revisión en función a la recomendación del revisor o revisora, el área de conocimiento a la que pertenece la revista, el tipo de revisión por pares que implementa la revista y el género de la persona que revisa. El análisis se realiza sobre un total de 472.449 textos de revisión, que pertenecen a un total de 61 revistas de la casa editorial Elsevier. De estas 61 revistas, 22 pertenecen al área de ciencias médicas y de la salud, 5 de ellas a ciencias de la vida, 30 al área de la física y 4 a ciencias sociales y económicas.

Para la caracterización de los textos de revisión, se emplearon diferentes técnicas, por un lado, técnicas más tradicionales basadas en diccionarios y por otro, técnicas basadas en aprendizaje máquina. Para la extracción de sentimientos, se utilizaron tres técnicas diferentes: i) Standford CoreNLP, un conjunto de herramientas de procesado de lenguaje natural y modelos computacionales de lenguaje, que incluyen el análisis de sentimientos mediante una red neuronal pre-entrenada que devuelve valores entre -1 (sentimiento negativo) y +4 (sentimiento positivo). ii) SentimentR, una librería de análisis de sentimientos para R, basada en léxico y diccionario que devuelve valores entre -1 (negativo) y +1 (positivo). iii) el diccionario LIWC, que extrae el tono emocional del texto en valores entre 0 % (negativo) y 100 % (positivo). Para la extracción de tono analítico, la autenticidad y la influencia, se utilizó la última versión disponible del diccionario Linguistic Inquiry and Word Count (LIWC). Por un lado, el tono analítico indica cuan lógico y jerarquizado es el estilo de escritura empleado. Por otro lado, la influencia, denota la sensibilidad personal, confianza y tono tentativo. Por último, la autenticidad hace referencia a la honestidad o superficialidad con la que está redactado el texto. Todas estas características vienen representadas en porcentaje de términos que aparecen sobre el total de palabras del texto. Para la caracterización de aspectos morales se utilizó el diccionario de la Moral Fundations[5], que devuelve los siguientes aspectos, cuidado/daño, equidad/engaño, lealtad/traición, autoridad/subversión y santidad/degradación.

Una vez caracterizados los textos, se emplearon modelos lineales de efectos mixtos para analizar la interacción de entre variables y comparar las diferencias entre los grupos poblacionales existentes. Así pues, se fijaron como efectos fijos las diferentes características lingüísticas extraídas, la recomendación del revisor, el tipo de revisión por pares de la revista y el género del revisor. Como efectos aleatorios, se fijaron el número de palabras del texto de revisión y la revista. Los resultados de este análisis demuestran que, por un lado, los textos de revisión acompañados de recomendaciones de rechazar o revisiones mayores utilizan un lenguaje menos emocional y más analítico. Por el otro, los resultados del modelo indican que no existen sesgos o diferencias significativas en el tipo de revisión, el área de la revista o el género del revisor.

Siguiendo la misma línea de caracterización lingüística, en la contribución al congreso *PUBMET*, titulada *Linguistic and semantic characteristics or articles and peer review reports in social and medical sciences: analysis of articles published in Open Research Central* y realizado sobre un conjunto de datos de open peer review, obtenido de Open Research Central[6]. Se utiliza el diccionario de LIWC para caracterizar los textos de revisión, además, también se utilizan representaciones vectoriales (*word embeddings*) como parte de la caracterización. En este caso, se estudia también la estructura de los artículos en dos áreas concretas, ciencias sociales y ciencias médicas, separando las diferentes secciones de la estructura de los artículos. En este aspecto, sus estructuras difieren según el área, la introducción y la sección de conclusiones tienen a ser más largas y elaboradas en ciencias sociales, y en general, presentan mayores valores de sensibilidad personal y confianza y un tono menos positivo que los artículos de ciencias médicas. No obstante, en este estudio y para modelo de revisión por pares, no se encontraron diferencias significativas en los textos de revisión.

Por otro lado, la contribución presentada al *International Congress on Peer Review and Scientific Publication*, titulada *Unprofessional Comments in Peer Review Reports Across Scholarly Disciplines*, trata de caracterizar comentarios no profesionales en los textos de revisión. En este caso, se utilizan técnicas cualitativas, donde, mediante un etiquetado manual, se identifican aquellos comentarios poco profesionales o que emplean lenguaje poco apropiado o malsonante. Sobre el conjunto de datos de PEERE, que

[5] https://moralfoundations.org/
[6] https://openresearchcentral.org/

contiene información de aproximadamente 300.000 artículos, se extrae una muestra aleatoria de 380 artículos, que se compone a su vez de 1.147 textos de revisión. Tras analizar manualmente la muestra, se detectó que el 1.1 % de los textos de revisión seleccionados contienen comentarios poco profesionales.

Por último, para la *International Conference of the Catalan Association for Artificial Intelligence*, se preparó una charla sobre los presentes problemas y nuevos retos en el ámbito del NLP aplicado al Science of Science, donde se destaca la importancia de generar modelos específicos, capaces de interpretar y trabajar con lenguaje científico y técnico.

3.6.2. Contribución B

Los trabajos agrupados en el bloque B estudian las características de las revisiones, más allá de la sintaxis de los textos de revisión. Teniendo en cuenta también el contenido de los mismos y midiendo el efecto que tienen las revisiones sobre los propios artículos. Este grupo de publicaciones esta formado por las siguientes contribuciones a congresos y publicaciones:

- Daniel Garcia-Costa, Anabel Forte, Emilia López-Iñesta, Flaminio Squazzoni, Francisco Grimaldo. Does peer review improve the statistical content of manuscripts? A study on 27,467 submissions to four journals. R. Society Open Science, 9:210681.210681, 2022. (Apéndice C)

- Federico Bianchi, Daniel García-Costa, Francisco Grimaldo, Flaminio Squazzoni. Measuring the effect of reviewers on manuscript change: A study on a sample of submissions to Royal Society journals (2006–2017). Journal of Informetrics, Volume 16, Issue 3, 2022. (Apéndice B)

- Anabel Forte, Daniel Garcia-Costa, Emilia López-Iñesta, Phil Hurst, Flaminio Squazzoni, Francisco Grimaldo. Change in statistical terms in peer-reviewed journals. PEERE International Conference on Peer Review, Valencia (Spain), 2020.

- Federico Bianchi, Daniel Garcia-Costa, Francisco Grimaldo, Flaminio Squazzoni. Peer Review Improves Manuscripts of Moderate Initial Quality. A Study on Ten Journals from the Royal Society (2006-2017). PEERE International Conference on Peer Review, Valencia (Spain), 2020.

- Ivan Buljan, Daniel Garcia-Costa, Francisco Grimaldo, Richard Klein, Marjan Bakker, Ana Marušić. Development of a List to Detect Statistical and Methodological Terms in Peer Reviews. International Congress on Peer Review and Scientific Publication, Chicago, (United States), 2022.

La publicación *Does peer review improve the statistical content of manuscripts? A study on 27,467 submissions to four journals*, y sus resultados previos presentados al PEERE International Conference on Peer Review se centran en estudiar la evolución en la cantidad de contenido estadístico en los artículos científicos al pasar por el proceso de revisión por pares como estimador de rigor metodológico.

Se analizaron 27,267 artículos enviados a 5 revistas de la Royal Society, comprendidos entre 2006 y 2017. Estos datos provienen de la COST Action PEERE y contienen información, tanto de los propios artículos y sus autores o autoras, como de las revisiones y personas que los revisaron. Partiendo de un glosario de términos estadísticos, se creó un diccionario para identificar y cuantificar la presencia o ausencia de estos. Este diccionario, en formato LIWC, se utilizó para contar el número de términos estadísticos diferentes, presentes en las distintas versiones de los artículos, así como en los textos de revisión. De modo que, se pudiera estudiar la evolución en el número de términos a lo largo de la vida del artículo. Los textos se procesaron excluyendo fórmulas, figuras y tablas, pero dejando los pies de las mismas para no perder el contenido descriptivo.

Con el objetivo de explorar el efecto de la revisión por pares sobre el contenido estadístico se utilizaron dos modelos: por un lado, una regresión de Poisson sobre el número de términos estadísticos diferentes presentes en la última versión del artículo y por otro lado, una regresión logística para la probabilidad de que el artículo sea aceptado después del proceso de revisión. Se aplicó una selección Bayesiana de variables para identificar aquellas que debían ser incluidas en las regresiones, considerando las probabilidades a posteriori para cada posible combinación de variables y calculando, para cada una de ellas, su Probabilidad a Posteriori de Inclusión (PIP). Finalmente, se seleccionaron las variables con una PIP superior a 0.5.

Los resultados de los modelos indican que la existencia de guías sobre revisiones estadísticas por parte de la revista no tiene un efecto significativo sobre la variación en el número de términos estadísticos. No obstante si existe una diferencia significativa entre las distintas revistas estudiadas, que pertenecen a su vez a diferentes áreas de la ciencia y tienen un público objetivo distinto. Los resultados sugieren que, por lo general, el proceso de revisión por pares, sobre artículos finalmente aceptados, incrementa el número de términos estadísticos, incrementando por tanto, el rigor metodológico de los mismos. En cambio, esto no sucede en aquellos artículos que se rechazan por el camino, donde un 93.1 % de los mismos no varían. Del mismo modo, el número de revisores y el grado de acuerdo entre estos, juegan un papel fundamental en la probabilidad de aceptación de los artículos.

De manera similar, la contribución presentada al congreso *International Congress on Peer Review and Scientific Publication*, titulada *Development of a List to Detect Statistical and Methodological Terms in Peer Reviews*, trata sobre el desarrollo de un diccionario de términos estadísticos aplicando técnicas de creación de diccionarios semiautomáticas. Partiendo de un conjunto de términos confeccionado a partir de diversos glosarios de estadística, se generó un modelo de *word embeddings* sobre un conjunto de textos de revisión y se extendió la lista de términos de manera iterativa. En cada iteración se buscaron términos similares en el espacio vectorial del modelo y se comprobaba manualmente la nueva selección de términos y verificando posibles solapamientos. Al final de este proceso se generó un diccionario en formato LIWC compuesto por 16 categorías y dos grupos de variables.

Por otro lado, el artículo *Measuring the effect of reviewers on manuscript change: A study on a sample of submissions to Royal Society journals (2006–2017)*, así como su trabajo previo presentado al *PEERE International Conference on Peer Review*, presentan un estudio sobre los cambios de redacción que sufren los artículos durante la fase de revisión y el efecto que estos tienen sobre la probabilidad de ser citados.

El estudio se realizó sobre 10.996 artículos enviados a 7 revistas de la Royal Society en el periodo comprendido entre 2006 y 2017. De este conjunto de datos, se extrajo información textual de los artículos, originalmente en diferentes formatos como PDF, DOCX o TEX y se transformó y estandarizó a texto plano, eliminando tablas, figuras, cabeceras y pies de página y otras marcas o caracteres causadas por la conversión de formatos. Además, se descargaron las versiones publicadas de los artículos que habían sido aceptados y se estandarizaron del mismo modo que las demás, descargando también algunas métricas, como el número de citas.

Las diferentes versiones de los artículos se agrupan según el artículo al que pertenecen para poder compararlas entre sí, extraer las diferencias y analizar los cambios sufridos. Para medir esos cambios, se computa la distancia Levenshtein (Levenshtein, 1966). Esta distancia, normalizada, arroja la proporción de cambios entre ambos documentos como un valor entre 0 y 1. Hacer uso de una distancia no basada en tokens permite medir absolutamente todos los cambios, incluso las reordenaciones de texto, cosa que no sería posible empleando una distancia basada en tokens (Augsten and Böhlen, 2014). Para cuantificar el cambio en la sección de referencias, se identifican mediante el uso de un conjunto de expresiones regulares, primero, para identificarlas y separar sus campos (lista de autores, título, etc.) y después se calcula la proporción de referencias diferentes como

$$1 - \frac{\textit{Número de referencias iguales}}{\textit{Número máximo de referencias en ambos documentos}}$$

Por otro lado, para estudiar el posible efecto de las recomendaciones de los revisores, se calcula un estimador de acuerdo entre revisores (Bravo et al., 2018).

Para el análisis, se emplean dos modelos estadísticos, por un lado, un modelo lineal de efectos mixtos con el que estimar el efecto del número de revisores, acuerdo entre revisores y longitud de la revisión sobre la proporción de cambios. Por otro lado, una regresión logística para estimar el efecto sobre la probabilidad de ser citado como mínimo una vez.

Los resultados sugieren que los revisores tienen un impacto considerable en el número de cambios producidos en los artículos, llegando estos a cambiar hasta en un 40 % en el texto y la sección de referencias. Estos cambios tienen una tendencia creciente en función al número de revisores. Además, estos cambios se producen independientemente de la calidad inicial del artículo y se pueden observar, tanto en aquellos inicialmente mejor valorados por los revisores, como en los que peores valoraciones iniciales reciben. A su vez, aquellos artículos que más cambios sufren durante el proceso de revisión reflejan una mayor probabilidad de ser citados una vez publicados.

3.6.3. Contribución C

La contribución C de esta tesis doctoral presenta una métrica para medir la completitud de los textos de revisión y poder evaluar el valor de desarrollo, es decir, el valor que el sistema de revisión por pares aporta a la mejora de los artículos. Esta contribución se refleja en la siguiente publicación.

- Daniel Garcia-Costa, Flaminio Squazzoni, Bahar Mehmani, Francisco Grimaldo. Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. PeerJ, 10:e13539, 2022. (Apéndice D)

Este trabajo busca generar una métrica con la que medir la completitud, como estimador de calidad, de los textos de revisión por pares, que permita a los diferentes actores de la comunidad científica evaluar las revisiones. El conjunto de datos empleado es, con mucha probabilidad, el mayor conjunto de textos de revisión utilizado hasta la fecha. Este conjunto de datos, cedido por la casa editorial Elsevier, contiene datos de más de 1.3 millones de textos de revisión, pertenecientes a 740 revistas que cubren las 4 grandes áreas de la ciencia, ciencias de la vida (LS), ciencias de la salud y medicina (HMS), física y ciencias puras (PS) y por último, ciencias sociales y económicas (SSE).

Como en todo trabajo con datos, primero es necesario estandarizar todas las variables disponibles para homogeneizar su contenido. Esta parte del proceso conlleva un gran trabajo de minería de datos, teniendo que limpiar y controlar múltiples variables para asegurar su integridad. De igual modo, es necesario enriquecer los datos con fuentes de datos externas. En este caso, a través del ICSR Lab se accedió a la base de datos de Scopus para recuperar datos sobre los revisores, tales como, la fecha de la primera publicación o el HIndex y otros datos sobre las revistas como el cuartil que ocupan en JCR. Otro aspecto importante sobre los revisores es el género de los mismos, si no se dispone de él, este puede inferirse utilizando el nombre y el país de procedencia para tener una estimación del género con un margen de error aceptable.

Para medir la calidad o completitud de los textos de revisión se partió de ARCADIA (Superchi et al., 2020), un estudio cualitativo que define una lista de aspectos a tener en cuenta para evaluar si una revisión es o no de calidad. De entre estos aspectos, después de analizarlos sobre el conjunto de datos, se decide seleccionar aquellos que son fácilmente medibles mediante técnicas cuantitativas: impacto, literatura, metodología, métodos estadísticos, conclusiones, limitaciones, aplicabilidad, presentación, disponibilidad de datos y por último, organización y escritura.

A falta de modelos de aprendizaje máquina diseñados para un lenguaje tan específico como los textos de revisión, se optó por una técnica semiautomática de construcción de diccionarios. Para ello se construyó un modelo lingüístico computacional utilizando como entrada el conjunto de textos de revisión y, mediante una lista inicial de palabras manualmente seleccionadas, se buscó en el espacio vectorial del modelo otros términos cercanos o similares que pudieran ser sinónimos de estos. Una vez obtenida la nueva lista de términos se revisaron y validaron por 3 personas y se seleccionaron los términos que tenía sentido mantener. Estos nuevos términos se añadieron a la lista y se volvió a repetir el proceso hasta obtener precisiones en las predicciones menores que un umbral establecido. El diccionario generado contiene un total de 1565 términos distribuidos en las 10 dimensiones a medir.

Este diccionario se aplicó sobre todo el conjunto de textos de revisión, extrayendo para cada uno de ellos, un valor que mide la proporción de términos que contiene el texto. Partiendo de los valores obtenidos para cada una de las 10 dimensiones, se calcula una métrica de completitud que unifica dichos valores en uno único. Sobre él se aplican diferentes modelos lineales generalizados de tipo Gamma, para caracterizar y comparar los diferentes grupos poblacionales de revisores en base al resto de características del conjunto de datos.

Los resultados de estos modelos sugieren diferencias entre revistas en cuanto a su factor de impacto, donde, por lo general, aquellas con mayor factor de impacto presentan a su vez una mayor puntuación. Con algunas excepciones, como por ejemplo para el área SSE, donde no se sigue esta premisa. A su vez, se aprecia una correlación positiva con el tiempo de revisión, es decir, aquellas revisiones que tardan más tiempo en ser entregadas son, a su vez, aquellas con mayor puntuación.

A nivel humano, las diferencias en base al género de los revisores y revisoras no es muy significativa a nivel global, aunque si lo es en las áreas de SSE y HMS, donde las mujeres obtienen puntuaciones aproximadamente un 8 % mayores que los hombres. En cuanto a la edad de los revisores, se aprecia una diferencia clara, transversal para todas las áreas, entre los revisores junior y los senior, donde los jóvenes obtienen puntuaciones de un 8 % mayores, llegando a casi el 10 % en SSE. Por último, a nivel geográfico, se aprecian claras diferencias entre las regiones de la institución de los revisores, siendo Europa oriental la región con mayor puntuación llegando a marcar diferencias de hasta el 15 % con regiones de Asia.

3.6.4. Contribución D

Esta última contribución no estaba contemplada inicialmente en la planificación de esta tesis doctoral. La necesidad de este estudio surge a raíz de la aparición de la COVID-19, para estudiar sus efectos sobre el sistema de publicaciones científicas y se compone de la siguiente publicación.

- Flaminio Squazzoni, Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa, Mike Farjam, Bahar Mehmani. Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals. PLoS One, 16(10): e0257919, 2021. (Apéndice E)

La pandemia de la COVID-19 provocó un cambio generalizado a nivel mundial en muchos aspectos de nuestras vidas, desde el ámbito personal hasta el ámbito laboral. De hecho, la ciencia también se vio afectada, durante los 10 primeros meses de la aparición de la COVID-19 se detectó un gran incremento en el número de publicaciones científicas. Dicho incremento fue notorio en todas las áreas de la ciencia y no solo en aquellos estudios relacionados con la propia COVID-19.

Para este estudio se firmó un contrato de compartición de datos con la casa editorial Elsevier, que aglutina información de 2329 revistas y más de 8 millones de artículos enviados en el periodo comprendido entre enero de 2018 y mayo de 2020. A su vez, recoge información de más de 5 millones de personas. Sobre este conjunto de datos, al igual que para otros estudios presentados, la información de las personas involucradas se enriquece realizando consultas, a través del ICSR Lab, a la base de datos de Scopus. Para estimar el género de los actores involucrados, se recurre la inferencia mediante el nombre y el país de procedencia, a través de varias librerías y servicios dedicados a este fin.

Los resultados de este estudio muestran un incremento en el envío de artículos, entre febrero y mayo de 2020, de un 30 % en comparación con el mismo periodo en años anteriores. Realizando el mismo análisis por áreas, el incremento es mucho más acusado en HMS, que llega al 63 % de incremento. Del mismo modo, el número de invitaciones de revisión aceptadas muestra un incremento general del 29 %, nuevamente llegando al 63 % en HMS.

En cuanto a factores humanos, los modelos arrojan resultados de un claro sesgo de género a favor de los hombres. Especialmente al comparar hombres jóvenes con mujeres jóvenes, donde la diferencia llega a ser de aproximadamente un 15 % en HMS. Del mismo modo existe una interacción positiva entre el género y la edad, las mujeres senior se ven menos afectadas por este sesgo que las mujeres jóvenes. En cambio, este sesgo no está presente en la proporción de revisiones aceptadas, es decir, las mujeres, pese no haber incrementado el número de envíos de artículos, si aumentaron el número de revisiones aceptadas en una proporción muy similar a sus homólogos hombres.

En el siguiente capítulo se resumen, discuten y destacan los resultados y hallazgos más relevantes de las contribuciones que componen esta tesis doctoral así como las posibles líneas de trabajo futuro derivadas de la misma.

Capítulo 4

Conclusiones y trabajo futuro

Dado que esta tesis doctoral se presenta con mención internacional al título de doctor y atendiendo a la normativa establecida por la Escuela de Doctorado de la Universitat de València, el presente capítulo, en el que se detallan las conclusiones y trabajo futuro ha sido redactado en Inglés.

4.1. Discussion and conclusions

The peer review process of scientific articles plays a paramount role in the quality assurance of science (Kharasch et al., 2021). In this sense, the credibility of scientific journals and publishers depends mainly on the quality of their review system (Edwards and Siddhartha, 2017) (Kharasch et al., 2021), which is why they take particular interest in publishing only studies with remarkable scientific rigor (Atjonen, 2019). These reviews help improve manuscripts' content and ensure quality standards that benefit the scientific community. The interest in evaluating and understanding how they work is continuously evident. The proliferation of Science of Science has led to many works in this regard in recent years.

This doctoral thesis presents different contributions in the field of the characterization of the peer review system of scientific articles, focusing not only on characterization but also on the calculation of metrics and on the different methodological approaches for the treatment of this information, as well as on analyzing the existing differences in the various socio-demographic groups involved.

On the one hand, studying the linguistic characteristics of the review texts, it is observed that the type of language used is clearly linked to the reviewer's recommendation (Buljan et al., 2020). In the case of more negative recommendations, such as major changes or rejection of the article, reviewers use more analytical and less emotional language while producing shorter reviews, thus being more concrete and specific in their writing (Buljan et al., 2020). This does not happen in the case of more positive reviews, such as minor changes or acceptance of the article, where the language used denotes a more emotional tone (Buljan et al., 2020).

On the other hand, the peer review system has a considerable effect on the changes in articles submitted for publication. Exposing an article to several rounds of review means that it accumulates, on average, 40 % of changes concerning its initial version between the text of the article and its references (Bianchi et al., 2022). This percentage of changes tends to increase with the number of reviewers, with a greater number of changes in those articles to which more different reviewers are assigned. In this aspect, changes occur regardless of the initial quality of the article, even in those that receive more positive evaluations from the reviewers in the first round. Furthermore, although the impact of an article will depend on many other factors (Coupé, 2013) (Seeber, 2020), there is a slight correlation between the number of changes that articles undergo during their review phase and the probability that they will be cited at least once after publication (Bianchi et al., 2022).

Likewise, some parts of the articles' content can serve as an estimator of methodological rigor. One of these is the statistical content of the articles. Correctly reporting the statistical analyses of a study guarantees its reproducibility, provided that the data used are provided. Reviewers should, therefore, not only focus on what is innovative but also on the methodology used for data and statistical analysis (Köhler et al., 2020). In this context, the results indicate that articles with statistical content, regardless of the amount of statistical content, increase the number of statistical terms used during the review process. The more statistical content they present, the higher the probability of acceptance of the article. Conversely, articles directly rejected by the editor generally have less statistical content than those that pass to the review phase (Garcia-Costa et al., 2022a). The existence of statistical review guidelines by the journals seems to guarantee a minimum of statistical content in the articles reviewed but does not seem to have any implication in the changes they undergo.

Analyzing the content of review texts allows, in addition to extracting the characteristics of the text, to understand what points the reviewer is focusing on to evaluate an article. A correct and complete review should address aspects such as the impact of the publication, the methodology used, the writing and presentation of the document, the results obtained, etc. (Superchi et al., 2020). In other words, not only articles should be methodologically rigorous, but also reviews. The quantification of these aspects by means of natural language processing techniques makes it possible to generate metrics with which to evaluate the standards of the peer review process. Specifically, the solution proposed in (Garcia-Costa et al., 2022b) brings together ten dimensions or aspects to be considered to evaluate the completeness of a review text. These aspects are measured using a dictionary generated by a semi-automatic dictionary generation technique, using a word embedding model created from the review texts. Thanks to this dictionary, it is possible to quantify each of these ten dimensions and calculate a completeness metric on the reviews themselves. Comparisons using this metric show that peer review standards are robust across the entire data set studied, although with certain nuances. Reviews belonging to the areas of social sciences and economics present higher scores, which coincides with historical trends in peer review in journals in this area (Merriman, 2020), where they tend to include very characteristic and specific elements that are not usually seen in other areas. Time also plays an essential role in the quality of the reviews; spending ten more days on a review results in an average increase of 3 % in the review score, indicating that spending more time on a review results in a more complete and more constructive text.

These differences are also reflected at the geographic level, where very heterogeneous results are observed among the different regions. Part of these differences could reflect certain linguistic and cultural particularities specific to some regions and, therefore, different ways of working and doing science and research. In this context, Asian countries stand out, which are quite distant from countries in Western Europe and the United States. These results suggest that it is necessary to strengthen training initiatives and diversity policies to reinforce the standards of the peer review system (Garcia-Costa et al., 2022b). In addition to the geographic level, it is observed that young researchers tend to perform more complete and constructive reviews, which is something that does not happen with their older counterparts. Likewise, focusing on the gender of the reviewers, women in the disciplines related to social and economic sciences and health sciences also perform better reviews than men in those same disciplines.

Focusing on the period of the COVID-19 pandemic, the difference between male and female researchers was clearly accentuated. During this period, there was a general increase in the number of articles submitted to scientific journals, resulting in a 30 % increase in the number of publications compared to the same period in previous years. This striking increase did not have the same connotations for men and women; on the contrary, the number of articles submitted by women was clearly lower than that submitted by men (Squazzoni et al., 2021). Specifically, the group that most distanced itself from this increase was younger women, with a difference of up to 15 % in the area of health sciences. In addition, an increase in the number of accepted reviews was also detected, obviously due to the rise in the number of submissions. In this case, in contrast to submissions, women generally accepted more articles for review than men, with the health sciences area showing the greatest difference (Squazzoni et al., 2021). These results denote a double bias since, despite not having increased the number of submissions, women did devote more effort to tasks with little or no recognition.

Throughout the development of this work, an unprecedented amount of data on the peer review process has been processed and provided by different scientific publishers. All this data and text mining work has made it possible to establish a workflow to transform this data into information and knowledge. This is reflected in the creation, by Elsevier, of the Peer Review Workbench, which is, to a large extent, the result of the knowledge acquired during the course of this doctoral thesis.

As a general conclusion, the characterization of the peer review system of scientific articles serves not only to understand how it works but also to understand how the different groups involved behave, what effect it has on the articles submitted for publication, and, consequently, what policies, actions or initiatives need to be taken to ensure its proper functioning or to improve it, if necessary. In this context, the results obtained are of great informative value for those bodies or institutions that supervise and design such policies and initiatives. Quantitatively measuring the functioning of the review system is extremely useful for the entire scientific community, where every one of its actors can benefit, in their own way, from this information.

This doctoral thesis has generated different results in various dimensions of the peer review system, also establishing procedures and methodologies when working with this information, opening and putting the focus on a complex problem, which requires further study to help understand its operation and, above all, its contribution to the scientific community.

4.2. Future work

Given the contributions presented in this doctoral thesis, there are different avenues to continue the work presented here and may constitute lines of future work.

On the one hand, in line with the characterization of review texts for the generation of metrics, one of the problems encountered when generating metrics from this information has been the lack of labeled datasets of an acceptable size. In this context, large-scale experimentation could be conducted to label a dataset. This would allow the use of machine learning techniques to generate models capable of classifying review texts according to their general features or aspects. Another possibility could be to use the dictionary already developed to pre-label the data and use these as input to train a classification model based on some current LLM that also considers the review texts'contextual information.

On the social aspects of science, another interesting starting point may be to study the burden that people place on the system with respect to their contribution to it. That is, how much people review what they themselves submit. If there are differences, it would be interesting to study them concerning gender, country or geographic region of origin, or age. The study of these characteristics would make it possible to find out whether the review system is equal for everyone or whether, on the contrary, there are groups that contribute more of the work they themselves generate or vice versa.

Regarding the type of peer review, doubts arise as to whether the behavior and characteristics found would also be the same in the open peer review model. In this context, it would be ideal to reproduce some of the studies presented in this doctoral thesis but using a data set of this type of review process with which to compare the results, looking for differences in behavior according to the type of review process.

Finally, concerning the pandemic produced by COVID-19, we could try to find out how the different countries have responded to the supposed increase in the number of publications produced, studying how it has affected each geographical region not only in terms of the number of submissions and reviews but also in review times, acceptance rates and reviewers'recommendations and how these factors have affected the publication system.

Apéndice A

Large-scale language analysis of peer review reports

eLife

FEATURE ARTICLE

META-RESEARCH

Large-scale language analysis of peer review reports

Abstract Peer review is often criticized for being flawed, subjective and biased, but research into peer review has been hindered by a lack of access to peer review reports. Here we report the results of a study in which text-analysis software was used to determine the linguistic characteristics of 472,449 peer review reports. A range of characteristics (including analytical tone, authenticity, clout, three measures of sentiment, and morality) were studied as a function of reviewer recommendation, area of research, type of peer review and reviewer gender. We found that reviewer recommendation had the biggest impact on the linguistic characteristics of reports, and that area of research, type of peer review and reviewer gender had little or no impact. The lack of influence of research area, type of review or reviewer gender on the linguistic characteristics is a sign of the robustness of peer review.

IVAN BULJAN*, DANIEL GARCIA-COSTA, FRANCISCO GRIMALDO, FLAMINIO SQUAZZONI AND ANA MARUŠIĆ

Introduction

Most journals rely on peer review to ensure that the papers they publish are of a certain quality, but there are concerns that peer review suffers from a number of shortcomings (***Grimaldo et al., 2018***; ***Fyfe et al., 2020***). These include gender bias, and other less obvious forms of bias, such as more favourable reviews for articles with positive findings, articles by authors from prestigious institutions, or articles by authors from the same country as the reviewer (***Haffar et al., 2019***; ***Lee et al., 2013***; ***Resnik and Elmore, 2016***).

Analysing the linguistic characteristics of written texts, speeches, and audio-visual materials is well established in the humanities and psychology (***Pennebaker, 2017***). A recent example of this is the use of machine learning by Garg et al. to track gender and ethnic stereotypes in the United States over the past 100 years (***Garg et al., 2018***). Similar techniques have been used to analyse scientific articles, with an early study showing that scientific writing is a complex process that is sensitive to formal and informal standards, context-specific canons and subjective factors (***Hartley et al., 2003***). Later studies found that fraudulent scientific papers seem to be less readable than non-fraudulent papers (***Markowitz and Hancock, 2016***), and that papers in economics written by women are better written than equivalent papers by men (and that this gap increases during the peer review process; ***Hengel, 2018***). There is clearly scope for these techniques to be used to study other aspects of the research and publishing process.

To date most research on the linguistic characteristics of peer review has focused on comparisons between different types of peer review, and it has been shown that open peer review (in which peer review reports and/or the names of reviewers are made public) leads to longer reports and a more positive emotional tone compared to confidential peer review (***Bravo et al., 2019***; ***Bornmann et al., 2012***). Similar techniques have been used to explore possible gender bias in the peer review of grant applications, but a consensus has not been reached yet (***Marsh et al., 2011***; ***Magua et al., 2017***). To date, however, these techniques have not been applied to the peer review process at a large scale, largely because most journals strictly limit access to peer review reports.

***For correspondence:** ibuljan@mefst.hr

Competing interests: The authors declare that no competing interests exist.

Funding: See page 8

Reviewing editor: Peter Rodgers, eLife, United Kingdom

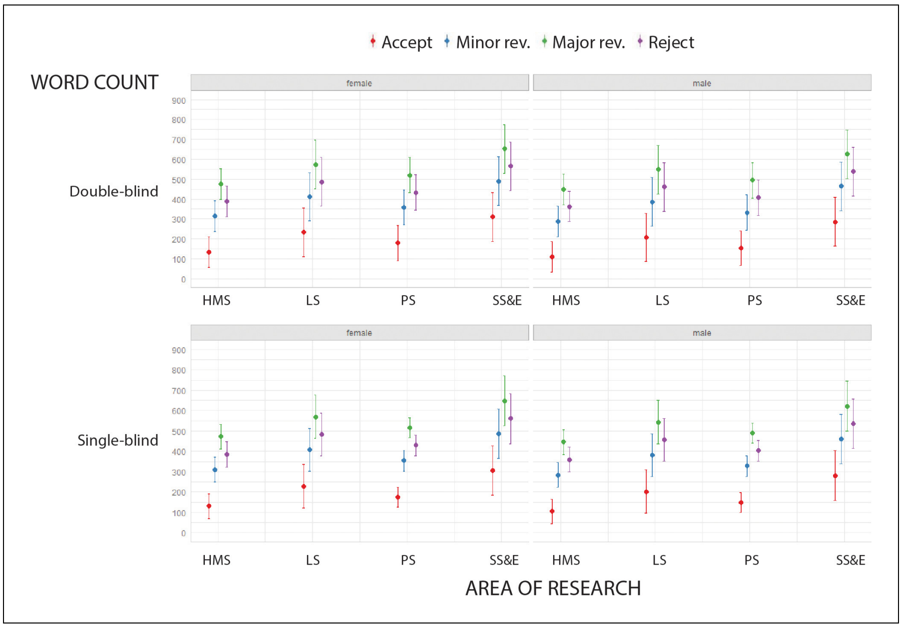

Figure 1. **Words counts in peer review reports.** Word count (mean and 95% confidence interval; LIWC analysis) of peer review reports in four broad areas of research for double-blind review (top) and single-blind review (bottom), and for female reviewers (left) and male reviewers (right). Reports recommending accept (red) were consistently the shortest, and reports recommending major revisions (green) were consistently the longest. See ***Supplementary file 1*** for summary data and mixed model linear regression coefficients and residuals. HMS: health and medical sciences; LS: life sciences; PS: physical sciences; SS&E: social sciences and economics.

Here we report the results of a linguistic analysis of 472,449 peer review reports from the PEERE database (***Squazzoni et al., 2017***). The reports came from 61 journals published by Elsevier in four broad areas of research: health and medical sciences (22 journals); life sciences (5); physical sciences (30); social sciences and economics (4). For each review we had data on the following: i) the recommendation made by the reviewer (accept [n = 26,387, 5.6%]; minor revisions required [134,858, 28.5%]; major revisions required [161,696, 34.2%]; reject [n = 149,508, 31.7%]); ii) the broad area of research; iii) the type of peer review used by the journal (single-blind [n = 411,727, 87.1%] or double-blind [n = 60,722, 12.9%]); and the gender of the reviewer (75.9% were male; 24.1% were female).

Results

We used various linguistic tools to examine the peer review reports in our sample (see Methods for more details). Linguistic Inquiry and Word Count (LIWC) text-analysis software was used to perform word counts and to return scores of between 0% and 100% for 'analytical tone', 'clout' and 'authenticity' (***Pennebaker et al., 2015***). Three different approaches were used to perform sentiment analysis: i) LIWC returns a score between 0% and 100% for 'emotional tone' (with more positive emotions leading to higher scores); ii) the SentimentR package returns a majority of scores between –1 (negative sentiment) and +1 (positive sentiment), with an extremely low number of results outside that range (0.03% in our sample); iii) the Stanford CoreNLP returns a score between 0 (negative sentiment) to +4 (positive sentiment). We also used LIWC to analyse the reports in terms of five foundations of morality (***Graham et al., 2009***).

Length of report

For all combinations of area of research, type of peer review and reviewer gender, reports recommending accept were shortest, followed by reports recommending minor revisions, reject, and major revisions (***Figure 1***). Reports written by reviewers for social sciences and economics

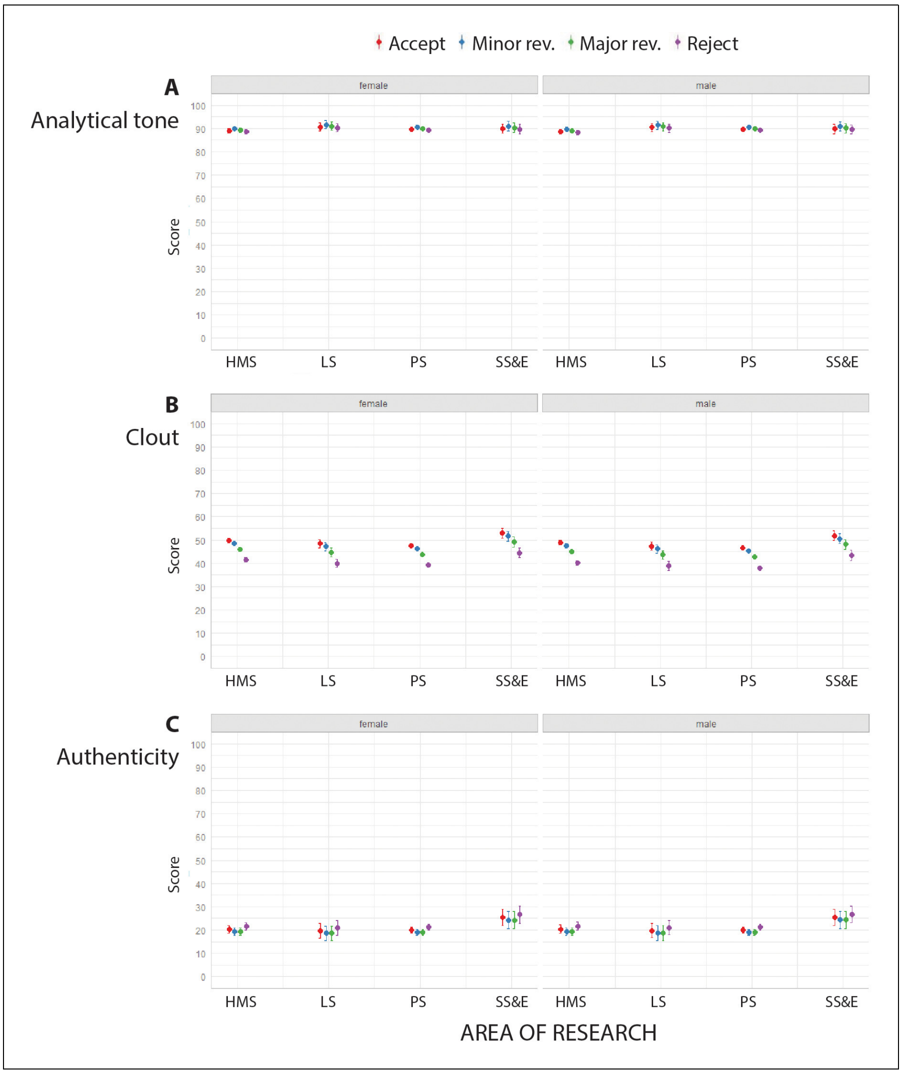

Figure 2. **Analytical tone, clout and authenticity and in peer review reports for single-blind review.** Scores returned by LIWC (mean percentages and 95% confidence interval) for analytical tone (**A**), clout (**B**) and authenticity (**C**) for peer review reports in four broad areas of research for female reviewers (left) and male reviewers (right) using single-blind review. Reports recommending accept (red) consistently had the most clout, and reports recommending reject (purple) consistently had the least clout. See ***Supplementary files 2–4*** for summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for analytical tone, clout and authenticity. HMS: health and medical sciences; LS: life sciences; PS: physical sciences; SS&E: social sciences and economics.
The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Analytical tone, clout and authenticity in peer review reports for double-blind review.

journals were significantly longer than those written by reviewers for medical journals; men also tended to write longer reports than women; however, the type of peer review (i.e., single- vs. double-blind) did not have any influence on the length of reports (see Table 2 in ***Supplementary file 1***).

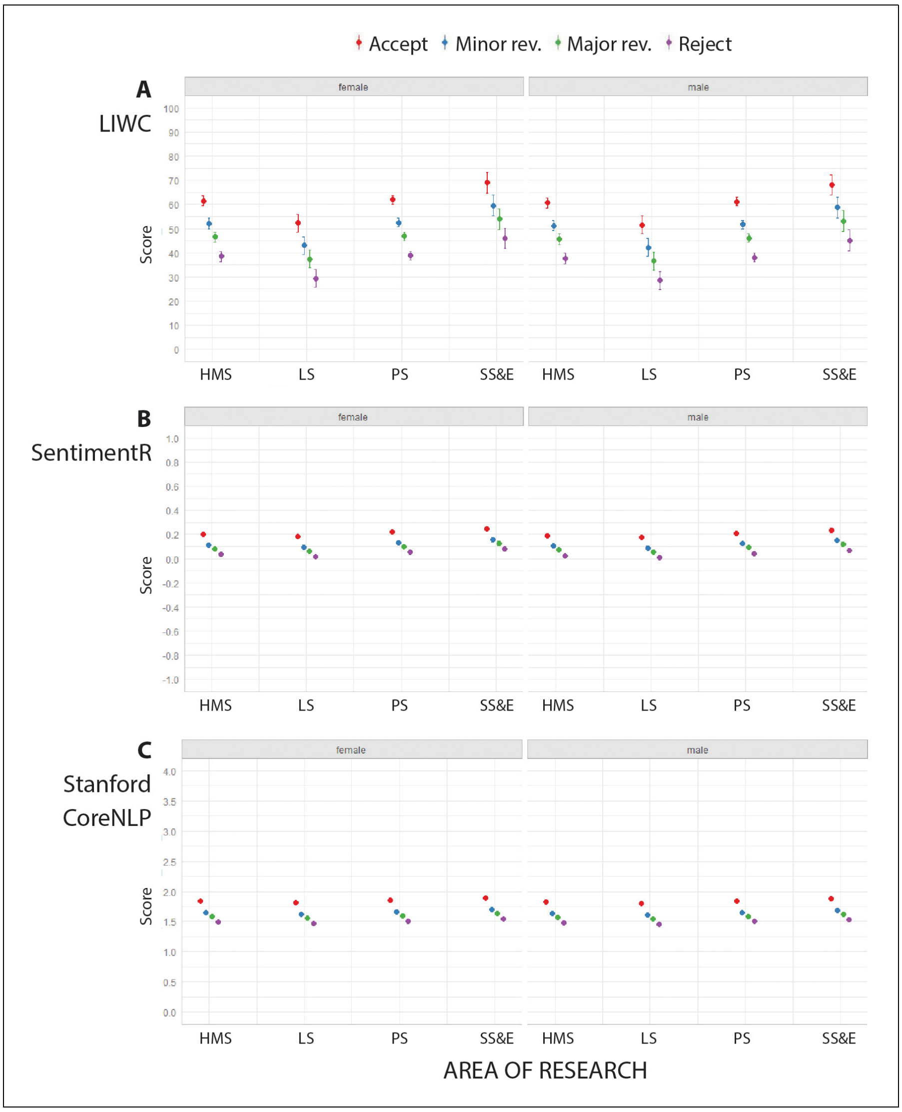

Figure 3. **Sentiment analysis of peer review reports for single-blind review.** Scores for sentiment analysis returned by LIWC (**A**; mean percentage and 95% confidence interval, CI), SentimentR (**B**; mean score and 95% CI), and Stanford CoreNLP (**C**; mean score and 95% CI) for peer review reports in four broad areas of research for female reviewers (left) and male reviewers (right) using single-blind review. See ***Supplementary files 5–7*** for summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for sentiment according to LIWC, SentimentR and Stanford CoreNLP analysis.
The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Sentiment analysis of peer review reports for double-blind review.

Analytical tone, clout and authenticity

LIWC returned high scores (typically between 85.0 and 91.0) for analytical tone, and low scores (typically between 18.0 and 25.0) for authenticity, for the peer review reports in our sample (***Figure 2A,C***; ***Figure 2—figure supplement 1A,C***). High authenticity of a text is defined as the use of more personal words (I-

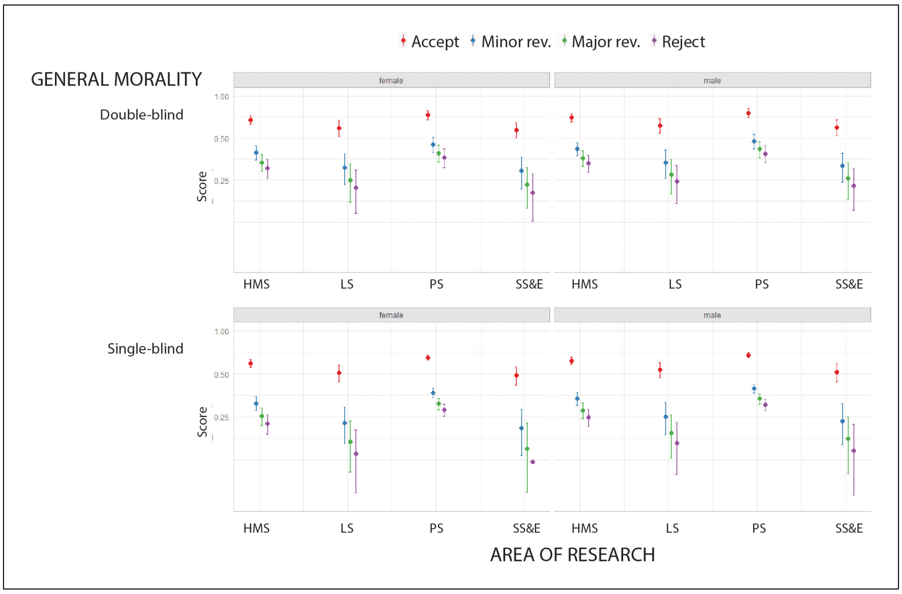

Figure 4. Moral foundations in peer review reports. Scores returned by LIWC (mean percentage on a log scale) for general morality in peer review reports in four broad areas of research for double-blind review (top) and single-blind review (bottom), and for female reviewers (left) and male reviewers (right). Reports recommending accept (red) consistently had the highest scores. See ***Supplementary file 8*** for lists of the ten most frequent words found in peer review reports for general morality and the five moral foundation variables. HMS: health and medical sciences; LS: life sciences; PS: physical sciences; SS&E: social sciences and economics.
The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Scores returned by LIWC (mean percentage on a log scale and 95% CI) for care/harm, one of the five foundations of Moral Foundations Theory.
Figure supplement 2. Scores returned by LIWC (mean percentage on a log scale and 95% CI) for fairness/cheating, one of the five foundations of Moral Foundations Theory.
Figure supplement 3. Scores returned by LIWC (mean percentage on a log scale and 95% CI) for loyalty/betrayal, one of the five foundations of Moral Foundations Theory.
Figure supplement 4. Scores returned by LIWC (mean percentage on a log scale and 95% CI) for authority/subversion, one of the five foundations of Moral Foundations Theory.
Figure supplement 5. Scores returned by LIWC (mean percentage on a log scale and 95% CI) for sanctity/degradation, one of the five foundations of Moral Foundations Theory.

words), present tense words, and relativity words, and fewer non-personal words and modal words (***Pennebaker et al., 2015***). Low authenticity and high analytical tone are characteristic of texts describing medical research (***Karačić et al., 2019***; ***Glonti et al., 2017***). There was some variation with reviewer recommendation in the scores returned for clout, with accept having the highest scores for clout, followed by minor revisions, major revisions and reject (***Figure 2B***; ***Figure 2—figure supplement 1B***).

When reviewers recommended major revisions, the text of the report was more analytical. The analytical tone was higher when reviewers were women and for single-blind peer review, but we did not find any effect of the area of research (see Table 4 in ***Supplementary file 2***).

Clout levels varied with area of research, with the highest levels in social sciences and economics journals (see Table 7 in ***Supplementary file 3***). When reviewers recommended rejection, the text showed low levels of clout, as it did when reviewers were men and when the journal useded single-blind peer review (see Table 7 in ***Supplementary file 3***).

The text of reports in social sciences and economics journals had the highest levels of authenticity. Authenticity was prevalent also when reviewers recommended rejection. There was no significant variation in terms of authenticity per reviewer gender or type of peer review (see Table 10 in ***Supplementary file 4***).

Sentiment analysis

The three approaches were used to perform sentiment analysis on our sample – LIWC, SentimentR and the Stanford CoreNLP – produced similar results. Reports recommending accept had the highest scores, indicating higher sentiment, followed by reports recommending minor revisions, major revisions and reject (***Figure 3***; ***Figure 3—figure supplement 1***). Furthermore, reports for social sciences and economics journals had the highest levels of sentiment, as did reviews written by women. We did not find any association between sentiment and the type of peer review (see Table 13 in ***Supplementary file 5***, Table 16 in ***Supplementary file 6*** and Table 19 in ***Supplementary file 7***).

Moral foundations

LIWC was also used to explore the morality of the reports in our sample (***Graham et al., 2009***). The differences between peer review recommendations were statistically significant. Reports recommending acceptance had the highest scores for general morality, followed by reports recommending minor revisions, major revisions and reject (***Figure 4***). Regarding the research area, we found a lowest proportion of words related to morality in the social sciences and economics, when reviewers were men, and when single-blind peer review was used (***Figure 4***).

We also explored five foundations of morality – care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation – but no clear patterns emerged (***Figure 4—figure supplements 1–5***). See the Methods section for more details, and ***Supplementary file 8*** for lists of the ten most common phrases from the LIWC Moral Foundation dictionary. In general, the prevalence of these words was minimal, with average scores lower than 1%. Moreover, these words tended to be part of common phrases and thus did not speak to the moral content of the reviews. This suggests that a combination of qualitative and quantitative methods, including machine learning tools, will be required to explore the moral aspects of peer review.

Conclusion

Our study suggests that the reviewer recommendation has the biggest influence on the linguistic characteristics (and length) of peer review reports, which is consistent with previous, case-based research (***Casnici et al., 2017***). It is probable that whenever reviewers recommend revision, they write a longer report in order to justify their requests and/or to suggest changes to improve the manuscript (which they do not have to do when they recommend to accept or reject). In our study, in the case of the two more negative recommendations (reject and major revisions), the reports were shorter, and language was less emotional and more analytical. We found that the type of peer review – single-blind or double-blind – had no significant influence on the reports, contrary to previous reports on smaller samples (***Bravo et al., 2019***; ***van Rooyen et al., 1999***). Likewise, area of research had no significant influence on the reports in the sample, and neither did reviewer gender, which is consistent with a previous smaller study (***Bravo et al., 2019***). The lack of influence exerted by the area of research, the type of peer review or the reviewer gender on the linguistic characteristics of the reports is a sign of the robustness of peer review.

The results of our study should be considered in the light of certain limitations. Most of the journals were in the health and medical sciences and the physical sciences, and most used single-blind peer review. However, the size, depth and uniqueness of our dataset helped us provide a more comprehensive analysis of peer review reports than previous studies, which were often limited to small samples and incomplete data (***van den Besselaar et al., 2018***; ***Sizo et al., 2019***; ***Falk Delgado et al., 2019***). Future research would also benefit from baseline data against which results could be compared, although our results match the preliminary results from a study at a single biomedical journal (***Glonti et al., 2017***), and from knowing more about the referees (such as their status or expertise). Finally, we did not examine the actual content of the manuscripts under review, so we could not determine how reliable reviewers were in their assessments. Combining language analyses of peer review reports with estimates of peer review reliability for the same manuscripts (via inter-reviewer ratings) could provide new insights into the peer review process.

Methods

The PEERE dataset

PEERE is a collaboration between publishers and researchers (***Squazzoni et al., 2020***), and the PEERE dataset contains 583,365 peer review reports from 61 journal published by Elsevier, with data on reviewer recommendation, area of research (health and medical sciences; life sciences; physical sciences; social sciences and economics), type of peer review (single blind or double blind), and reviewer gender for each report. Most of the reports (N = 481,961) are for original research papers, with the rest (N = 101,404) being for opinion pieces, editorials and letters to the editor. The database was first filtered to exclude reviews that included reference to manuscript revisions, resulting in 583,365 reports. We eliminated 110,636 due to the impossibility to determine reviewer gender, and 260 because we did not have data on the recommendation. Our analysis was performed on a total number of 472,449 peer review reports.

Gender determination

To determine reviewer gender, we followed a standard disambiguation algorithm that has already been validated on a dataset of scientists extracted from the Web of Science database covering a similar publication time window (***Santamaría and Mihaljević, 2018***). Gender was assigned following a multi-stage gender inference procedure consisting of three steps. First, we performed a preliminary gender determination using, when available, gender salutation (i.e., Mr, Mrs, Ms...). Secondly, we queried the Python package gender-guesser about the extracted first names and country of origin, if any. Gender-guesser has demonstrated to achieve the lowest misclassification rate and introduce the smallest gender bias (***Paltridge, 2017***). Lastly, we queried the best performer gender inference service, Gender API (https://gender-api.com/), and used the returned gender whenever we found a minimum of 62 samples with, at least, 57% accuracy, which follows the optimal values found in benchmark 2 of the previous research (***Santamaría and Mihaljević, 2018***). This threshold for the obtained confidence parameters was suitable to ensure that the rate of misclassified names did not exceed 5% (***Santamaría and Mihaljević, 2018***). This allowed us to determine the gender of 81.1% of reviewers, among which 75.9% were male and 24.1% female. With regards to the three possible gender sources, 6.3% of genders came from scientist salutation, 77.2% from gender-guesser, and 16.5% from the Gender API. The remaining 18.9% of reviewers were assigned an unknown gender. This level of gender determination is consistent with the non-classification rate for names of scientists in previous research (***Santamaría and Mihaljević, 2018***).

Analytical tone, authenticity and clout

We used a version of the Linguistic Inquiry and Word Count (LIWC) text-analysis software with standardized scores (http://liwc.wpengine.com/) to analyze the peer review reports in our sample. LIWC measures the percentage of words related to three psychological features (so scores range from 0 to 100): 'analytical tone'; 'clout'; and "authenticity. A high score for analytical tone indicates a report with a logical and hierarchical style of writing. Clout reveals personal sensitivity towards social status, confidence or leadership: a low score for clout is associated with insecurities and a less confident and more tentative tone (***Kacewicz et al., 2014***). A high score for authenticity indicates a report written in a style that is honest and humble, whereas a low score indicates a style that is deceptive and superficial (***Pennebaker et al., 2015***). The words people use also reflect how authentic or personal they sound. People who are authentic tend to use more I-words (e.g. I, me, mine), present-tense verbs, and relativity words (e.g. near, new) and fewer she-he words (e.g. his, her) and discrepancies (e.g. should, could) (***Pennebaker et al., 2015***).

Sentiment analysis

We used three different methodological approaches to assess sentiment. (i) LIWC measures 'emotional tone', which indicates writing dominated by either positive or negative emotions by counting number of words from a pre-specified dictionary. (ii) The SentimentR package (***Rinker, 2019***) classifies the proportion of words related to sentiment in the text, similarly to the 'emotional tone' scores in LIWC but using a different vocabulary. The SentimentR score is the valence of words related with the specific sentiment, majority of scores (99.97%) ranging from −1 (negative sentiment) +1 (positive sentiment). (iii) Stanford CoreNLP is a deep language analysis program that uses machine learning to determine the emotional valence of the text (***Socher et al., 2013***), and score ranges from 0 (negative sentiment) to +4 (positive sentiment). Examples of characteristic text variables from the peer review reports analysed with these approaches are given in ***Supplementary files 5–7***.

Moral foundations

We used LIWC and Moral Foundations Theory (https://moralfoundations.org/other-materials/) to analyse the reports in our sample according to five moral foundations: care/harm (also known as care-virtue/care-vice); fairness/cheating (or fairness-virtue/fairness-vice); loyalty/betrayal (or loyalty-virtue/loyalty-vice); authority/subversion (authority virtue/authority-vice); and sanctity/degradation (or sanctity-virtue/sanctity-vice).

Statistical methods

Data were analysed using the R programming language, version 3.6.3. (***R Development Core Team, 2017***). To test the interaction effects and compare different peer review characteristics, we conducted a mixed model linear analysis on each variable (analytical tone, authenticity, clout; the measures of sentiment; and the measures of morality) with reviewer recommendation, area of research, type of peer review (single- or double-blind) and reviewer gender as fixed factors (predictors) and the journal, word count and article type as the random factor. This was to control across-journal interactions, number of words and article type.

Acknowledgements

We thank Dr Bahar Mehmani from Elsevier for helping us with data collection.

Ivan Buljan is in the Department of Research in Biomedicine and Health, University of Split School of Medicine, Split, Croatia
ibuljan@mefst.hr
https://orcid.org/0000-0002-8719-7277

Daniel Garcia-Costa is in the Department d'Informàtica, University of Valencia, Burjassot-València, Spain
https://orcid.org/0000-0002-8939-8451

Francisco Grimaldo is in the Department d'Informàtica, University of Valencia, Burjassot-València, Spain
https://orcid.org/0000-0002-1357-7170

Flaminio Squazzoni is in the Department of Social and Political Sciences, University of Milan, Milan, Italy
https://orcid.org/0000-0002-6503-6077

Ana Marušić is in the Department of Research in Biomedicine and Health, University of Split School of Medicine, Split, Croatia
https://orcid.org/0000-0001-6272-0917

Author contributions: Ivan Buljan, Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Daniel Garcia-Costa, Data curation, Software, Formal analysis, Investigation, Visualization, Writing - original draft, Writing - review and editing; Francisco Grimaldo, Conceptualization, Data curation, Software, Formal analysis, Funding acquisition, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Flaminio Squazzoni, Conceptualization, Resources, Data curation, Supervision, Investigation, Methodology, Writing - original draft, Project administration, Writing - review and editing; Ana Marušić, Conceptualization, Resources, Supervision, Funding acquisition, Methodology, Writing - original draft, Project administration

Received 01 November 2019
Accepted 16 July 2020
Published 17 July 2020

Funding

Funder	Grant reference number	Author
Ministerio de Ciencia e Innovación	RTI2018-095820-B-I00	Daniel Garcia-Costa Francisco Grimaldo
Spanish Agencia Estatal de Investigación	RTI2018-095820-B-I00	Daniel Garcia-Costa Francisco Grimaldo
European Regional Development Fund	RTI2018-095820-B-I00	Daniel Garcia-Costa Francisco Grimaldo
Croatian Science Foundation	IP-2019-04-4882	Ana Marušić

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.53249.sa1
Author response https://doi.org/10.7554/eLife.53249.sa2

Additional files

Supplementary files

- Supplementary file 1. Word count (***Figure 1***): summary data and mixed model linear regression coefficients and residuals.
- Supplementary file 2. Analytical tone (***Figure 2A***): summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for LIWC analytical tone.
- Supplementary file 3. Clout (***Figure 2B***): summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for LIWC clout.
- Supplementary file 4. Authenticity (***Figure 2C***): summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for LIWC authenticity.

- Supplementary file 5. Sentiment/LIWC emotional tone (*Figure 3A*): summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for sentiment (LIWC emotional tone).
- Supplementary file 6. Sentiment/SentimentR score (*Figure 3B*): summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for sentiment (SentimentR scores).
- Supplementary file 7. Sentiment/Stanford CoreNLP score (*Figure 3C*): summary data, mixed model linear regression coefficients and residuals, and examples of reports with high and low scores for sentiment (Stanford CoreNLP score).
- Supplementary file 8. Ten most frequent words found in peer review reports for general morality and the five moral foundation variables.
- Transparent reporting form

Data availability

The journal dataset required a data sharing agreement to be established between authors and publishers. A protocol on data sharing entitled 'TD1306 COST Action New frontiers of peer review (PEERE) PEERE policy on data sharing on peer review' was signed by all partners involved in this research on 1 March 2017, as part of a collaborative project funded by the EU Commission. The protocol established rules and practices for data sharing from a sample of scholarly journals, which included a specific data management policy, including data minimization, retention and storage, privacy impact assessment, anonymization, and dissemination. The protocol required that data access and use were restricted to the authors of this manuscript and data aggregation and report were done in such a way to avoid any identification of publishers, journals or individual records involved. The protocol was written to protect the interests of any stakeholder involved, including publishers, journal editors and academic scholars, who could be potentially acted by data sharing, use and release. The full version of the protocol is available on the peere.org website. To request additional information on the dataset and for any claim or objection, please contact the PEERE data controller at info@peere.org.

References

Bornmann L, Wolf M, Daniel H-D. 2012. Closed versus open reviewing of journal manuscripts: how far do comments differ in language use? *Scientometrics* **91**:843–856. DOI: https://doi.org/10.1007/s11192-011-0569-5

Bravo G, Grimaldo F, López-Iñesta E, Mehmani B, Squazzoni F. 2019. The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications* **10**:322. DOI: https://doi.org/10.1038/s41467-018-08250-2

Casnici N, Grimaldo F, Gilbert N, Squazzoni F. 2017. Attitudes of referees in a multidisciplinary journal: An empirical analysis. *Journal of the Association for Information Science and Technology* **68**:1763–1771. DOI: https://doi.org/10.1002/asi.23665

Falk Delgado A, Garretson G, Falk Delgado A. 2019. The language of peer review reports on articles published in the BMJ, 2014–2017: an observational study. *Scientometrics* **120**:1225–1235. DOI: https://doi.org/10.1007/s11192-019-03160-6

Fyfe A, Squazzoni F, Torny D, Dondio P. 2020. Managing the growth of peer review at the Royal Society journals, 1865-1965. *Science, Technology & Human Values* **45**:405–429. DOI: https://doi.org/10.1177/0162243919862868

Garg N, Schiebinger L, Jurafsky D, Zou J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS* **115**:E3635–E3644. DOI: https://doi.org/10.1073/pnas.1720347115

Glonti K, Hren D, Carter S, Schroter S. 2017. Linguistic features in peer reviewer reports: how peer reviewers communicate their recommendations. *Proceedings of the International Congress on Peer Review and Scientific Publication*. https://peerreviewcongress.org/prc17-0234 [Accessed April 20, 2020].

Graham J, Haidt J, Nosek BA. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* **96**:1029–1046. DOI: https://doi.org/10.1037/a0015141, PMID: 19379034

Grimaldo F, Marušić A, Squazzoni F. 2018. Fragments of peer review: A quantitative analysis of the literature (1969-2015). *PLOS ONE* **13**:e0193148. DOI: https://doi.org/10.1371/journal.pone.0193148

Haffar S, Bazerbachi F, Murad MH. 2019. Peer review bias: a critical review. *Mayo Clinic Proceedings* **94**:670–676. DOI: https://doi.org/10.1016/j.mayocp.2018.09.004, PMID: 30797567

Hartley J, Pennebaker JW, Fox C. 2003. Abstracts, introductions and discussions: how far do they differ in style? *Scientometrics* **57**:389–398. DOI: https://doi.org/10.1023/A:1025008802657

Hengel E. 2018. Publishing while female: are women held to higher standards? Evidence from peer review. University of Cambridge. https://www.repository.cam.ac.uk/handle/1810/270621

Kacewicz E, Pennebaker JW, Davis M, Jeon M, Graesser AC. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology* **33**:125–143. DOI: https://doi.org/10.1177/0261927X13502654

Karačić J, Dondio P, Buljan I, Hren D, Marušić A. 2019. Languages for different health information readers: multitrait-multimethod content analysis of Cochrane systematic reviews textual summary formats. *BMC Medical Research Methodology* **19**:75. DOI: https://doi.org/10.1186/s12874-019-0716-x

Lee CJ, Sugimoto CR, Zhang G, Cronin B. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology* **64**:2–17. DOI: https://doi.org/10.1002/asi.22784

Magua W, Zhu X, Bhattacharya A, Filut A, Potvien A, Leatherberry R, Lee YG, Jens M, Malikireddy D, Carnes M, Kaatz A. 2017. Are female applicants disadvantaged in National Institutes of Health peer review? Combining algorithmic text mining and qualitative methods to detect evaluative differences in R01 reviewers' critiques. *Journal of Women's Health* **26**:560–570. DOI: https://doi.org/10.1089/jwh.2016.6021, PMID: 28281870

Markowitz DM, Hancock JT. 2016. Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology* **35**:435–445. DOI: https://doi.org/10.1177/0261927X15614605

Marsh HW, Jayasinghe UW, Bond NW. 2011. Gender differences in peer reviews of grant applications: A substantive-methodological synergy in support of the null hypothesis model. *Journal of Informetrics* **5**:167–180. DOI: https://doi.org/10.1016/j.joi.2010.10.004
Paltridge B. 2017. *The Discourse of Peer Review: Reviewing Submissions to Academic Journals*. London: Palgrave Macmillan. DOI: https://doi.org/10.1057/978-1-137-48736-0
Pennebaker JW, Boyd RL, Jordan K, Blackburn K. 2015. The development and psychometric properties of LIWC2015. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf [Accessed July 18, 2020].
Pennebaker JW. 2017. Mind mapping: Using everyday language to explore social & psychological processes. *Procedia Computer Science* **118**:100–107. DOI: https://doi.org/10.1016/j.procs.2017.11.150
R Development Core Team. 2017. R: a language and environment for statistical computing. 3.6.3. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/
Resnik DB, Elmore SA. 2016. Ensuring the quality, fairness, and integrity of journal peer review: a possible role of editors. *Science and Engineering Ethics* **22**:169–188. DOI: https://doi.org/10.1007/s11948-015-9625-5, PMID: 25633924
Rinker TW. 2019. sentimentr: Calculate text polarity sentiment. *Github*. version 2.7.1. http://github.com/trinker/sentimentr
Santamaría L, Mihaljević H. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* **4**:e156. DOI: https://doi.org/10.7717/peerj-cs.156
Sizo A, Lino A, Reis LP, Rocha Á. 2019. An overview of assessing the quality of peer review reports of scientific articles. *International Journal of Information Management* **46**:286–293. DOI: https://doi.org/10.1016/j.ijinfomgt.2018.07.002
Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing 1631–1642.
Squazzoni F, Grimaldo F, Marušić A. 2017. Publishing: Journals could share peer-review data. *Nature* **546**:352. DOI: https://doi.org/10.1038/546352a
Squazzoni F, Ahrweiler P, Barros T, Bianchi F, Birukou A, Blom HJJ, Bravo G, Cowley S, Dignum V, Dondio P, Grimaldo F, Haire L, Hoyt J, Hurst P, Lammey R, MacCallum C, Marušić A, Mehmani B, Murray H, Nicholas D, et al. 2020. Unlock ways to share data on peer review. *Nature* **578**:512–514. DOI: https://doi.org/10.1038/d41586-020-00500-y
van den Besselaar P, Sandström U, Schiffbaenker H. 2018. Studying grant decision-making: a linguistic analysis of review reports. *Scientometrics* **117**:313–329. DOI: https://doi.org/10.1007/s11192-018-2848-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**:23–27. DOI: https://doi.org/10.1136/bmj.318.7175.23

Apéndice B

Measuring the effect of reviewers on manuscript change: A study on a sample of submissions to Royal Society journals (2006–2017)

Journal of Informetrics 16 (2022) 101316

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Measuring the effect of reviewers on manuscript change: A study on a sample of submissions to Royal Society journals (2006–2017)

Federico Bianchi[a], Daniel García-Costa[b], Francisco Grimaldo[b], Flaminio Squazzoni[a,*]

[a] *Department of Social and Political Sciences, University of Milan, Via Conservatorio 7, Milan, 20122, Italy*
[b] *Department of Computer Science, University of Valencia, Avinguda de la Universitat s/n, Burjassot, 46100, Spain*

ARTICLE INFO

Keywords:
Peer review
Journals
Manuscripts
Reviewers
Manuscript changes

ABSTRACT

Peer review is key for public trust of academic journals. It ensures that only rigorous research is published but also helps authors to increase the value of their manuscripts through feedback from reviewers. However, measuring the developmental value of peer review is difficult as it requires fine-grained manuscript data on various stages of the editorial process, which are rarely available. To fill this gap, we accessed complete data from Royal Society journals from 2006 to 2017, and measured manuscript changes during peer review from their initial submissions. We then estimated the effect of the number of reviewers and the evaluation of reviewers on manuscript development and their citations after publication. We found that the number of reviewers had an almost linear effect on manuscript change although with decreasing marginal effects whenever more than two reviewers were involved. This effect did not depend on the initial quality of manuscripts. We also found that changes due to reviewers tended to increase a manuscript's probability of being cited at least once after publication. While our findings show the multiple functions of peer review for manuscript development, research with larger and more representative journal samples is needed to develop more systematic measures that reflect the complexity of peer review.

1. Introduction

The digital age has witnessed an explosion of the means of scientific dissemination (Tennant et al., 2017). The proliferation of preprints, websites and online repositories has contributed to enhance the curation function of academic journals for scientific records (Squazzoni et al., 2020). The fact that we consider journals as synonymous of the quality of scientific records depends on the rigour of their internal evaluation standards and their capacity of adding value to submitted manuscripts (Baldwin, 2018). These standards can be achieved only if journals ensure rigorous selection of manuscripts and improve them through intensive collaboration between authors, reviewers and editors (Bornmann, 2011). Indeed, collaboration between editors, board members and external experts has greatly varied over time. This in turn has ensured that manuscript quality-screening and improvements have always been an intrinsic part of peer review at least since the 1950s in many research areas (Fyfe et al., 2020; Merriman, 2020; Moxham and Fyfe, 2018).

Understanding whether and how these activities are performed by journals requires the examination of a variety of complex factors (Publons, 2018). Screening manuscripts and weeding out low-quality research require the involvement of reviewers and editors, who reflect the best standards of research (Siler et al., 2015). Developing manuscripts depends on a journal's capacity to create contexts within which a constructive dialogue between reviewers and authors is both fair and disinterested (Dondio et al., 2019).

* Corresponding author.
E-mail address: flaminio.squazzoni@unimi.it (F. Squazzoni).

https://doi.org/10.1016/j.joi.2022.101316
Received 11 January 2022; Received in revised form 17 June 2022; Accepted 18 July 2022
Available online 1 August 2022

Unfortunately, examining these factors jointly and empirically is difficult for various reasons, the most significant of which is the lack of fine-grained data from journals. For instance, research on peer review reports from repositories, such as Publons, helps to identify certain socio-demographic characteristics of reviewers and the choice of journals for which scholars typically review (Severin et al., 2021) or the connection between peer review activities and research productivity (Ortega, 2017). Recent research on a sample of peer review reports from Elsevier journals reconstructed the linguistic characteristics of reports depending on the type of recommendations and certain reviewer characteristics (Buljan et al., 2020). Similarly, a recent study on a large-sample of reports from Elsevier journals found interesting heterogeneity in standards of reports depending on reviewer characteristics and areas of research (Garcia-Costa et al., 2022). However, interlinking reports and manuscripts is impossible with a peer review report database, thus undermining the possibility of gauging the effect of peer review on manuscripts and on the journals themselves.

Research on the screening function of peer review typically concentrates on the reviewers' capability of predicting the quality of manuscripts (Casnici et al., 2017). It has generally used ex-post measurements as an indirect proxy of reviewer reliability, including citations of different versions of manuscripts, e.g., published articles vs. rejected manuscripts later published in other journals, as well as differences in the impact factors of journals rejecting/publishing different versions of the same manuscripts (Rigby et al., 2018). Unfortunately, only rarely have these studies included data on peer review reports and tracked manuscript change within the editorial process.

We believe that this is key to assess the developmental value of peer review as it allows us to examine how manuscripts change throughout the process of peer review (Atjonen, 2019; Bedeian, 2004; Matsui et al., 2021; Rigby et al., 2018; Teplitskiy, 2016). For instance, the tendency of reducing the curation function of peer review to the goal of identifying impactful manuscripts via post-publication indicators (e.g., altmetrics, citations and other indicators), does not help to assess the quality of internal journal processes (Pontille and Torny, 2015; Seeber, 2020). However, without measuring how and how much manuscripts change throughout the process due to reviewer feedback, it is impossible to understand whether peer review adds anything relevant to the final manuscripts (Cowley, 2015).

Research examining these factors jointly is also essential to understand how journals harmonise different peer review functions for the benefit of their various stakeholders. The mechanics of peer review implies at least a triadic relationship with various expectations (Lugosi, 2021). Editors rely on reviewers to avoid publishing manuscripts of low quality and defend the prestige and position of their journals in a competitive, continually evolving environment (Liu et al., 2018; Taşkın et al., 2021). Authors expect that reviewers share constructive feedback for manuscript improvements, even when their manuscript is eventually rejected (Huisman and Smits, 2017). Reviewers expect authors to consider their comments and suggestions seriously to avoid being exploited while enforcing the highest scientific standards (Horbach and Halffman, 2018). The biases and inefficiencies of peer review are presently under the spotlight (Squazzoni et al., 2021; Tomkins et al., 2017) and many publishers are exploring innovative models to increase the transparency and accountability of the process, e.g., open peer or post-publication peer review, which require careful assessment (Eyre-Walker and Stoletzki, 2013; Harms and Credé, 2020; Thelwall et al., 2021). Thus, understanding manuscript change during peer review with data from multiple journals – and not only from individual cases (Grimaldo et al., 2018) – can help us evaluate the importance of this fundamental academic institution more systematically (Horbach, 2021; Tennant and Ross-Hellauer, 2020).

Our paper aims to contribute to empirical research on peer review by presenting an explorative measurement of the developmental function of peer review. While previous research has investigated only specific journals and only rarely with complete data on manuscripts from each stage of the editorial process (Matsui et al., 2021; Teplitskiy, 2016), here, we have tested manuscript change during the editorial process with a large-scale, across-journal dataset and estimated possible effects on article citations. We aimed to test the effect of the number of reviewers and their evaluation on manuscript change within the editorial process and on later citations.

For this study, we first signed a confidential data sharing agreement with The Royal Society (Squazzoni et al., 2017), the world's oldest independent scientific academy. The Royal Society pioneered the concepts and practices of academic journals, editorial responsibility and peer review (Fyfe et al., 2015). Their journals include 11 titles, including *Philosophical Transactions A* and *Proceedings A*, which publish research on physical, mathematical and engineering sciences, *Philosophical Transactions B, Proceedings B* and *Biology Letters*, with a readership in biological sciences, as well as cross-disciplinary outlets, such as *Interface*, for cross-disciplinary research at the interface between the physical and life sciences, and *Royal Society Open Science*, the Royal Society's most recent open access journal in science, engineering and mathematics.

This agreement permitted us to collect complete and fully comparable temporal data on their journals from 2006 to 2017, including more than 10,000 manuscripts (see Methods). In order to ensure full comparability in terms of type of manuscripts and journals, we excluded all manuscripts submitted to the following four journals: *Open Biology, Interface Focus, Notes and Records* and *Biographical Memoirs*. Manuscripts from these journals were only weakly comparable with the rest of the sample, being mostly commentaries, short notes or reviews rather than research articles. We also restricted our sample to research articles, thus excluding any comments, reviews or notes.

After transforming all manuscript and review files of various format into text files, we calculated the Levenshtein distance (Levenshtein, 1966) between different versions of manuscripts to track any changes occurring throughout the process. Following Bravo et al. (2018), we built a *review score* that measured reviewer recommendations for each manuscript consistently, regardless of the different number of reviewers and rounds of reviews per manuscript. We considered this as a proxy of the initial quality of manuscripts as perceived by reviewers. We also calculated citations of published manuscripts to check whether changes during peer review could increase an article's probability of being cited after publication.

2. Methods

Our dataset included 10,996 manuscripts submitted to seven journals from the Royal Society from 2006 to 2017. Data included complete information regarding initial and revised versions of each submitted manuscript, including full text, reviewers' recommendations and editorial decisions.

In order to quantify the length of each manuscript, we converted each document into plain text files using dedicated *Python* libraries (i.e., '*docx*' for .doc and .docx files and '*pdfminer*' for .pdf files). We removed tables, figures, marks, rare characters, page headers and footers, as well as any irrelevant marks caused by document conversion. We then removed all non-ASCII characters. We downloaded the final version of all published articles from the Royal Society website. In the case of published articles, we divided their text into different portions and excluded images, figures and tables, thus standardizing their format with their related submission files. This allowed us to assign a unique ID to different files of the same manuscript (e.g., original submissions and published articles).

We measured the *text changes* by computing the difference between the originally submitted manuscript and either the published or the rejected version. We computed the Levenshtein distance (Levenshtein, 1966) between different text versions, i.e., the number of changes needed to convert one text string into another, thus detecting any change of the text throughout the various stages of peer review. We preferred this measurement to token-based distances, such as cosine or Jaccard distance, as the latter would not have permitted us to consider certain changes, such as the syntax or rephrasing using the same words.

When calculating *text changes* with the Levenshtein distance, we also calculated the difference between the originally submitted manuscript and the final version (either the published article or the rejected manuscript) in their listed references. In order to identify references, we used various regular expressions (*regex*) which were shared by different referencing styles (e.g., IEEE, Vancouver, APA). We defined the regex to extract separately the publication year, the title and the list of authors. We then calculated a similarity ratio that considered two references as equal when: (i) both sources reported the same publication year; (ii) the cosine distance between titles was smaller than 0.1; and (iii) either both references had the same number of authors or the cosine distance between the list of authors was smaller than 0.1. We set this threshold to 0.1 after manual experimentation on the data. We used the cosine distance as any token-based distances was less sensitive to small spelling changes when comparing references.

We calculated the *reference changes* as follows:

$$1 - \frac{\textit{Number of similar references}}{\textit{Max number of references in either documents}}.$$

For the sake of interpretation, we re-scaled both *text changes* and *reference changes* to a 0–100 range.

We then calculated the *number of reviewers* for each manuscript by counting the total number of reviewers involved in all rounds of reviews. For instance, assume that in the first round, a manuscript was reviewed by reviewers 1 and 2 and that in the second round, reviewer 2 was not involved while the editor contacted reviewer 3. In these cases, we counted a total number of reviewers = 3. This was to reflect the fact that a manuscript can change due to the effect of each individual reviewer to whom it was exposed.

By following Bravo et al. (2018), we calculated a *review score* for each manuscript as a proxy of the manuscript's quality resulting from reviewers' recommendations. This score allowed us to compare the evaluation of manuscripts submitted to various journals regardless of differences in the number of reviewers per manuscript. We built a set of all possible unique combinations of recommendations for each manuscript (e.g., *{accept, accept}, {accept, minor revision}, {accept, major revision}*,..., *{reject, reject}*) and counted the number of combinations that were less favourable (*# worse*) or more favourable (*# better*) than the recommendation received by the manuscript (e.g., *{accept, accept}* was better than *{accept, major revision}*). We handled combinations which could not be considered as clearly better or worse as reported in Bravo et al. (2018, Table 2). After testing all possible (better or worse) combinations per manuscript and verifying lack of differences on the outcomes, we calculated the review scores as follows:

$$\textit{review score} = \frac{\#\ \textit{worse}}{\#\ \textit{worse} + \#\ \textit{better}}.$$

We measured inter-reviewer *agreement* by calculating the number of similar recommendations divided by the total number of reviews per manuscript at the first round (e.g., 2/3 agreement in case of three reviewers recommending *{minor revision, major revision, major revision}*). Finally, we measured the *impact* of published manuscripts by calculating the number of citations for each article using the DOI obtained from the Royal Society journal platform to query Altmetrics API on Dimensions.ai database (Khan et al., 2021).

3. Results

The length of the text of originally submitted manuscripts was highly left-skewed. The median length was 21,773 characters. Figure 1 shows that the final version of both published and rejected manuscripts changed considerably in terms of Levenshtein distance compared to their initial version. This was true for both *text changes* (M = 40.72 %, SD = 15.67 %) and *reference changes* (M = 41.33 %, SD = 21.42 %). Most manuscripts were reviewed by at least two reviewers (65.60 %), with only a minority reviewed by three or more reviewers.

We tested the impact of the *number of reviewers* on manuscript change by estimating linear mixed-effect models with random intercepts on *text changes* and *reference changes*. With regard to the former, results showed that the number of reviewers tended to increase manuscript changes (see Fig. 2a). Changes increased almost linearly with the number of reviewers. However, a greater effect was found when shifting from one to two reviewers evaluating the same manuscript in various rounds of the process. Note that whenever manuscripts were evaluated by five or more reviewers, we found decreasing marginal effects compared to the case of

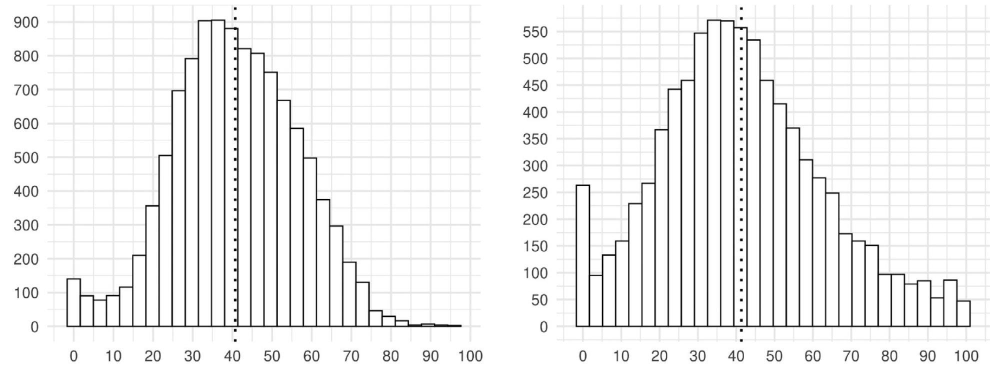

Fig. 1. Distribution of *text changes* (left) and *reference changes* (right) among sampled manuscripts, measured by the Levenshtein distance (%) between the original submission and the final version (either published articles or rejected manuscripts). Vertical dashed lines indicate mean values.

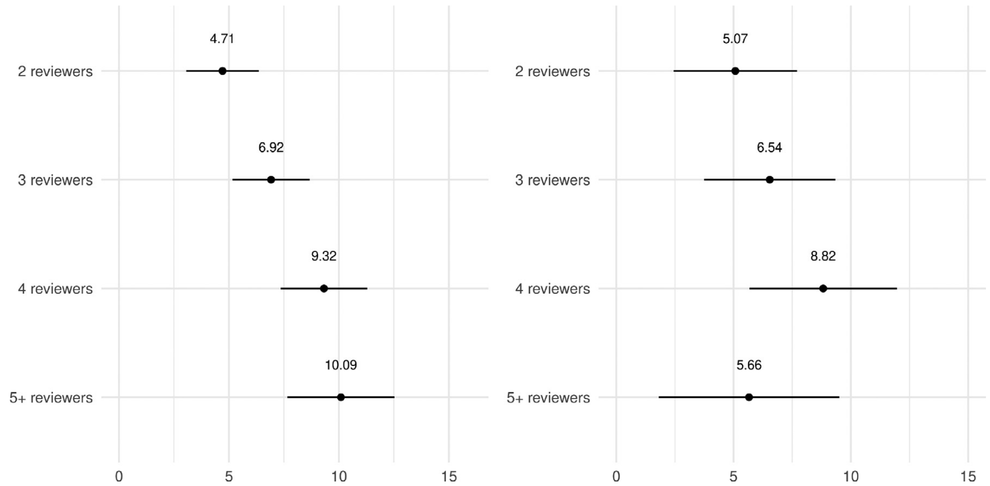

Fig. 2. Linear mixed-effects models of *text changes* (left) and *reference changes* (right), measured through Levenshtein distance (%): Estimated fixed effects of *number of reviewers* (*dots*, reference category: “1 reviewer”) with 95% confidence intervals (*lines*)). The models include all control variables presented in Table 1 and random intercepts of different journals.

manuscripts evaluated by four reviewers. We found a similarly positive effect on *reference changes* when manuscripts were assessed by up to four reviewers. Note that this effect decreased whenever manuscripts were assessed by more than four reviewers (see Fig. 2b). In both models, the effect of the *number of reviewers* was estimated by controlling for journal-specific heterogeneity (random intercepts), the length of the originally submitted manuscripts, the *review score*, i.e., the quality of manuscripts in reviewers’ opinion, and the inter-reviewer agreement (see Table 1).

With regard to the effect of manuscript change on published articles’ *impact*, Fig. 3 shows that the distribution of manuscripts cited at least one time after being published was relatively heterogeneous across the journals. Overall, the average number of citations was 22.64 (SD = 39.82).

Tables 2 and 3 show two logistic regression models estimating a small positive effect of *text changes* and *reference changes* respectively on *impact*. In both models, we controlled for differences between journals, which significantly varied in terms of impact factor, and time exposure of articles, which could affect citation dynamics. Note that the distribution of the number of citations was highly skewed (5.39), thus making linear regression models poorly informative. This led us to consider a binary variable, i.e., whether articles had received at least one citation or not. We also estimated zero-inflated negative binomial regression models (Hilbe, 2014), which suggested that evidence of a small effect of *text* and *reference changes* could be found only in changing between receiving no citations or receiving at least one (see Additional analysis for more details), adjusting for across-journal differences and years from publication. However, note that estimating the effect of changes due to peer review on citations is problematic because of other

Table 1
Linear mixed-effects models estimating the effect of the *number of reviewers* (reference category: "1 reviewer") on *text changes* and *reference changes* with journal-specific random intercepts. Note that the number of observations varied due to cases of manuscript files without correctly formatted or reported references.

	Text changes				Reference changes			
	$\hat{\beta}$	S.E.	95% C.I.	*p*	$\hat{\beta}$	S.E.	95% C.I.	*p*
Fixed effects								
2 reviewers	4.71	0.84	[3.06, 6.36]	0.00	5.07	1.34	[2.45, 7.70]	0.00
3 reviewers	6.92	0.89	[5.16, 8.67]	0.00	6.54	1.43	[3.74, 9.34]	0.00
4 reviewers	9.32	1.00	[7.36, 11.29]	0.00	8.82	1.60	[5.67, 11.96]	0.00
5+ reviewers	10.09	1.24	[7.66, 12.53]	0.00	5.66	1.96	[1.81, 9.51]	0.00
Length of original submission	0.00	0.00	[0.00, 0.00]	0.01	0.00	0.00	[0.00, 0.00]	0.03
Review score	0.04	0.01	[0.03, 0.06]	0.00	0.11	0.01	[0.09, 0.13]	0.00
Reviewer agreement	-0.06	0.01	[-0.07, -0.05]	0.00	-0.04	0.01	[-0.06, -0.02]	0.00
Constant	37.78	2.10	[33.67, 41.90]	0.00	33.83	2.77	[28.40, 39.26]	0.00
Random effects								
SD (Intercept)			4.56				5.19	
Number of observations			10,308				7,777	

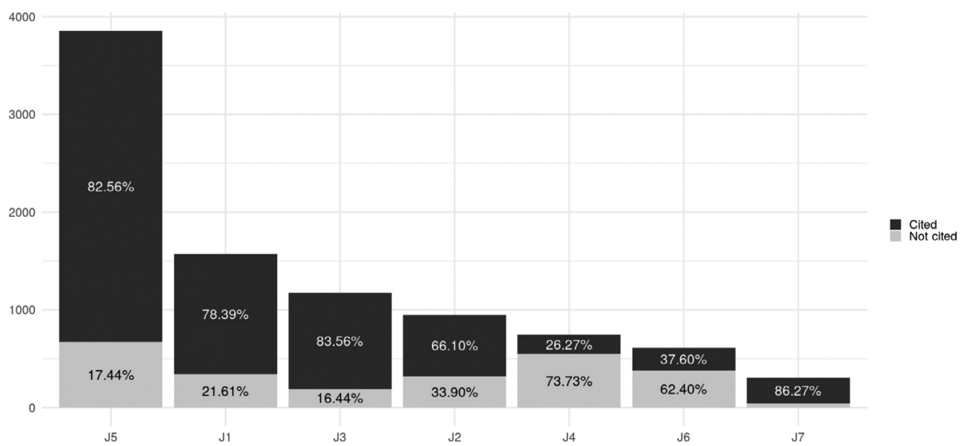

Fig. 3. Distribution of published articles which had received at least one citation (dark) vs. those which had not received any citation (light) across journals. Values reported inside bar relate to within-journal percentages.

Table 2
Logistic regression model estimating the effect of *text changes* on an article's probability of being cited at least once after publication. (Reference category of *journal*: "Journal 1").

	Odds	C.I. 95%	S.E.	*p*
Text changes	1.01	[1.01, 1.02]	(0.00)	0.00
Review score	1.00	[1.00, 1.01]	(0.00)	0.08
Journal 2	0.32	[0.25, 0.40]	(0.12)	0.00
Journal 3	0.28	[0.22, 0.36]	(0.13)	0.00
Journal 4	0.04	[0.03, 0.05]	(0.14)	0.00
Journal 5	1.29	[1.09, 1.52]	(0.08)	0.00
Journal 6	0.03	[0.02, 0.04]	(0.16)	0.00
Journal 7	0.59	[0.33, 1.15]	(0.31)	0.09
Years published	0.69	[0.68, 0.71]	(0.01)	0.00
Constant	25.94	[17.93, 37.67]	(0.19)	0.00
Number of observations		8,589		
Log likelihood		-3429.04		

Table 3
Logistic regression model estimating the effect of *reference changes* on an article's probability of being cited at least once after publication. (Reference category of *journal*: "Journal 1").

	Odds	C.I. 95%	S.E.	*p*
Reference changes	1.01	[1.01, 1.01]	(0.00)	0.00
Review score	1.00	[1.00, 1.00]	(0.00)	0.32
Journal 2	0.32	[0.25, 0.42]	(0.13)	0.00
Journal 3	0.41	[0.31, 0.55]	(0.15)	0.00
Journal 4	0.04	[0.03, 0.06]	(0.02)	0.00
Journal 5	1.43	[1.20, 1.70]	(0.09)	0.00
Journal 6	0.04	[0.03, 0.06]	(0.02)	0.00
Journal 7	0.67	[0.36, 1.35]	(0.33)	0.22
Years published	0.71	[0.69, 0.72]	(0.01)	0.00
Constant	24.42	[17.02, 35.23]	(0.19)	0.00
Number of observations		6,653		
Log likelihood		-2844.44		

possible confounding factors, including authors' reputation or particular characteristics of the published study (e.g., the popularity of the topic).

4. Discussion and conclusions

The credibility of academic journals greatly depends on the quality of peer review (Bornmann, 2011; Edwards and Siddhartha, 2017; Kharasch et al., 2021). Screening manuscripts without providing constructive feedback to authors to help them improving their manuscripts is not a good practice, especially whenever journals must ensure that only rigorous science is published (Atjonen, 2019; Teplitskiy, 2016). Although this may come at the price of delaying publications, constructive and elaborated peer review is also key for expert learning (Rigby et al., 2018).

Our study contributes to research on the developmental function of peer review (Atjonen, 2019; Garcia-Costa et al., 2022; Matsui et al., 2021; Seeber, 2020; Strang and Siler, 2015) by exploring a large dataset of manuscripts, editorial decisions and peer review outcomes from journals from the Royal Society. Our results showed that reviewers had a considerable impact on manuscript changes. Exposing manuscripts to reviewer evaluations in various peer review rounds led to an average level of about 40% of changes in manuscript text and references. Manuscript change tended to increase with the number of reviewers assessing the same manuscript and this effect was independent of the initial quality of manuscripts. Not only were manuscripts of moderate initial quality improved during peer review, but also manuscripts initially receiving more positive evaluations from reviewers, as well as those determining lowest inter-reviewer agreement, were refined and changed throughout the process. Furthermore, this effect was found regardless of any journal specificity.

Unfortunately, our analysis could not focus on details on the content of reviewer requests. While reference changes would indicate that reviewers requested authors to add relevant literature, only a linguistic analysis of the content of reports could help us to disentangle requests for conceptual developments or methodological improvements. A comprehensive analysis would also require us to match requests by reviewers and revisions made by authors, which could be made only by reducing the sample size at the expense of generalisation (Eve et al., 2021).

With all due caveats regarding possible confounding factors, we found that manuscript changes increased the probability that a published article was cited at least once after publication. However, this finding should be considered with caution. Previous research has showed mixed evidence on the link between peer review and article citations, suggesting that reviewers do not systematically predict the future impact of articles in terms of citations (Teplitskiy, 2016). The essential element of developmental peer review is to help authors improve their manuscripts, whereas the impact of articles depends on various factors (Coupé, 2013). For instance, it is difficult to estimate whether citations of manuscripts are related to the quality of manuscripts as outcome of peer review, the reputation of authors or to interest in the manuscripts' topics (Seeber, 2020). Here, future research on the developmental function of peer review should consider these complex factors more systematically, although fine-grained data required to study these aspects are rarely available, e.g., the integration of journal data with scientific records of authors prior to submitting their manuscripts (Squazzoni et al., 2020).

Finally, as suggested by a recent systematic review on experimental interventions (Gaudino et al., 2021), improving the developmental function of peer review calls for problems of sustainability and publication time delay (Merrill, 2014). There is a clear trade-off between peer review functions, including quality and efficient use of reviewer time (Bianchi et al., 2018). Unfortunately, there is still scant knowledge on these multiple functions of peer review, including the effect of reviewer guidelines (or lack of), the role ambiguity of editors and reviewers with often unclear editorial decision-making responsibility (Seeber, 2020; Song et al., 2021; Tennant and Ross-Hellauer, 2020). More research is needed to assess these trade-offs and examine the effect of peer review on the quality and recognition of manuscripts. This will mostly depend on our collective capability of removing obstacles of data sharing between publishers, journals and the scientific community (Squazzoni et al., 2020).

Data accessibility

The dataset for full replication of our study is provided here: https://dataverse.harvard.edu/privateurl.xhtml?token=6bde093d-dc44-4702-92ef-741f2e166e83. As mentioned in the text, data access required a confidentiality agreement to be signed with the Royal Society, which included journal anonymization.

Funding

DG-C and FG were partially supported by the Spanish Ministry of Science, Innovation and Universities (MCIU), the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF) under project RTI2018-095820-B-I00. FB was supported by a PRIN-MUR (Progetti di Rilevante Interesse Nazionale – Italian Ministry of University and Research) grant (Grant Number: 20178TRM3F001 "14All"). FS was supported by a Transition Grant from the University of Milan (PSR2015-17).

Declaration of Competing Interest

The authors declare we have no competing interests.

CRediT authorship contribution statement

Federico Bianchi: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Daniel García-Costa:** Conceptualization, Formal analysis, Data curation, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Francisco Grimaldo:** Conceptualization, Resources, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Writing – review & editing. **Flaminio Squazzoni:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Acknowledgements

We gratefully acknowledge Phil Hurst and the team of The Royal Society for providing data and covering the cost of their extraction from manuscript submission systems. We would like to thank the journal reviewers for helpful comments.

Appendix A. Additional analysis

Figure A.1 (left) shows the distribution of the number of citations received by published articles. The average number of citations was 22.87 (SD = 39.93), while the median number was 11. The right side of Fig. A.1 shows the distribution of the log-transformed

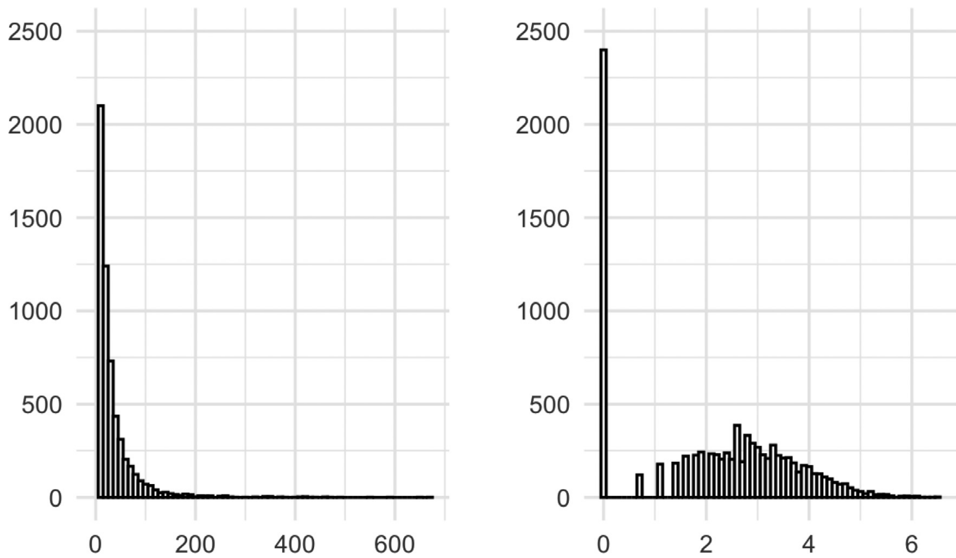

Fig. A.1. Distribution of the number of citations (left) and the logarithmic transformation (right) among published articles.

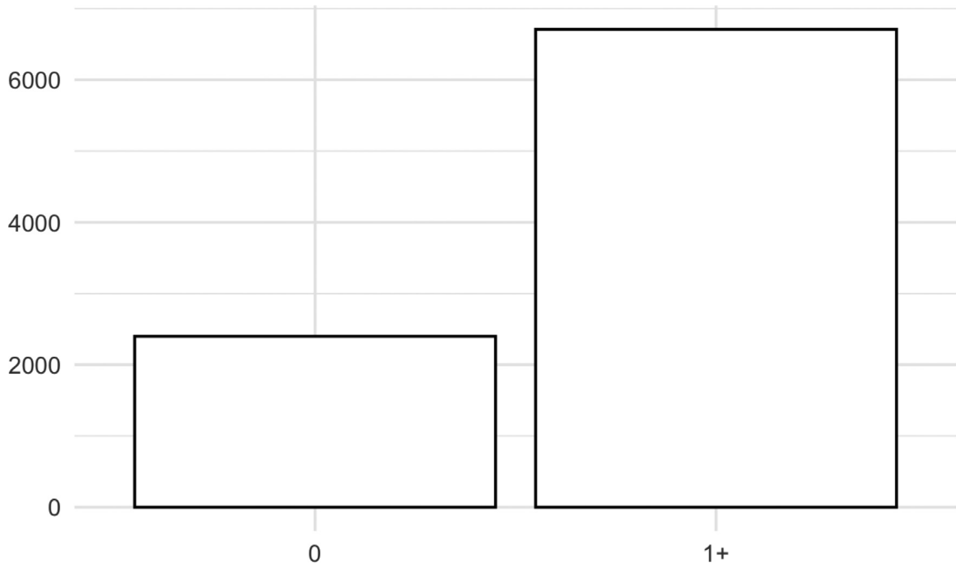

Fig. A.2. Number of published articles with zero vs. at least one citations.

Table A.1
Zero-Inflation Negative Binomial Regression model of the number of citations as a function of *text changes* and the same covariates as in Table 2.

	Count model			Zero-inflation model		
	$\hat{\beta}$	S.E.	Pr(> |z|)	$\hat{\beta}$	S.E.	Pr(> |z|)
Text changes	0.00	0.00	0.32	-0.01	0.00	0.00
Review score	0.00	0.00	0.00	0.00	0.00	0.12
Journal 2	0.08	0.04	0.06	1.10	0.12	0.00
Journal 3	-0.08	0.04	0.03	1.11	0.15	0.00
Journal 4	-0.19	0.07	0.00	3.23	0.15	0.00
Journal 5	0.39	0.03	0.00	-0.26	0.09	0.00
Journal 6	-0.11	0.08	0.17	3.53	0.17	0.00
Journal 7	0.41	0.06	0.00	0.32	0.45	0.48
Years published	0.19	0.00	0.00	0.39	0.01	0.00
Constant	2.00	0.06	0.00	-3.49	0.20	0.00
Log(θ)	0.51	0.02	0.00			

number of citations, according to $\ln(\textit{number of citations} + 1)$. A Kolmogorov-Smirnov test of normality reported strong evidence against the log-linearity of the distribution of the number of citations ($D = 0.62$, $p = 0.00$).

Figure A.2 shows the number of published articles with zero citations (25.30%) compared to the number of articles which were cited at least once.

Tables A.1 and A.2 show the estimates of Zero-Inflated Negative Binomial (ZINB) regression models (Hilbe, 2014) in which the number of article citations was considered as a function of the same set of regressors reported in Tables 2 and 3, respectively. ZINB regressions consider the binary event of scoring 0 (zero-inflation model) separately from the count scores of an outcome (count model). The reported models show that both *text* and *reference changes* implied a small negative effect on the probability of receiving 0 citations against those of receiving at least one. With regard to the count models, we did not find any evidence of an effect of *text changes*, while we found a null effect of *reference changes*.

Furthermore, we modelled the number of citations as a 4-level ordinal variable based on quartiles. Tables A.3 and A.4 show results from ordinal logistic regression models (McCullagh, 1980) as a function of the same set of regressors reported in Tables 2 and 3, respectively. In both models, we found a small effect of *text* and *reference changes*.

Table A.2
Zero-Inflation Negative Binomial Regression model of the number of citations as a function of *reference changes* and the same covariates as in Table 3.

	Count model			Zero-inflation model		
	$\hat{\beta}$	S.E.	Pr(> |z|)	$\hat{\beta}$	S.E.	Pr(> |z|)
Reference changes	0.00	0.00	0.00	-0.01	0.00	0.00
Review score	0.00	0.00	0.00	0.00	0.00	0.44
Journal 2	1.11	0.05	0.5	1.11	0.14	0.00
Journal 3	-0.05	0.05	0.29	0.67	0.18	0.00
Journal 4	0.25	0.09	0.01	3.17	0.18	0.00
Journal 5	0.45	0.03	0.00	-0.35	0.09	0.00
Journal 6	0.00	0.10	0.98	3.39	0.20	0.00
Journal 7	0.44	0.08	0.00	0.27	0.44	0.53
Years published	0.19	0.00	0.00	0.37	0.01	0.00
Constant	2.00	0.06	0.00	-3.49	0.20	0.00
Log(θ)	0.54	0.02	0.00			

Table A.3
Ordinal logistic regression model of quartiles of number of citations as a function of *text changes* and the same covariates as in Table 3.

	$\hat{\beta}$	S.E.	Pr(> |t|)
Text changes	0.01	0.00	0.00
Review score	0.00	0.00	0.00
Journal 2	-0.59	0.08	0.00
Journal 3	-0.55	0.07	0.00
Journal 4	-2.36	0.10	0.00
Journal 5	0.62	0.06	0.00
Journal 6	-1.67	0.13	0.00
Journal 7	0.08	0.13	0.52
Years published	-0.04	0.01	0.00
1—2	-1.08	0.12	0.00
2—3	0.25	0.12	0.03
3—4	1.46	0.12)	0.00

Table A.4
Ordinal logistic regression model of quartiles of number of citations as a function of *reference changes* and the same covariates as in Table 3.

	$\hat{\beta}$	S.E.	Pr(> |t|)
Reference changes	0.01	0.00	0.00
Review score	0.00	0.00	0.00
Journal 2	-0.69	0.10	0.00
Journal 3	-0.39	0.08	0.00
Journal 4	-2.56	0.13	0.00
Journal 5	0.63	0.07	0.00
Journal 6	-1.56	0.16)	0.00
Journal 7	0.14	0.15	0.33
Years published	-0.04	0.01	0.00
1—2	-0.81	0.12	0.00
2—3	0.33	0.12	0.00
3—4	1.46	0.12	0.00

References

Atjonen, P. (2019). Peer review in the development of academic articles: Experiences of finnish authors in the educational sciences. *Learned Publishing, 32*(2), 137–146. 10.1002/leap.1204.

Baldwin, M. (2018). Scientific autonomy, public accountability, and the rise of "peer review" in the cold war united states. *Isis; An International Review Devoted to the History of Science and its Cultural Influences, 109*(3), 538–558. 10.1086/700070.

Bedeian, A. G. (2004). Peer review and the social construction of knowledge in the management discipline. *Academy of Management Learning & Education, 3*(2), 198–216. 10.5465/amle.2004.13500489.

Bianchi, F., Grimaldo, F., Bravo, G., & Squazzoni, F. (2018). The peer review game: An agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics, 116*, 1401–1420. 10.1007/s11192-018-2825-4.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*(1), 197–245. 10.1002/aris.2011.1440450112.

Bravo, G., Farjam, M., Grimaldo, F., Birukou, A., & Squazzoni, F. (2018). Hidden connections: Network effects on editorial decisions in four computer science journals. *Journal of Informetrics, 12*(1), 101–112. 10.1016/j.joi.2017.12.002.

Buljan, I., Garcia-Costa, D., Grimaldo, F., Squazzoni, F., & Marušić, A. (2020). Meta-research: Large-scale language analysis of peer review reports. *eLife, 9*, e53249. 10.7554/eLife.53249.

Casnici, N., Grimaldo, F., Gilbert, N., & Squazzoni, F. (2017). Attitudes of referees in a multidisciplinary journal: An empirical analysis. *Journal of the Association for Information Science and Technology, 68*(7), 1763–1771. 10.1002/asi.23665.

Coupé, T. (2013). Peer review versus citations – an analysis of best paper prizes. *Research Policy, 42*(1), 295–301. 10.1016/j.respol.2012.05.004.

Cowley, S. J. (2015). How peer-review constrains cognition: On the frontline in the knowledge sector. *Frontiers in Psychology, 6*, 1706. 10.3389/fpsyg.2015.01706.

Dondio, P., Casnici, N., Grimaldo, F., Gilbert, N., & Squazzoni, F. (2019). The "invisible hand" of peer review: The implications of author-referee networks on peer review in a scholarly journal. *Journal of Informetrics, 13*(2), 708–716. 10.1016/j.joi.2019.03.018.

Edwards, M. A., & Siddhartha, R. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science, 34*(1), 51–61. 10.1089/ees.2016.0223.

Eve, M. P., Neylon, C., O'Donnell, D. P., Moore, S., Gadie, R., Odeniyi, V., & S, P. (2021). *Reading peer review. PLOS ONE and institutional change in academia*. Cambridge University Press.

Eyre-Walker, A., & Stoletzki, N. (2013). The assessment of science: The relative merits of post-publication review, the impact factor, and the number of citations. *PLOS Biology, 11*(10), 1–8. 10.1371/journal.pbio.1001675.

Fyfe, A., McDougall-Waters, J., & Moxham, N. (2015). 350 Years of scientific periodicals. *Notes and Records: The Royal Society Journal of the History of Science, 69*(3), 227–239. 10.1098/rsnr.2015.0036.

Fyfe, A., Squazzoni, F., Torny, D., & Dondio, P. (2020). Managing the growth of peer review at the royal society journals, 1865–1965. *Science, Technology, & Human Values, 45*(3), 405–429. 10.1177/0162243919862868.

Garcia-Costa, D., Squazzoni, F., Mehmani, B., & Grimaldo, F. (2022). Measuring the developmental function of peer review: A multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. *PeerJ, 10*(e13539). 10.7717/peerj.13539.

Gaudino, M., Robinson, N. B., Franco, A. D., Hameed, I., Naik, A., Demetres, M., Girardi, L. N., Frati, G., Fremes, S. E., & Biondi-Zoccai, G. (2021). Effects of experimental interventions to improve the biomedical peer-review process: A systematic review and meta-analysis. *Journal of the American Heart Association, 10*(15), e019903. 10.1161/JAHA.120.019903.

Grimaldo, F., Marušić, A., & Squazzoni, F. (2018). Fragments of peer review: A quantitative analysis of the literature (1969–2015). *PLOS ONE, 13*(2), 1–14. 10.1371/journal.pone.0193148.

Harms, P. D., & Credé, M. (2020). Bringing the review process into the 21st century: Post-publication peer review. *Industrial and Organizational Psychology, 13*(1), 51–53. 10.1017/iop.2020.13.

Hilbe, J. M. (2014). *Modeling count data*. New York, NY: Cambridge University Press.

Horbach, S., & Halffman, W. (2018). The changing forms and expectations of peer review. *Research Integrity and Peer Review, 3*(8). 10.1186/s41073-018-0051-5.

Horbach, S. P. J. M. (2021). No time for that now! Qualitative changes in manuscript peer review during the Covid-19 pandemic. *Research Evaluation, 30*(3), 231–239. 10.1093/reseval/rvaa037.

Huisman, J., & Smits, J. (2017). Duration and quality of the peer review process: The author's perspective. *Scientometrics, 113*, 633–650. 10.1007/s11192-017-2310-5.

Khan, D., Arjmandi, M. K., & Yuvaraj, M. (2021). Most cited works on cloud computing: The 'citation classics' as viewed through dimensions.ai. *Science & Technology Libraries, 0*(0), 1–14. 10.1080/0194262X.2021.1951424.

Kharasch, E. D., Avram, M. J., Clark, J. D., Davidson, A. J., Houle, T. T., Levy, J. H., London, M. J., Sessler, D. I., & Vutskits, L. (2021). Peer review matters: Research quality and the public trust. *Anesthesiology, 134*(1), 1–6. 10.1097/ALN.0000000000003608.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*(8), 707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965

Liu, M., Hu, X., Wang, Y., & Shi, D. (2018). Survive or perish: Investigating the life cycle of academic journals from 1950 to 2013 using survival analysis methods. *Journal of Informetrics, 12*(1), 344–364. 10.1016/j.joi.2018.02.001.

Lugosi, P. (2021). The value creation cycle of peer review. *Annals of Tourism Research, 86*, 103092. 10.1016/j.annals.2020.103092.

Matsui, A., Chen, E., Wang, Y., & Ferrara, E. (2021). The impact of peer review on the contribution potential of scientific papers. *PeerJ, 9*, e11999. 10.7717/peerj.11999.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B, 42*(2), 109–142.

Merrill, E. (2014). Reviewer overload and what can we do about it. *The Journal of Wildlife Management, 78*(6), 961–962. 10.1002/jwmg.763.

Merriman, B. (2020). Peer review as an evolving response to organizational constraint: Evidence from sociology journals, 1952–2018. *The American Sociologist, 52*, 341—-366. 10.1007/s12108-020-09473-x.

Moxham, N., & Fyfe, A. (2018). The Royal Society and the prehistory of peer review, 1665–1965. *The Historical Journal, 61*(4), 863–889. 10.1017/S0018246X17000334.

Ortega, J. (2017). Are peer-review activities related to reviewer bibliometric performance? A scientometric analysis of Publons. *Scientometrics, 112*, 947–962. 10.1007/s11192-017-2399-6.

Pontille, D., & Torny, D. (2015). From manuscript evaluation to article valuation: The changing technologies of journal peer review. *Human Studies, 38*, 57–79. 10.1007/s10746-014-9335-z.

Publons (2018). 2018 Global state of peer review. Clarivate Analytics.

Rigby, J., Cox, D., & Julian, K. (2018). Journal peer review: A bar or bridge? An analysis of a paper's revision history and turnaround time, and the effect on citation. *Scientometrics, 114*(3), 1087–1105. 10.1007/s11192-017-2630-5.

Seeber, M. (2020). How do journals of different rank instruct peer reviewers? Reviewer guidelines in the field of management. *Scientometrics, 122*, 1387–1405. 10.1007/s11192-019-03343-1.

Severin, A., Strinzel, M., Egger, M., Domingo, M., & Barros, T. (2021). Characteristics of scholars who review for predatory and legitimate journals: Linkage study of cabells scholarly analytics and Publons data. *BMJ Open, 11*(7). 10.1136/bmjopen-2021-050270.

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences, 112*(2), 360–365. 10.1073/pnas.1418218112.

Song, E., Ang, L., Park, J.-Y., Jun, E.-Y., Kim, K. H., Jun, J., Park, S., & Lee, M. S. (2021). A scoping review on biomedical journal peer review guides for reviewers. *PLOS ONE, 16*(5), 1–18. 10.1371/journal.pone.0251440.

Squazzoni, F., Ahrweiler, P., Barros, T., Bianchi, F., Birukou, A., Blom, H. J. J., Bravo, G., Cowley, S., Dignum, V., Dondio, P., Grimaldo, F., Haire, L., Hoyt, J., Hurst, P., Lammey, R., MacCallum, C., Marušić, A., Mehmani, B., Murray, H., Nicholas, D., Pedrazzi, G., Puebla, I., Rodgers, P., Ross-Hellauer, T., Seeber, M., Shankar, K., Van Rossum, J., & Willis, M. (2020). Unlock ways to share data on peer review. *Nature, 578*, 512–514. 10.1038/d41586-020-00500-yz.

Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., Birukou, A., Dondio, P., & Grimaldo, F. (2021). Peer review and gender bias: A study on 145 scholarly journals. *Science Advances, 7*(2), eabd0299. 10.1126/sciadv.abd0299.

Squazzoni, F., Grimaldo, F., & Marusic, A. (2017). Publishing: Journals could share peer-review data. *Nature, 546*(352). 10.1038/546352a.

Strang, D., & Siler, K. (2015). Revising as reframing: Original submissions versus published papers in administrative science quarterly, 2005 to 2009. *Sociological Theory, 33*(1), 71–96. 10.1177/0735275115572152.

Taşkın, Z., Doğan, G., Kulczycki, E., & Zuccala, A. A. (2021). Self-citation patterns of journals indexed in the journal citation reports. *Journal of Informetrics, 15*(4), 101221. 10.1016/j.joi.2021.101221.

Tennant, J., Dugan, J., Graziotin, D., Jacques, D. C., Waldner, F., Mietchen, D., Elkhatib, Y., B. Collister, L., Pikas, C., Crick, T., Masuzzo, P., Caravaggi, A., Berg, D., Niemeyer, K., Ross-Hellauer, T., Mannheimer, S., Rigling, L., Katz, D., Greshake Tzovaras, B., Pacheco-Mendoza, J., Fatima, N., Poblet, M., Isaakidis, M., Irawan, D., Renaut, S., Madan, C., Matthias, L., Nørgaard Kjær, J., O'Donnell, D., Neylon, C., Kearns, S., Selvaraju, M., & Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review [version 3; referees: 2 approved]. *F1000 Research, 6*(1151). 10.12688/f1000research.12037.3.

Tennant, J., & Ross-Hellauer, T. (2020). The limitations to our understanding of peer review. *Research Integrity & Peer Review, 6*. 10.1186/s41073-020-00092-1.

Teplitskiy, M. (2016). Frame search and re-search: How quantitative sociological articles change during peer review. *The American Sociologist, 47*, 264–288. 10.1177/0162243919862868.

Thelwall, M., Allen, L., Papas, E.-R., Nyakoojo, Z., & Weigert, V. (2021). Does the use of open, non-anonymous peer review in scholarly publishing introduce bias? Evidence from the f1000research post-publication open peer review publishing model. *Journal of Information Science, 47*(6), 809–820. 10.1177/0165551520938678.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences, 114*(48), 12708–12713. 10.1073/pnas.1707323114.

Apéndice C

Does peer review improve the statistical content of manuscripts? A study on 27,467 submissions to four journals

ROYAL SOCIETY
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Garcia-Costa D, Forte A, Lòpez-Iñesta E, Squazzoni F, Grimaldo F. 2022 Does peer review improve the statistical content of manuscripts? A study on 27 467 submissions to four journals. *R. Soc. Open Sci.* **9**: 210681. https://doi.org/10.1098/rsos.210681

Received: 6 May 2021
Accepted: 23 August 2022

Subject Category:
Computer science and artificial intelligence

Subject Areas:
artificial intelligence/statistics/psychology

Keywords:
peer review, manuscripts, reviewers, statistics, academic journals

Author for correspondence:
Francisco Grimaldo
e-mail: francisco.grimaldo@uv.es

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6174474.

THE ROYAL SOCIETY
PUBLISHING

Does peer review improve the statistical content of manuscripts? A study on 27 467 submissions to four journals

Daniel Garcia-Costa[1], Anabel Forte[2], Emilia Lòpez-Iñesta[3], Flaminio Squazzoni[4] and Francisco Grimaldo[1]

[1]Department of Computer Science, and [2]Department of Statistics and Operational Research, University of Valencia, Burjassot, Spain
[3]Department of Mathematics Education, University of Valencia, Valencia, Spain
[4]Department of Social and Political Sciences, University of Milan, Milan, Italy

FS, 0000-0002-6503-6077

Improving the methodological rigour and the quality of data analysis in manuscripts submitted to journals is key to ensure the validity of scientific claims. However, there is scant knowledge of how manuscripts change throughout the review process in academic journals. Here, we examined 27 467 manuscripts submitted to four journals from the Royal Society (2006–2017) and analysed the effect of peer review on the amount of statistical content of manuscripts, i.e. one of the most important aspects to assess the methodological rigour of manuscripts. We found that manuscripts with both initial low or high levels of statistical content increased their statistical content during peer review. The availability of guidelines on statistics in the review forms of journals was associated with an initial similarity of statistical content of manuscripts but did not have any relevant implications on manuscript change during peer review. We found that when reports were more concentrated on statistical content, there was a higher probability that these manuscripts were eventually rejected by editors.

1. Introduction

Peer review is key for public trust in the scientific community [1]. By exposing manuscripts to scrutiny by independent experts, it ensures that scientific claims are grounded on reliable evidence. This requires reviewers to screen the rigour and quality of methods and analysis reported in manuscripts submitted to journals for publication. Although reviewers are expected to check various aspects of a manuscript, this attention to rigour and methodology includes one of the most important imperatives of science as an institutional system—what the famous sociologist of science Robert K. Merton called 'organized skepticism' [2]. While the purposes and practices of peer review have varied considerably with time, place and discipline [3,4], collaboration between unrelated experts in improving the rigour and reliability of scientific findings is of paramount importance especially in the current climate of academic hyper-competition, where scientists are exposed to perverse incentives that maximize the 'publishability' of research rather than its methodological rigour [5–7].

While author–reviewer collaboration during peer review can have different forms, some of which are potentially dysfunctional, e.g. collusion and parochialism [8,9], one of the most important functions of reviewers is to ensure that journals achieve the highest methodological rigour and statistical standards by improving manuscripts. On the one hand, this developmental function of peer review is pivotal in helping authors improve their manuscripts throughout the process [10]. On the other, it enhances the legitimacy and credibility of journals as gatekeepers of scholarly communication [11,12].

Unfortunately, there is little understanding of how this developmental function actually works [13–16]. While research on specific journals has shown that exposure to different rounds of peer review could increase the quality of manuscripts—including later submissions to other journals if rejected [17], other studies have suggested that reviewers are keen to preferably concentrate on theoretical aspects rather than rigour, methodology and statistical content [11,18]. While reviewers are expected to comment on various aspects as well as assisting editors in judging about the suitability of work for publication, exclusively considering background theory, novelty and implications could be detrimental for peer review quality, as reported in the current debate on the quality of peer review during the COVID-19 pandemic [19].

To ensure that reviewers do not only consider novelty as opposed to rigour, journals have introduced guidelines and instructions to ensure they focus on data analysis and statistical testing [14,20]. These often include instructions on how reviewers should provide valid assessments of methods and statistics reported in articles, including measurement validity, outcome sensitivity and findings replicability [21]. While assessing the effective use of these instructions is difficult [22,23], measuring the effect of peer review on how manuscripts change from initial submission to the published version is even more challenging given the system's confidentiality and lack of data on internal editorial processes [24].

To fill this gap, we established a confidential agreement with the Royal Society to access manuscript and peer review data from their journals. The world's oldest independent scientific academy, with the first publication of *Philosophical Transactions* in 1665, the Royal Society pioneered the concepts and practices of academic journals, editorial responsibility and peer review [25]. The Royal Society journals include prestigious titles, such as *Philosophical Transactions A* and *Proceedings A*, which publish research on physical, mathematical and engineering sciences, *Philosophical Transactions B*, *Proceedings B* and *Biology Letters*, with a readership in biological sciences, as well as cross-disciplinary outlets, such as *Interface*, for cross-disciplinary research at the interface between the physical and life sciences, and *Royal Society Open Science*, the Royal Society's most recent open access journal in science, engineering and mathematics.

Data included complete manuscript files and (when available) peer review reports over the same time frame (2006–2017) from all these journals. However, after careful analysis of the database, we restricted our sample to four journals to ensure full comparability of manuscripts (see detail in the Methods section). We concentrated on 27 467 manuscripts from four journals and built a glossary of statistical terms to analyse the text of manuscripts and review reports. Note that in compliance with the agreement signed by all authors of this study, journals were fully anonymized to avoid identification. While other research has examined review reports, e.g. studying their linguistic properties [26–28], our rich and original dataset allowed us to link manuscripts and reports, thus providing a more comprehensive, contextual picture of the collaboration between authors and reviewers in improving manuscripts. Our aim here was to measure the change of the statistical content of manuscripts during peer review, i.e. one of the most relevant functions of reviewers (at least in hard sciences), to estimate conditions and contexts that could stimulate collaborative improvement of manuscripts between authors and reviewers. We first measured the statistical content of manuscripts by scanning their text with a Linguistic Inquiry and Word Count style dictionary built upon a well-known statistics glossary. We assumed that the number of statistical terms included in

Table 1. Data overview.

journal ID	J1	J7	J8	J11	all
guidelines for statistics	yes	yes	yes	no	—
peer-reviewed manuscripts	7742	350	2420	731	*11 243*
rejection rate	59.2%	47.1%	57.9%	49.8%	*58.0%*
median number of rounds	1	2	2	2	*1*
mean number of statistical terms	12.65	12.41	7.95	11.14	*11.53*
desk-rejection or acceptance	8627	957	2481	1551	*13 616*
mean number of statistical terms	11.80	7.61	7.40	10.11	*10.51*
manuscripts with no review report	963	429	626	590	*2608*
rejection rate	25.5%	15.9%	17.4%	14.2%	*19.4%*
median number of rounds	2	2	3	3	*3*
mean number of statistical terms	12.79	11.28	8.33	10.51	*10.95*
number of research manuscripts	*17 332*	*1736*	*5527*	*2872*	*27 467*

the text was a proxy of their statistical content. We then measured the statistical content of manuscripts from their initial submissions to their revisions by comparing different versions of the same manuscripts. We also similarly measured the statistical content of review reports. By controlling for important factors, such as the reviewer score received by manuscripts, the number of rounds of peer review and the number of reviewers commenting on the same manuscript, we tried to estimate the effect of peer review on manuscript change and examine the most relevant peer review-related factors shaping the final editorial decision.

Note that we did not assume that any change of the statistical content of manuscripts during peer review would always lead to manuscript improvements in terms of methodological rigour. We also did not assume that any change of statistical terms in the text would necessarily mean the improvement of the quality and rigour of manuscripts. Here, we assumed that the change of statistical content of manuscripts throughout the peer review process as proxied by text revisions may reveal a joint attention effort by reviewers and authors on the methodological content of manuscripts, which is one of the most important functions of peer review. As suggested by recent research, exploring the text of manuscript and peer review reports quantitatively is key to understand the scholarly communication landscape and reconstruct the complex, indirect, collaborative relationship between authors and reviewers, which typically occurs behind the confidentiality of the journal editorial process [29].

2. Methods

2.1. Data

Data were obtained thanks to a confidential agreement with the Royal Society and were extracted in a comparable time-frame (2006–2017). The original dataset included 60 240 manuscripts submitted to 13 journals. However, in order to ensure full comparability, we concentrated on four journals, which ensured similar standards in terms of number, type of submissions and rejection rates. We also excluded from the sample any manuscript without a clear submission date, being reviewed by multiple journals, changing its status during re-submission, being assigned an unclear final decision in the manuscript submission system (e.g. rejected after accepted or accepted twice), or with missing files. This implied removing more than 24 000 manuscripts from the sample. The remaining 34 781 manuscripts (see table S1 in the electronic supplementary material) were further filtered by selecting all research articles and excluding review papers, opinion pieces, reports, memoirs, recollections, replies etc. We restricted our analysis to journals J1, J7, J8 and J11, since these journals contributed to 97.1% of peer-reviewed manuscripts in our dataset (see electronic supplementary material, table S2). We excluded

Table 2. Selected statistical terms for each category.

royalsocietypublishing.org/journal/rsos *R. Soc. Open Sci.* **9**: 210681

category	list of terms
descriptive	binomial distribution, box plot, density, geometric distribution, histogram, negative-binomial distribution, normal distribution, outlier, percentile, Poisson distribution, quantile, quartile
contrast	alternative hypothesis, anova, chi-square, control group, Fisher, multiplicity, null hypothesis, odds, *p*-value, power, rejection region, significant, size effect, *t*-test, *z*-score, *z*-test
estimation	average, bias, confidence interval, correlation, estimate, estimation, estimator, expectation, expected value, probability, standard deviation, standard error
modelization	area under the curve, association, causality, confounding, cross-sectional study, extrapolation, interaction, interpolation, Kaplan Meier, longitudinal, model, regression
generics	Bayes, boostrap, central limit theorem, confidence level, independence, kernel, law of large numbers, likelihood, parameters, population, random, sample, variable

data from the rest of the journals since they marginally contribute with less than 1% of peer-reviewed articles.

This led us to consider 27 467 manuscripts (table 1), including:

— 11 243 manuscripts that were peer-reviewed,
— 13 616 manuscripts that were desk-rejected or accepted without any round of peer review, and
— 2608 manuscripts without any available review report (i.e. missing or not recorded in the journal submission system).

We then checked whether journals included any guidelines for assessing statistics in their forms sent to reviewers, i.e. an explicit question asking reviewers to assess the quality of a manuscript's statistical analysis in the review form.

To map the statistical content of manuscripts, we selected a list of commonly used statistical terms from a statistics glossary developed by the University of Berkeley (https://www.stat.berkeley.edu/~stark/SticiGui/Text/gloss.htm). Table 2 shows our selected list of terms, which were then aggregated into five categories for the sake of simplicity. We checked for between-terms orthogonality over the full list of terms, thereby ensuring that each term represented different, not overlapping concepts. Electronic supplementary material, figure S10, shows that between-terms mutual overlapping was rare, except for the term 'model', which has multiple meanings and so was kept in the dictionary.

We applied our dictionary to map the presence of these concepts in the text of manuscripts and review reports by using an R library called `quanteda.dictionaries`. Our study considered all categories together since our main focus was the whole statistical content, regardless of the changing nature of statistical concepts within manuscripts (either descriptive, inferential or both).

We considered the presence of statistical terms within the text of manuscripts and excluded equations, tables and figures, while keeping their captions and recurrences in the text. This allowed us to consider also equations, tables or figures while achieving full comparison of manuscripts and journals and minimizing bias due to either journal- or manuscript-specific features (e.g. different file format, such as PDF, LaTeX, Word, RTF).

2.2. Statistical models

To explore the potential effect of peer review on the statistical content of manuscripts and on editorial decisions, we built two models: a Poisson regression for the number of different statistical terms in the final version of each manuscript which underwent revisions during peer review, and a logistic linear regression for the probability of editorial acceptance of manuscripts after peer review.

We applied a Bayesian variable selection to identify the variables to be included in these linear predictors. To do so, we considered posterior probabilities for each possible combination of variables shown in electronic supplementary material, tables S3 and S4. More specifically, we considered 2^p models, p being the potential co-variates in each linear predictor (6 and 7, respectively). We then calculated the posterior inclusion probability (PIP) for each variable as the sum of the posterior probabilities in all models.

This required us to specify the prior distributions involved in the Bayes theorem, which were priors for each model and their parameters, and calculate 2^p posterior probabilities which usually need numerical integration (e.g. [30]). Due to the generalized linear nature of our models, we followed [31] and their numerical approximation to the solution, which was implemented in the R package BAS (Bayesian model averaging using Bayesian adaptive sampling) [32].

Tables S3 and S4 in the electronic supplementary material show the results of our model implementations. We selected variables with PIPs greater than 0.5. For the first model, the statistical content of the final version of manuscripts was mainly associated with: the statistical content of the review reports received by manuscripts (max_stats_rev), the level of statistical content in the initial version of manuscripts (initial_stats) and the total number of review rounds undergone by manuscripts (*nrounds*). For the second model, the probability of a manuscript's acceptance was associated with: the statistical content of the associated review reports (max_stats_rev), the number of rounds (nrounds), the number of reviewers (nreviewers) and the review score of manuscripts (score) as defined in [33].

After selecting our model variables, in order to estimate the final number of different statistical terms of manuscripts, we added a random effect per journal to reflect possible differences between journals (figure 1*c*). However, for the sake of clarity, we excluded this effect when examining the probability of each manuscript's acceptance as these probabilities were similar across journals.

Considering all aspects, the final model of the number of different statistical terms in the final version of manuscripts (i) was as follows:

$$y_i \sim \text{Poiss}(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 \text{max\_stats\_rev}_i + \beta_2\ \text{initial\_stats}_i + \beta_3\ \text{nrounds}_i + b_{\text{journal}_i}$$
$$b_j \sim N(0, \sigma) \quad \text{for } j = 1, 7, 8, 11.$$

The selected model for the probability of a manuscript's acceptance (i) was as follows:

$$\text{accept}_i \sim \text{Bernoulli}(\pi_i)$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1\ \text{max\_stats\_rev}_i + \beta_2\ \text{nrounds}_i + \beta_3\ \text{nreviewers}_i + \beta_4\ \text{score}_i.$$

Following the Bayesian paradigm, all model parameters were considered as random variables and assigned a prior distribution. For the regression coefficients β_j, we used a normal prior distribution at 0 and with large variance. For the standard deviation of the random effect associated with each journal, σ, we used a uniform distribution from 0 to 10.

These models were estimated using Bayesian inference through the software JAGS (Just another Gibbs Sampler) and its R interface `rjags` [34]. JAGS performs Markov chain Monte Carlo (MCMC) methods to simulate from desired posterior distributions. After a burning and a thinning MCMC process with one chain, we kept a total of 3000 samples of the posterior distribution of the model parameters.

3. Results

Figure 1 shows that initial submissions had a relatively homogeneous statistical content, except for manuscripts directly accepted by editors without any peer review (see the red solid line, which corresponded to 42 manuscripts). The availability of guidelines on statistics for reviewers did not have any qualitative effect on the variation of the initial statistical content of manuscripts submitted for publication (note that journals J1, J7, J8 included these questions in the review form, whereas journal J11 did not). However, we found certain differences between journals, which reflected their different academic audiences. For instance, initial submissions to J7 showed the greatest variability of statistical

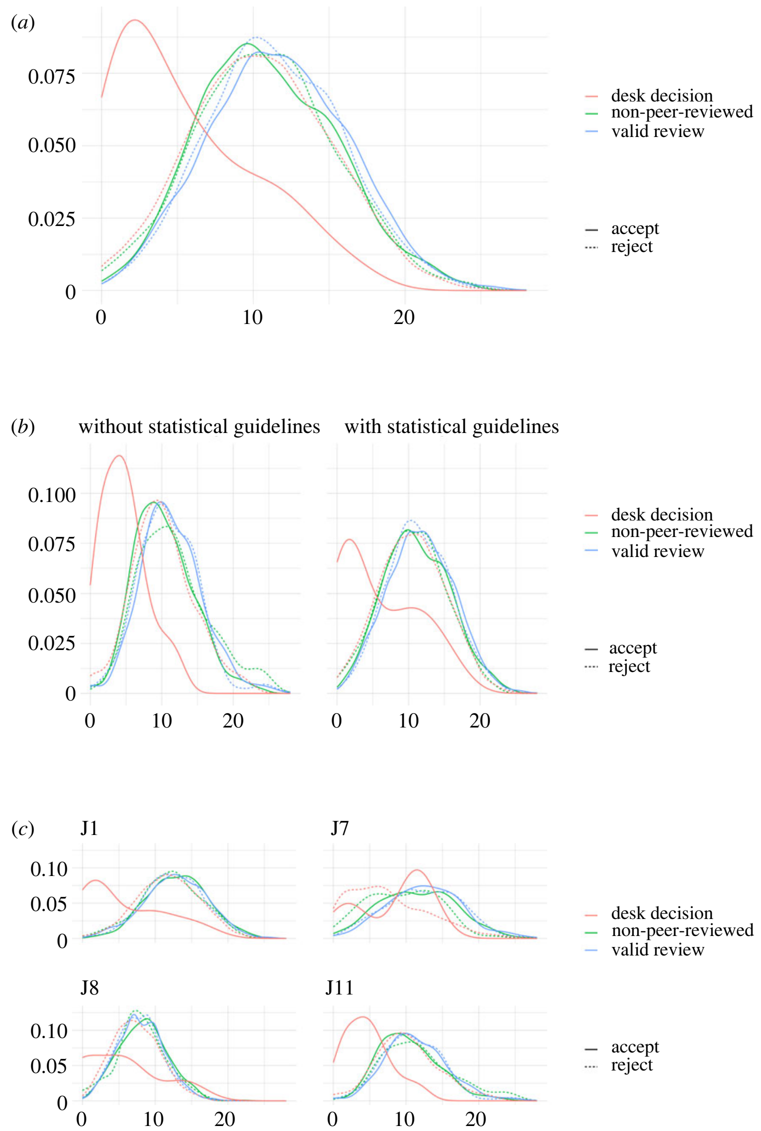

Figure 1. Number of different statistical terms (*x*-axis) in initial submissions for rejected (dotted line) or accepted (solid line) manuscripts, in cases of not peer-reviewed (green), desk rejected/accepted (red) and peer-reviewed (blue) manuscripts (*a*), per journal with or without guidelines for statistics (*b*) and per journal (*c*).

content among journals, whereas initial submissions to J8 showed the lowest level of statistical content in the manuscript sample.

We then considered all 11 243 manuscripts that survived the editorial desk and were eventually reviewed multiple times (note that 50.7% of these 11 243 manuscripts were rejected after the first round). We compared their initial statistical content with the final version of manuscripts after peer review. We found that 13.8% of these did not vary their statistical content (i.e. the number of different statistical terms in these manuscripts was the same). For the remaining 35.4%, 23.9% of these manuscripts increased their statistical content, whereas 11.6% reduced it. Regarding the final editorial decision, half of manuscripts accepted for publication increased their statistical content during peer review, 25% decreased it, whereas the remaining 25% did not vary. A proportion of 93.1% of manuscripts which were eventually rejected after peer review did not change in terms of statistical content, 5% increased it, whereas 1.9% decreased it (see figure S1 in the electronic supplementary material).

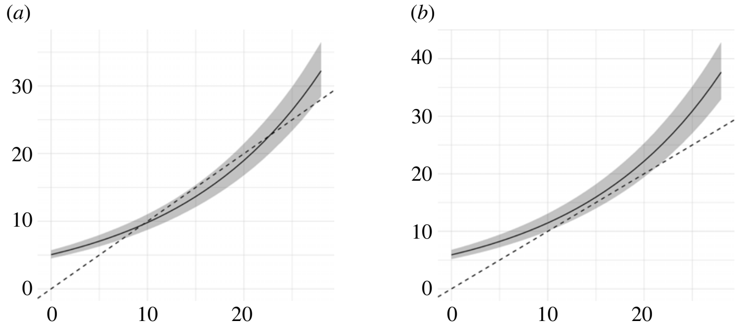

Figure 2. Initial (*x*-axis) versus final (*y*-axis) statistical content of manuscripts by moderate (five terms) statistical content of reports (*a*) or strong (25 terms) statistical content of reports (*b*).

We then considered other variables, which could affect the difference of statistical content during manuscript revisions, including:

— the availability of guidelines to assess the statistical content of manuscripts in the review form of some journals;
— the number of rounds of peer review undergone by manuscripts before the final editorial decision;
— the number of reviewers who jointly or sequentially assessed the same manuscript; and
— the reviewer score, i.e. the quality of manuscripts as assessed by reviewers.

We found that the availability of guidelines in the review form did not have any significant effect on the statistical content of manuscripts (see figure S2 in the electronic supplementary material). We found a positive effect of the peer review on the statistical content of manuscripts: more rounds implied more substantial changes (see figure S3 in the electronic supplementary material). We also found that being assessed by more than two reviewers led to an increase of manuscripts' statistical content (with a significant χ^2-test) for both accepted and rejected manuscripts (see figure S4 in the electronic supplementary material).

We then measured each reviewer's focus on statistics by analysing the statistical content of their comments to authors. Given that this required the availability of review text, we had to restrict our analysis to 11 050 manuscripts (out of 11 243). Results showed that reviewers varied their opinion on the statistical content of manuscripts (see figure S5 in the electronic supplementary material). We found a wide variability in the maximum number of different statistical terms in reviewer reports. Reports with less statistical content were associated with smaller changes in the statistical content of manuscripts (e.g. see the lowest median of statistical terms in review reports—*y*-axis—associated with the value 0 in changes in statistical content of manuscripts—*x*-axis—in figure S5 in the electronic supplementary material).

Following [27,33,35], we used the review score as a proxy of the quality of manuscripts, which is typically a robust predictor of editorial decisions (see detail on the review score in the Methods sections of the references cited above). As expected, editorial decisions on manuscripts depended greatly on review scores: manuscripts rejected after peer review had a lower and more variable review score, whereas manuscripts accepted for publication had higher review scores. Results showed that manuscripts eventually accepted for publication but receiving lowest review scores were also those increasing their statistical content the most during peer review (see figure S6 in the electronic supplementary material).

Results of our models showed that the statistical content of a manuscript's final version was related to the level of statistical content of its initial version submitted for publication, the statistical content of review reports and the number of peer review rounds (see table S3 and figure S7 in the electronic supplementary material, where we report posterior distributions of the exponential of the coefficients associated with each variable). Furthermore, when considering random effects at a journal level, results confirmed that manuscripts submitted to journal J8 generally had lower levels of statistical content (see figure S8 in the electronic supplementary material).

More importantly, we found that reviewers contributed to increase the statistical content of manuscripts regardless of the statistical content of review reports (figure 2). However, it is worth noting that manuscripts with moderate levels of initial statistical content (i.e. about 15 words compared to the maximum number of different statistical terms, which was 30 terms as shown in (figure 1) had fewer variations throughout the peer review process than those with either a small or large number of different statistical terms in their initial version. In short, manuscripts with initial low or high levels of statistical content were those which improved the most during peer review.

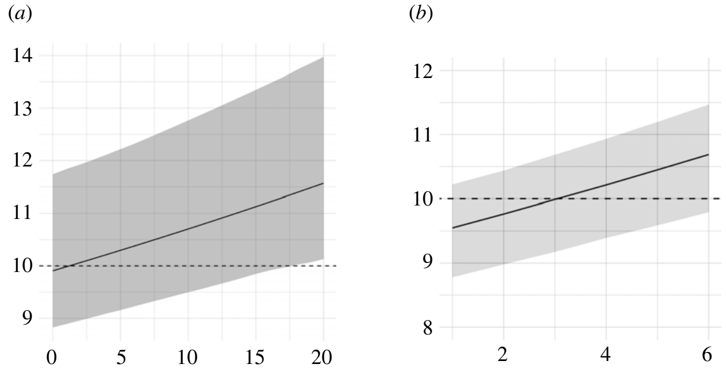

Figure 3. Number of different statistical terms in the final version of manuscripts (*y*-axis) as due to (*x*-axis) the maximum number of different statistical content in the report (*a*) and the number of rounds of peer review (*b*).

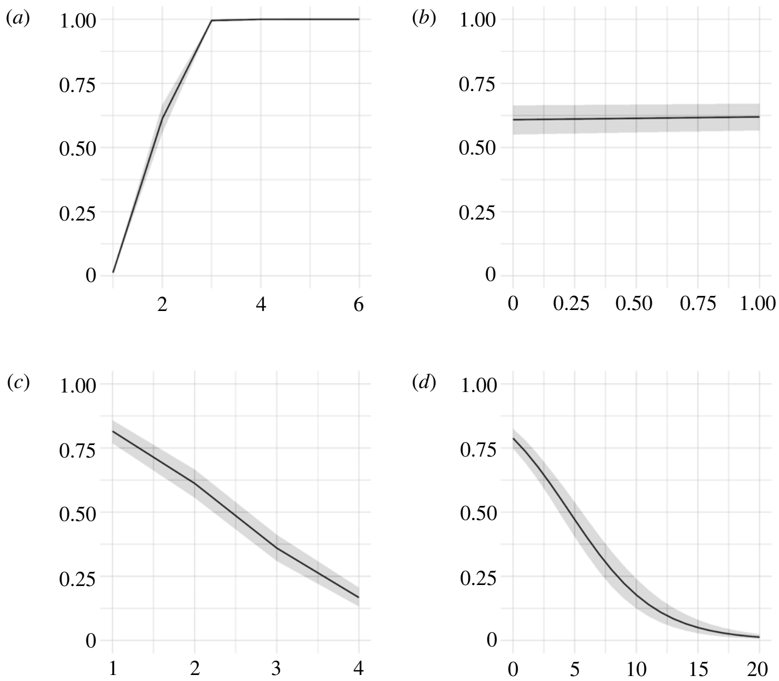

Figure 4. The probability of a manuscript's acceptance (*y*-axis) due to the number of peer review rounds (*a*), the review score for papers following two rounds of review (*b*), the number of reviewers (*c*) and the maximum number of different statistical terms in the review reports (*d*).

Figure 3 shows that the effect of the number of different statistical terms in the reports and the number of rounds of peer review on the final statistical content of manuscripts is increasing and linear. Though, changes were mostly marginal, e.g. adding one new term to the average increase of 10 different statistical terms (see dotted line) in the final version of the manuscript.

Electronic supplementary material, table S4, shows the results of our logistic regression model (see electronic supplementary material). The probability of a manuscript being accepted for publication was related to the number of reviewers who assessed it, the statistical content of review reports, the overall opinion of reviewers (i.e. the review score received by the manuscript in all rounds of peer review), and the number of rounds of reviews (see the posterior distributions of the exponential of the coefficients associated with each of the variables in figure S9 in the electronic supplementary material).

Figure 4 shows that a manuscript that underwent more than two review rounds was eventually accepted by the editor (figure 4*a*). The review score was increasingly instrumental for a manuscript's final acceptance as it is closely related to the number of rounds. For instance, manuscripts undergoing only one round of reviews had a median review score of 0.12 and those undergoing more than one round had a median review score of 0.67. Although the effect of the number of rounds was strongly associated with the review score (e.g. see the marginal effect for manuscripts undergoing two review rounds, figure 4*b*), we considered both variables to build a better model, as indicated by their posterior inclusion probabilities (see electronic supplementary material, table S4). We also found a decreasing effect of the number of reviewers (figure 4*c*) and the statistical content of reviews (figure 4*d*). This would suggest that the more the reviewers were concentrated on statistics in their reports, the less likely a manuscript was eventually accepted for publication by the editor.

4. Discussion

The role of peer review in improving the quality of scientific publications has been subject to increasing scrutiny in recent research [28], which has led especially to more examination of current practices and standards [27,36,37]. However, this type of research only rarely integrates full data on manuscripts during each stage of the editorial process and data on review reports, at the same time covering different journals [38,39]. Integrating data on manuscripts and reports is key to providing a context-specific picture of peer review and editorial processes, not to mention the possibility of assessing changes and revisions of manuscripts due to peer review [28,40]. Although difficult, pooling across-journal data is instrumental to examine the emergence of peer review practices that are shared in various communities [24,35].

Here, we aimed to fill this gap by examining manuscript changes and peer review reports in a sample of manuscripts submitted to four journals from the Royal Society in the same time frame (2006–2017). We concentrated on the statistical content of manuscripts as a proxy of the rigour of the analysis supporting scientific claims and findings in published manuscripts. While this can be irrelevant in certain areas of research, e.g. the humanities, robust quantitative methodologies and statistical tests are key to corroborate findings in 'hard sciences'. Furthermore, our database allowed us to consider various factors that could influence manuscript development, including the number of rounds of peer review undergone by manuscripts, the number of reviewers who jointly or sequentially assessed them, the reviewer score, reflecting a manuscript's perceived quality by reviewers, and the availability of guidelines in the reviewer form.

Our results suggest that manuscripts with both initial lowest or highest levels of statistical content increased their statistical content during the process, whereas desk-rejected manuscripts had comparatively fewer statistical terms in their text. We found that these developments were associated with a higher probability of a manuscript's acceptance. The availability of reviewer guidelines on statistics on review forms seems to ensure similar initial levels of statistical content among submitted manuscripts but did not have any qualitative implication on manuscript change during peer review. We found that editors were more likely to reject manuscripts when reviewers concentrated more on the statistical content of manuscripts in their reports.

Note that our developmental measurements of peer review here did not consider the possible developments of manuscripts rejected by these four journals but later submitted to and possibly published by other journals. Although authors can disregard advice from reviewers after rejection and rejections are costly to the system and are often a source of academic frustration [41], research on the fate of rejected manuscripts has found that manuscripts are often developed across journals via subsequent, multiple submissions [17,42]. Review reports are of a great benefit to authors' learning and a source of scientific improvement, especially when reviewers spot flaws in methodology and lack of rigour in analysis, i.e. amendable weaknesses [43].

This said, our study has certain limitations. First, in order to analyse the text of manuscripts and review reports, we started from a glossary of statistical terms, selected those relevant to our purposes and measured the occurrence of these terms throughout manuscripts and reports. In our opinion, this was an appropriate design strategy considering the type of journals and areas of research in our dataset and the fact that statistics is a standardized field. However, integrating our measurements with qualitative analysis of the text by human experts would be a significant step forward [40]. This would also help to assess the potentially negative effect of reviewer requests on manuscript change as well as inform us about the link between increased statistical content and methodological quality and rigour of reported studies. Furthermore, applying supervised machine learning techniques could also be helpful to test alternative measurements. Unfortunately, yet large, and complete, our dataset was not sufficiently large to use supervised machine learning techniques, e.g., neural networks, which require large-scale, training datasets.

Secondly, although the four journals from the Royal Society covered here allowed us a certain degree of variety in terms of fields and journals, extending our research to other fields where statistics and statistical models are important, such as medicine, engineering, economics and social sciences, could help provide a more comprehensive picture of the developmental function of peer review in terms of rigour and methodology. This would also increase the in-depth definition of rigour: in certain areas, it is expected that the concept of rigour could extend to hypothesis testing and data collection, thereby suggesting that looking at statistical terms is only an approximation.

Finally, note that this type of research on language and content analysis of manuscripts and reports is in its infancy [26,28,44–46]. This implies that any measurement is only explorative and caution must be used when drawing any conclusions from a study's findings. On the one hand, even research on manuscript change in preprint–publication pairs estimates the potential effects of peer review only indirectly as the link between manuscripts and reports is missing [40,47,48]. On the other hand, research on the content of peer review reports from available report repositories, e.g. Publons, cannot help to estimate the effect of reports on manuscript change due to lack contextual information on associated manuscripts [49]. To improve this type of research, removing obstacles against data sharing from publishers to the community and increasing interdisciplinary, multi-approach studies combining qualitative and quantitative research is needed [24]. Not only would this help us assess the developmental role of peer review more systematically, but also this type of research could inform guidelines and arrangements to improve the fairness of peer review [28,29] and improve our understanding of the multiple functions and dimensions of this complex social institution called peer review [4,50].

Data accessibility. The dataset used for this study and the code for replication are available at https://doi.org/10.7910/DVN/MOKJED.

Supplementary material is available online [51].

Authors' contributions. D.G.-C: conceptualization, data curation, formal analysis, methodology, writing—original draft, writing—review and editing; E.L.-I.: conceptualization, formal analysis, writing—original draft, writing—review and editing; A.F.: conceptualization, formal analysis, writing—original draft, writing—review and editing; F.S.: conceptualization, methodology, supervision, writing—original draft, writing—review and editing; F.G.: conceptualization, data curation, project administration, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. The authors declare no competing interests.

Funding. This work was partially supported by the Spanish Ministry of Science and Innovation (MCINN), the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF) under projects RTI2018-095820-B-I00 and PID2019-104790GB-I00. F.S. was supported by a grant of MIUR-Italian Minister for Education, University and Research (20178TRM3F_002) and a grant from the University of Milan (PSR2015-17 transition grant). Funders had no role in the design of this study.

Acknowledgements. We gratefully acknowledge Phil Hurst and the team of the Royal Society for providing data and covering the cost of their extraction from manuscript submission systems.

References

1. Kharasch ED, Avram MJ, Clark JD, Davidson AJ, Houle TT. 2021 Peer review matters: research quality and the public trust. *Anesthesiology* **134**, 1–6. (doi:10.1097/ALN.0000000000003608)
2. Merton R. 1973 [1942] The normative structure of science. In *The sociology of science: theoretical and empirical investigations* (ed. R Merton), pp. 267–278. Chicago, IL: University of Chicago Press.
3. Moxham N, Fyfe A. 2018 The Royal Society and the prehistory of peer review, 1665–1965. *Hist. J.* **61**, 863–889. (doi:10.1017/S0018246 17000334)
4. Fyfe A, Squazzoni F, Torny D, Dondio P. 2020 Managing the growth of peer review at the Royal Society journals, 1865–1965. *Sci. Technol. Hum. Values* **45**, 405–429. (doi:10.1177/0162243919862868)
5. Flaherty M. 2016 Sociology as a conversation: the present circumstances and future prospects of peer review. *Am. Sociol.* **47**, 253–263. (doi:10.1007/s12108-015-9299-0)
6. Edwards MA, Siddhartha R. 2017 Academic research in the 21st century: maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environ. Eng. Sci.* **34**, 51–61. (doi:10.1089/ees.2016.0223)
7. Kiai A. 2019 To protect credibility in science, banish 'publish or perish'. *Nat. Hum. Behav.* **3**, 1017–1018. (doi:10.1038/s41562-019-0741-0)
8. Rigby J, Cox D, Julian K. 2018 Journal peer review: a bar or bridge? An analysis of a paper's revision history and turnaround time, and the effect on citation. *Scientometrics* **114**, 533–546. (doi:10.1007/s11192-017-2630-5)
9. Dondio P, Casnici N, Grimaldo F, Gilbert N, Squazzoni F. 2019 The 'invisible hand' of peer review: the implications of author-referee networks on peer review in a scholarly journal. *J. Informetrics* **13**, 708–716. (doi:10.1016/j.joi.2019.03.018)
10. Bedeian AG. 2004 Peer review and the social construction of knowledge in the management discipline. *Acad. Manage. Learn. Edu.* **3**, 198–216. (doi:10.5465/amle.2004.13500489)
11. Siler K, Lee K, Bero L. 2015 Measuring the effectiveness of scientific gatekeeping. *Proc. Natl Acad. Sci. USA* **112**, 360–365. (doi:10.1073/pnas.1418218112)

12. Paine CET, Fox CW. 2018 The effectiveness of journals as arbiters of scientific impact. *Ecol. Evol.* **8**, 9566–9585. (doi:10.1002/ece3.4467)
13. Atjonen P. 2019 Peer review in the development of academic articles: experiences of Finnish authors in the educational sciences. *Learned Publishing* **32**, 137–146. (doi:10.1002/leap.1204)
14. Seeber M. 2020 How do journals of different rank instruct peer reviewers? Reviewer guidelines in the field of management. *Scientometrics* **122**, 1387–1405. (doi:10.1007/s11192-019-03343-1)
15. Garcia-Costa D, Squazzoni F, Mehmani B, Grimaldo F. 2022 Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. *PeerJ* **10**, e13539. (doi:10.7717/peerj.13539)
16. Bianchi F, García-Costa D, Grimaldo F, Squazzoni F. 2022 Measuring the effect of reviewers on manuscript change: a study on a sample of submissions to Royal Society journals (2006–2017). *J. Informetrics* **16**, 101316. (doi:10.1016/j.joi.2022.101316)
17. Casnici N, Grimaldo F, Gilbert N, Dondio P, Squazzoni F. 2017 Assessing peer review by gauging the fate of rejected manuscripts: the case of the Journal of Artificial Societies and Social Simulation. *Scientometrics* **113**, 533–546. (doi:10.1007/s11192-017-2241-1)
18. Teplitskiy M. 2016 Frame search and re-search: how quantitative sociological articles change during peer review. *Am. Sociol.* **47**, 264–288. (doi:10.1007/s12108-015-9288-3)
19. Horbach SPJM. 2021 No time for that now! Qualitative changes in manuscript peer review during the COVID-19 pandemic. *Res. Eval.* **30**, rvaa037. (doi:10.1093/reseval/rvaa037)
20. Köhler T, González-Morales MG, Banks GC, O'Boyle EH, Allen JA, Sinha R, Woo SE, Gulick LMV. 2020 Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: introducing a competency framework for peer review. *Ind. Organ. Psychol.* **13**, 1–27. (doi:10.1017/iop.2019.121)
21. Davis WE, Giner-Sorolla R, Lindsay DS, Lougheed JP, Makel MC, Meier ME, Sun J, Vaughn LA, Zelenski JM. 2018 Peer-review guidelines promoting replicability and transparency in psychological science. *Adv. Methods Practices Psychol. Sci.* **1**, 556–573. (doi:10.1177/2515245918806489)
22. Cobo E *et al.* 2011 Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. *BMJ* **343**, d6783. (doi:10.1136/bmj.d6783)
23. Hirst A, Altman DG. 2012 Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS ONE* **7**, e35621. (doi:10.1371/journal.pone.0035621)
24. Squazzoni F *et al.* 2020 Unlock ways to share data on peer review. *Nature* **578**, 512–514. (doi:10.1038/d41586-020-00500-yz)
25. Fyfe A, McDougall-Waters J, Moxham N. 2015 350 years of scientific periodicals. *Notes Records* **69**, 227–239. (doi:10.1098/rsnr.2015.0036)
26. Paltridge B. 2015 Referees' comments on submissions to peer-reviewed journals: when is a suggestion not a suggestion? *Stud. Higher Edu.* **40**, 106–122. (doi:10.1080/03075079.2013.818641)
27. Bravo G, Grimaldo F, López-Iñesta E, Mehmani B, Squazzoni F. 2019 The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nat. Commun.* **10**, 322. (doi:10.1038/s41467-018-08250-2)
28. Eve MP, Neylon C, O'Donnell DP, Moore S, Gadie R, Odeniyi V, Parvin S. 2021 *Reading peer review. PLOS ONE and institutional change in academia*. Cambridge, UK: Cambridge University Press.
29. Ghosal T, Kumar S, Bharti PK, Ekbal A. 2022 Peer review analyze: a novel benchmark resource for computational analysis of peer reviews. *PLoS ONE* **17**, e0259238. (doi:10.1371/journal.pone.0259238)
30. Bayarri MJ, Berger JO, Forte A, García-Donato G. 2012 Criteria for Bayesian model choice with application to variable selection. *Ann. Stat.* **40**, 1550–1577. (doi:10.1214/12-AOS1013)
31. Li Y, Clyde MA. 2015 Mixtures of g-priors in generalized linear models. (https://arxiv.org/abs/1503.06913)
32. Clyde M. 2020 *BAS: Bayesian variable selection and model averaging using bayesian adaptive sampling*. R package version 1.5.5.
33. Bravo G, Farjam M, Grimaldo F, Birukou A, Squazzoni F. 2018 Hidden connections: network effects on editorial decisions in four computer science journals. *J. Informetrics* **12**, 101–112. (doi:10.1016/j.joi.2017.12.002)
34. Plummer M. 2019 *rjags: Bayesian graphical models using MCMC*. R package version 4–10.
35. Squazzoni F, Bravo G, Farjam M, Marusic A, Mehmani B, Willis M, Birukou A, Dondio P, Grimaldo F. 2021 Peer review and gender bias: a study on 145 scholarly journals. *Sci. Adv.* **7**, eabd0299. (doi:10.1126/sciadv.abd0299)
36. Casnici N, Grimaldo F, Gilbert N, Squazzoni F. 2017 Attitudes of referees in a multidisciplinary journal: an empirical analysis. *J. Assoc. Inf. Sci. Technol.* **68**, 1763–1771. (doi:10.1002/asi.23665)
37. Wolfram D, Wang P, Abuzahra F. 2021 An exploration of referees' comments published in open peer review journals: the characteristics of review language and the association between review scrutiny and citations. *Res. Eval.* **30**, rvab005. (doi:10.1093/reseval/rvab005)
38. Sabaj Meruane O, Gonzalez Vergara C, Pina-Stranger A. 2016 What we still don't know about peer review. *J. Sch. Publishing* **47**, 180–212. (doi:10.3138/jsp.47.2.180)
39. Squazzoni F, Grimaldo F, Marušić A. 2017 Publishing: journals could share peer-review data. *Nature* **546**, 352. (doi:10.1038/546352a)
40. Stephen D. 2022 Peer reviewers equally critique theory, method, and writing, with limited effect on the final content of accepted manuscripts. *Scientometrics* **127**, 3413–3435. (doi:10.1007/s11192-022-04357-y)
41. Horn SA. 2016 The social and psychological costs of peer review: stress and coping with manuscript rejection. *J. Manage. Inquiry* **25**, 11–26. (doi:10.1177/1056492615586597)
42. Crijns TJ, Ottenhoff JSE, Ring D. 2021 The effect of peer review on the improvement of rejected manuscripts. *Account. Res.* **28**, 517–527. (doi:10.1080/08989621.2020.1869547)
43. Hesterman CM, Szperka CL, Turner DP. 2018 Reasons for manuscript rejection after peer review from the journal Headache. *Headache* **58**, 1511–1518. (doi:10.1111/head.13343)
44. Falk Delgado A, Garretson G. 2019 The language of peer review reports on articles published in the BMJ, 2014–2017: an observational study. *Scientometrics* **120**, 1225–1235. (doi:10.1093/reseval/rvab005)
45. Buljan I, Garcia-Costa D, Grimaldo F, Squazzoni F, Marušić A. 2020 Meta-research: large-scale language analysis of peer review reports. *eLife* **9**, e53249. (doi:10.7554/eLife.53249)
46. Sueur HL, Dagliati A, Buchan I, Whetton AD, Martin GP, Dornan T, Geifman N. 2020 Pride and prejudice—what can we learn from peer review? *Med. Teach.* **42**, 1012–1018. (doi:10.1080/0142159X.2020.1774527)
47. Akbaritabar A, Stephen D, Squazzoni F. 2022 A study of referencing changes in preprint-publication pairs across multiple fields. *J. Informetrics* **16**, 101258. (doi:10.1016/j.joi.2022.101258)
48. Nicholson DN, Rubinetti V, Hu D, Thielk M, Hunter LE, Greene CS. 2022 Examining linguistic shifts between preprints and publications. *PLoS Biol.* **20**, e3001470. (doi:10.1371/journal.pbio.3001470)
49. Ortega J. 2017 Are peer-review activities related to reviewer bibliometric performance? A scientometric analysis of Publons. *Scientometrics* **112**, 947–962. (doi:10.1007/s11192-017-2399-6)
50. Severin A, Chataway J. 2021 Purposes of peer review: a qualitative study of stakeholder expectations and perceptions. *Learned Publishing* **34**, 144–155. (doi:10.1002/leap.1336)
51. García-Costa D, Forte A, López-Iñesta E, Squazzoni F, Grimaldo F. 2022 Does peer review improve the statistical content of manuscripts? A study on 27 467 submissions to four journals. Figshare. (doi:10.6084/m9.figshare.c.6174474)

Apéndice D

Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals

Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals

Daniel Garcia-Costa[1], Flaminio Squazzoni[2], Bahar Mehmani[3] and Francisco Grimaldo[1]

[1] Department of Computer Science, University of Valencia, Valencia, Spain
[2] Department of Social and Political Sciences, University of Milan, Milan, Lombardy, Italy
[3] STM Journals, Elsevier, Amsterdam, The Netherlands

ABSTRACT

Reviewers do not only help editors to screen manuscripts for publication in academic journals; they also serve to increase the rigor and value of manuscripts by constructive feedback. However, measuring this developmental function of peer review is difficult as it requires fine-grained data on reports and journals without any optimal benchmark. To fill this gap, we adapted a recently proposed quality assessment tool and tested it on a sample of 1.3 million reports submitted to 740 Elsevier journals in 2018–2020. Results showed that the developmental standards of peer review are shared across areas of research, yet with remarkable differences. Reports submitted to social science and economics journals show the highest developmental standards. Reports from junior reviewers, women and reviewers from Western Europe are generally more developmental than those from senior, men and reviewers working in academic institutions outside Western regions. Our findings suggest that increasing the standards of peer review at journals requires effort to assess interventions and measure practices with context-specific and multi-dimensional frameworks.

Submitted 23 December 2021
Accepted 13 May 2022
Published 7 June 2022

Corresponding author
Francisco Grimaldo,
francisco.grimaldo@uv.es

Academic editor
Harry Hochheiser

Additional Information and Declarations can be found on page 19

DOI 10.7717/peerj.13539

Subjects Science Policy, Data Science
Keywords Peer review, Standards, Reviewers, Academic journals, Natural language processing

INTRODUCTION

Peer review is key for public trust in science (*Bornmann, 2011*). Vetting scientific claims from authors who can often be over-confident and biased towards their own findings before publication is one of the main functions of academic journals. This ensures that only rigorous research reaches public visibility and informs medical treatment, technology innovations and public decisions (*Kharasch et al., 2021*). However, by ensuring high standards of review reports, journals also contribute to improve the value of manuscripts, so enhancing mutual learning between experts (*Rigby, Cox & Julian, 2018*). These two functions of peer review can be called: "quality screening" and "developmental" function (*Lewin, 2014*; *Seeber, 2020*; *Akbaritabar, Stephen & Squazzoni, 2022*).

While the 'publish or perish' academic culture and obsession for rapid dissemination of scientific findings are posing several challenges to peer review (*Edwards & Siddhartha,*

How to cite this article Garcia-Costa D, Squazzoni F, Mehmani B, Grimaldo F. 2022. Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. *PeerJ* **10**:e13539 http://doi.org/10.7717/peerj.13539

2017), including the recent impact of the fast track publication of COVID-19 pandemic research (*Squazzoni et al., 2021b*; *Horbach, 2021*; *Sullivan et al., 2022*), there is no consensus on how to measure the standards of peer review. While research on attitudes, practices and writing styles of peer reviewers has recently grown thanks to the availability of original data from Publons, various open peer review repositories or single journals (*Casnici et al., 2017*; *Buljan et al., 2020*; *Wolfram, Wang & Abuzahra, 2021*; *Stephen, 2022*; *Rice et al., 2022*; *Thelwall, 2022*), only in a few cases, data include full information on manuscripts and reviewers from different journals and research areas (*Squazzoni et al., 2020*). Furthermore, measuring the quality of peer review is difficult (*Cowley, 2015*). Indeed, efforts have been made to measure the quality of review reports since the end of the 1990s in biomedical research (*Van Rooyen, Black & Godlee, 1999*; *Jefferson, Wager & Davidoff, 2002*; *Van Rooyen, 2001*; *Schroter et al., 2004*; *Schroter et al., 2006*), especially as a means to estimate the efficacy of interventions, *e.g.*, reviewer training. However, there is no systematic measurement of peer review standards that can help assess the state-of-the-art of peer review in various areas of research. This does not permit a rigorous assessment of innovations on peer review at journals, thus undermining the adoption of an evidence-based approach to peer review reforms (*Squazzoni et al., 2020*).

In order to fill this gap, *Superchi et al. (2020)* have recently proposed Arcadia (Assessment of Review reports with a Checklist Available to eDItors and Authors), a tool to assess the quality of peer review reports in biomedical research. By surveying 446 biomedical editors and authors, they identified a checklist for the quality of review reports including five domains as follows: Importance of the study (*i.e.*, the contribution and relevance of the study); robustness of the study methods (*i.e.*, the soundness of the study methods); interpretation and discussion of the study results (*i.e.*, coherence of the study conclusions compared to research questions, external validity and study limitations); reporting and transparency of the manuscript (*i.e.*, data sharing, report guidelines and reproducibility); and characteristics of the peer reviewer's comments (*i.e.*, clarity, objectivity and constructiveness of reviewer comments). They defined quality as: "the extent to which a peer review report helps, first, editors make an informed and unbiased decision about the manuscripts' outcome and, second, authors improve the quality of the submitted manuscript", thus combining both functions of peer review, *i.e.*, quality screening and developmental function.

While this study has improved our understanding of peer review compared to previous research (*Superchi, González & Solá, 2019*), especially in terms of external validation of measurements, their sample of respondents was limited to biomedical experts and the validation test was only subjective, *i.e.*, reflecting opinion of experts rather than current practices (*Pranić et al., 2021*). Given that practices, norms and models of peer review are heterogeneous and field-specific (*Horbach & Halffman, 2018*; *Merriman, 2020*), there is no optimal benchmark to assess current practices and behaviors of peer review across different areas of research. Here, extending measurements to various research areas is key to increase comparability and provide a baseline for future research.

To fill this gap, we have adapted Arcadia and tested it against a rich database of 1.3 million review reports from 740 Elsevier journals. Data from Elsevier journal management systems were further enriched with data on reviewers and journals from Scopus and other sources. Our aim was to use Arcadia to examine the developmental function of peer review by including multi-dimensional measurements and developing a score that would permit systematic comparisons of peer review reports from different research areas.

We first translated Arcadia into a vocabulary to map the text of peer review reports. This allowed us to provide a multi-dimensional developmental score to compare and assess reports per area of research and reviewer characteristics. We guessed each reviewer's gender, and reconstructed their seniority and geographical/institutional location. We classified journals in quartiles of impact factor using Web of Science. This was to estimate the effects of various factors, either reviewer, field, or journal specific, on report standards. Rather than using humans to rate the quality dimensions of peer review as in previous research (*Superchi et al., 2020*), we used data to measure the current standards and practices of reporting in various journals (*Bianchi, Grimaldo & Squazzoni, 2019*). While the concept of 'quality' is hard to quantify due to its complexity and the co-existence of various goals and stakeholders, measuring standards of reports by means of natural language processing techniques on contents can help us to consider multi-dimensional factors without restricting the observation sample for the sake of human raters (*Ghosal et al., 2022*).

MATERIALS AND METHODS

Dataset

Data access required a confidential agreement to be signed on 12th May 2020 between Elsevier and each author of this study. The agreement was inspired by the PEERE protocol for data sharing and included anonymization, privacy, data management and security policies jointly determined by all partners (*Squazzoni, Grimaldo & Marusic, 2017*).

The whole dataset included 1,331,941 reviewer reports from 740 Elsevier journals in all areas of research: Life sciences (hereafter, LS), physical sciences (hereafter, PS), health and medical sciences (hereafter, HMS), and social sciences and economics (hereafter, SSE). Reports referred to the first round of peer review and were related to research manuscripts. Thanks to an ex-post integration and data enrichment from multiple sources, including Elsevier journal data, Scopus for additional information on reviewers and Web of Science for information on journals, we inferred each reviewer's gender, seniority, and country of affiliation. We also had information on the final editorial decision associated with each manuscript, the report time, and the review status. Given the relatively few cases of journals listed among the fourth quartile of impact factor and for the sake of our analysis, we decided to merge Q4 and non-indexed journals in the same category.

Tables 1, 2 and 3 show the number of journals and reports per area of research, journal quartile and reviewers' geographical location. Table 4 shows that women ensured only about 22% of reports, confirming recent findings on the weak involvement of women as reviewers (*Helmer et al., 2017*; *Publons, 2018*; *Stockemer, 2022*).

Each review report was cleaned and standardized by converting to lowercase, removing all non-alphanumerical characters, standardizing breaklines and separator characters and

PeerJ

Table 1 **Number of journals per quartile of impact factor and area of research.**

	PS	SSE	HMS	LS	Total
Journals	333	99	174	134	740
Journals Q1	161	45	40	38	283
Journals Q2	110	20	40	49	219
Journals Q3	29	17	32	27	105
Journals Q4	8	3	7	3	21
Journals NI	25	14	55	18	112

Table 2 **Number of reviews per journal quartile and area of research.**

	PS	SSE	HMS	LS	Total
Reviews	825.247	171.070	150.296	185.328	1.331.941
Reviews Q1	602.763	146.088	51.860	88.089	888.800
Reviews Q2	165.506	18.422	40.733	61.104	285.765
Reviews Q3	29.743	5.147	46.596	26.375	107.861
Reviews Q4	2.104	468	3.236	978	6.786
Reviews NI	25.131	945	7.871	8.782	42.729

Table 3 **Number of reviews per reviewers' geographical location and area of research.** (Note: Countries are classified according to ISO 3166 country codes, while their aggregation complies with the United Nation M49 standard).

	PS	SSE	HMS	LS	Total
Northern America	120392	64254	52027	52763	289436 (21.73%)
Western Europe	64603	16539	14798	17923	113863 (8.55%)
Eastern Asia	290125	32140	20583	37496	380344 (28.56%)
Southern Asia	57994	3880	6450	7124	75448 (5.66%)
Northern Europe	46505	16048	13235	12387	88175 (6.62%)
Eastern Europe	37165	1722	3622	5935	48444 (3.64%)
Latin America and the Caribbean	34886	2791	6329	11713	55719 (4.18%)
Southern Europe	85495	11726	15388	21733	134342 (10.09%)
South-East Asia	19079	2158	1995	3378	26610 (2.00%)
Western Asia (Middle East)	27071	5211	5653	5121	43056 (3.23%)
Australia and New Zealand	24925	12463	5716	5756	48860 (3.67%)
Northern Africa	8006	317	2300	1383	12006 (0.90%)
Central Asia	306	16	13	35	370 (0.03%)
Sub-Saharan Africa	5254	750	955	1201	8160 (0.61%)
Micronesia	134	24	74	52	284 (0.02%)
Melanesia	70	10	7	17	104 (0.01%)
Polynesia	23	3	2	5	33 (0.00%)
Missing	3214	1018	1149	1306	6687 (0.50%)

Table 4 Number of reviews per gender, seniority and area of research.

	PS	SSE	HMS	LS	Total (%)
Women	148807	44927	41754	59562	295050 (22.15%)
Men	645547	120529	106096	121488	993660 (74.60%)
Missing gender	30893	5614	2446	4278	43231 (3.25%)
<5 years	21365	9463	3867	4070	38765 (2.91%)
5 to 18 years	435892	101008	67912	85074	689886 (51.80%)
>18 years	335270	51557	69659	86483	542969 (40.77%)
Missing seniority	32720	9042	8858	9701	60321 (4.53%)

removing repeated white spaces, converting webpage links and reference citations to tokens, removing stop words and words stemming only from the root of each word. Note that after estimating the length of each report, we decided to remove outliers to avoid biasing our analysis. The final dataset included 1,331,941 review reports.

Standard measurements

In order to estimate peer review standards, we started from Arcadia, a recently released checklist to assess the quality of peer review reports in biomedical research (*Superchi et al., 2020*). Arcadia considers five domains and 14 items, including: Contribution; Relevant literature; Study methods; Statistical methods; Study conclusions; Study limitations; Applicability and generalizability; Study protocol; Reporting; Presentation and organization; Data availability; Clarity; Constructiveness; and Objectivity.

However, considering the specific purposes of Arcadia and its focus restricted to biomedical journals, we added modifications necessary to reflect the characteristics of our dataset, including journals from different areas of research. After translating items into words and running some preliminary test, we decided to merge 'Reporting' with 'Applicability and generalizability' and separe 'Presentation' from 'Organization'. We extracted 'Clarity' by means of readability metrics and decided to disregard 'Constructiveness and objectivity' because these dimensions were hardly quantifiable in our dataset.

This led us to concentrate on the following developmental dimensions:

- **Impact**, *i.e.*, comments from reviewers on the impact of findings or any other manuscript content on society, the economy or whatever external stakeholders, and the study contribution.
- **Relevant literature (literature)**, *i.e.*, comments of reviewers concerning the state-of-the-art and the manuscript references.
- **Study methods (methods)**, *i.e.*, comments from reviewers on materials, methods, and the study design.
- **Statistical methods (statistics)**, *i.e.*, comments from reviewers regarding the statistical content of the study.
- **Study conclusions (conclusions)**, *i.e.*, comments from reviewers on results and conclusions.
- **Limitations**, *i.e.*, comments from reviewers regarding study limitations.

- **Applicability**, comments from reviewers concerning the applicability, generalizability and reproducibility of the study.
- **Presentation**, *i.e.*, comments from reviewers about the presentation of the manuscript, and the quality/readability of tables, figures, and other visualizations.
- **Data availability (data)**, *i.e.*, comments from reviewers regarding data availability.
- **Organization and writing (writing)**, *i.e.*, comments from reviewers about the organization and the linguistic content and style of writing of the manuscript.

Dictionary building

In order to build a dictionary and also given the characteristics of our dataset, we decided to follow a semi-automatic dictionary building approach, which mostly ensured similar results to manually built dictionaries (*Muresan & Klavans, 2002*; *Godbole et al., 2010*; *Deng et al., 2017*; *Deng et al., 2019*; *Mpouli, Beigbeder & Largeron, 2020*). Given the very large corpus of textual data and the possibility of relying on a predefined list of developmental dimensions extracted from Arcadia, we used manual checks on the output of each iteration to verify the process and minimize possible mistakes.

We followed five steps: (1) corpus creation, (2) pre-processing and cleaning, (3) vector representation of the corpus, (4) term extraction and (5) validation, which included steps 4 and 5 to be repeated several times (see the full process in Fig. 1). In step 2, we converted the text into lower case, removed non-alphanumeric characters, trimmed white spaces and line breaks, tokenised web links and citations, removed stop words and finally applied stemming to standardize words. In step 3, we built an unsupervised Word2Vec model using the H2O API (https://www.h2o.ai/) in R (https://www.r-project.org/) to create a vector representation of our corpus. We departed from an initial list of manually defined terms by revising a sample of review reports and selecting ten terms for each dimension (see Table 5). By using bigrams, we minimised context-specific ambiguities while categorizing individual words.

In step 4, we used the Word2Vec model to search for near terms in all review texts. We extracted new terms by running the method 'findSynonyms' from the H2O API and selected the most frequent similar terms (*i.e.*, those with a normalized score, returned by this method, higher than 0.75) and listed among list candidates. The identification of non-existing unigram and bigram terms required different procedures: whenever a new bigram term was selected, we checked if any of its words already existed as unigram terms and, if so, the term was dropped out.

In step 5, we validated the list of new terms. We used a KWIC method to validate each new term, by checking the context in which the term was used throughout the corpus, obtaining some examples and assessing whether the term was appropriate or not. Given that this was context-dependent, we opted for a manual validation performed by a male PhD student (val1) and a female Master degree student (val2), with a male senior researcher (val3) decisive in case of any conflicting assessments. Note that these were all domain experts. During such a validation step, these experts were allowed to manually check when an unigram was dropped out due to its ambiguity by reconstructing the context and eventually converting the unigram to the correct bigram.

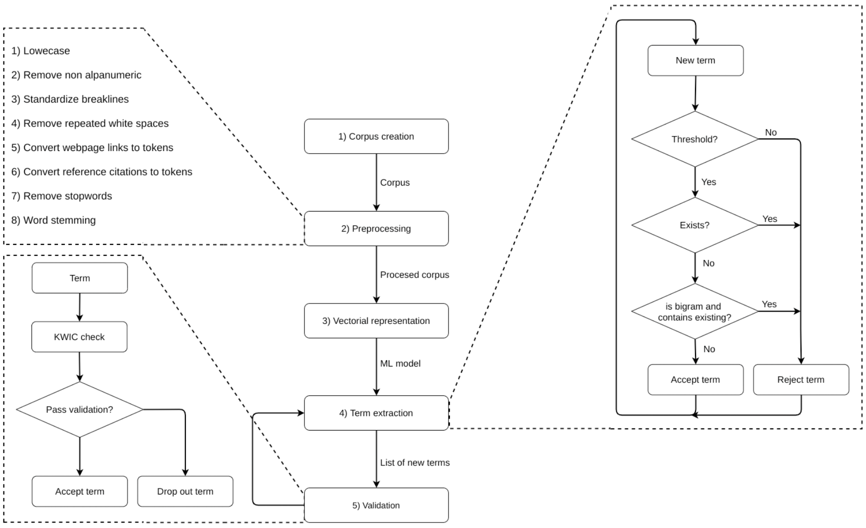

Figure 1 Steps of the dictionary building process.

Full-size DOI: 10.7717/peerj.13539/fig-1

This allowed us to use the output list of terms from the previous iteration as input for each new iteration. Any extraction and validation step was repeated until all new terms had low frequency values. This allowed us to obtain a total of 1,565 terms (see Table 6 for the distribution of terms of each dimension of the developmental score).

These final list of terms was then used to build a LIWC (Linguistic Inquiry and Word Count) (https://liwc.wpengine.com/) style dictionary. While our dictionary could be used in LIWC or any other program or library which accepts LIWC style dictionaries, here we used the package "quanteda.dictionaries" (https://github.com/kbenoit/quanteda.dictionaries) to estimate the developmental values for each review report in our dataset. These values reflected the number of words found from each category in the text reports.

Developmental score

Given that the distribution of developmental terms followed a Zipfian distribution with discrete different scales, aggregating all dimensions into a single score required to avoid that any specific dimension would predominate over others. To avoid this, we normalized each dimension by using the empirical cumulative distribution function (ECDF), thus transforming the discrete word count values into a real scale between 0 and 1. We used the arithmetic mean of these standardized values to aggregate them and generated a unique score for each report. Whenever a report did not contain any word from a certain dimension, its assigned value was 0.

The calculus of the score followed this formula:

$$Score = \frac{1}{n} * \sum_{i=1}^{n} F_{D_i}(v_i)$$

Table 5 Initial seed terms for each developmental dimension.

Developmental dimensions	Initial seed terms
Impact	relevant, impact, novel, original, innovator paper, interest paper, disappointing paper, important topic, relevant paper, research community
Relevant Literature	cite, consider reference, require reference, reference paper, related work, literature, bibliography, similar work, previous work, existing work
Study Methods	methodology, approach, experiment, techniques, analysis, procedures, provide justification, provide comparison, exploratory, meticulous
Statistical Methods	statistics, null hypothesis, regression, coefficient, significance, correlation, deviation, Bayesian, response variable, effect size
Study Conclusions	result, discussion, conclusion, findings, research question, unjustified, evidence, inconsistency, unsolved problem, explanation
Limitations	limitations, weakness, robustness, future work, lack acknowledg, acknowledg limit, expertise, under-investigated, flaws, bottleneck
Applicability	work applicability, application domain, reproducible, generalizable results, generalizable study, scalable, transferable, irreproducible, reusable, universal method
Presentation	table, figure, row, column, image, axis, caption, legend, graph, footer
Data Availability	database, data available, accessible data, experiment data, publish data, repository, source code, opaque, secrecy, available resources
Organization and Writing	rewrite, well written, poor written, reorganize, move, spelling, page, line, sentence, paragraph

Table 6 Number of terms for each dimension of the developmental score.

Item	Num of terms	Item	Num of terms
Impact	175	Literature	235
Methods	240	Statistics	122
Conclusions	283	Limitations	71
Applicability	139	Presentation	72
Data	128	Writing	168

where F_{D_i} was the cumulative distribution function of the discrete variable, while D_i was the Zipfian variable starting in 1. Note that in case $F_{D_i}(v_i) = 0$, no term was found in the report regarding a given dimension. Figure 2 shows the distribution of our developmental score.

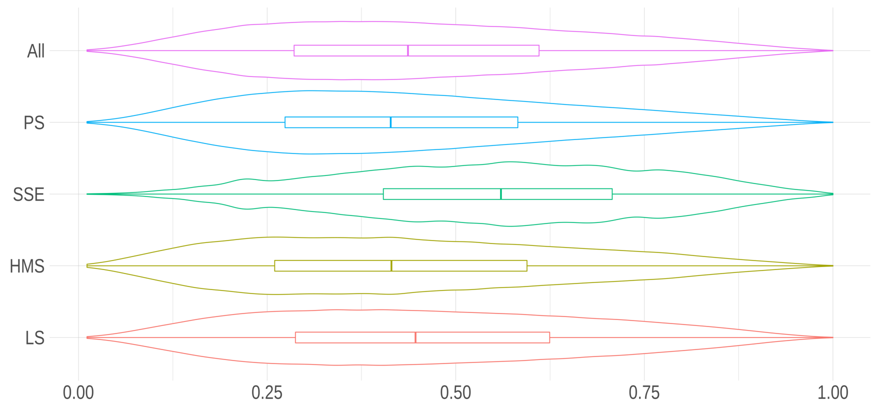

Figure 2 **Distribution of the developmental score per research area.** The density curves in this violin plot show the distribution of the score for all research areas and, separately, for PS, SEE, HMS or LS.
Full-size DOI: 10.7717/peerj.13539/fig-2

Table 7 **Explained variance by each principal component.**

	% of variance	Cumulative % of variance
PC1	39.38	39.38
PC2	9.88	49.26
PC3	8.76	58.02
PC4	7.18	65.20
PC5	6.99	72.19
PC6	6.63	78.82
PC7	5.99	84.81
PC8	5.54	90.36
PC9	5.24	95.59
PC10	4.41	100.00

Score internal validity

By using the R package FactoMineR, we performed a principal component analysis to check the amount of variance for each dimension. Table 7 shows that up to nine principal components were needed to explain at least 95% of our variance. This indicates that there was no correlation between our dimensions, so confirming the main finding from Arcadia (*Superchi et al., 2020*).

Internal consistency

In order to assess the internal consistency of our developmental score, we estimated Cronbach's alphas and the total-item correlation by using the R package Psych. Note that that threshold of acceptance should be greater than 0.70 for alpha, while item-total correlation should be greater than 0.30. Table 8 shows that we achieved a global Cronbach alpha of 0.82. This indicates that there is no developmental dimension that could have

Table 8 Cronbach alphas and Item-total correlations.

Dimension	α if item was dropped	Item-total correlation
Impact	0.81	0.53
Relevant literature	0.82	0.47
Study Methods	0.81	0.60
Statistical Methods	0.80	0.63
Study Conclusions	0.79	0.72
Limitations	0.81	0.55
Applicability	0.81	0.58
Presentation	0.82	0.43
Data availability	0.82	0.49
Organization and writing	0.80	0.65

Table 9 CFA factor loadings for each developmental item.

Indicator	Estimate	Std.Err	P(> |z|)
Impact	1.00	0.00	0.00
Relevant literature	1.01	0.00	0.00
Study methods	1.10	0.00	0.00
Statistical methods	1.19	0.00	0.00
Study conclusions	1.28	0.00	0.00
Limitations	1.19	0.00	0.00
Applicability	1.20	0.00	0.00
Presentation	0.88	0.00	0.00
Data availability	1.06	0.00	0.00
Organization and writing	1.13	0.00	0.00

been dropped that would have increased the value of alpha. Note also that the item-total correlation for each dimension was greater than the recommended minimum value of 0.30. This test demonstrates that our developmental dimensions were consistent throughout the whole sample, without any dimension biasing our measurements.

We also applied an additional method to evaluate consistency, *i.e.*, the Confirmatory Factor Analysis (CFA). Our CFA showed a good fit between model and data, with a CFI value of 0.93, which was greater than the recommended minimum of 0.90, and a RMSEA of 0.07, which was smaller than the recommended maximum of 0.08. As regards coefficients, note that all developmental items had significant p-values (see Table 9).

Gender guessing

Gender was guessed as previously described in *Squazzoni et al. (2021b)*. Specifically, we queried the Python package gender-guesser about the first names and countries of origin, if any. Gender-guesser allowed us to minimize gender bias and achieve the lowest misclassification rate (less than 3% for Benchmark 1 in *Santamaría & Mihaljević (2018)*). For names classified by gender-guesser as 'mostly_male', 'mostly_female', 'andy' (androgynous) or 'unknown' (name not found), we used GenderAPI (https://gender-api.com/), which ensures that the level of mis-classification is around 5% (see Table 4 in *Santamaría & Mihaljević (2018)*) and has the highest coverage on multiple name origins (see Table 5 in *Santamaría & Mihaljević, 2018*). This procedure allowed us to guess the gender of 94.5% of academics in our sample, 45.1% coming from gender-guesser and 49.2% from GenderAPI. The remaining 5.5% of academics were assigned an unknown gender. Note that this level of gender guessing is consistent with the non-classification rate for names of academics in previous research (*Santamaría & Mihaljević, 2018*). Note also that while we were aware that any gender binary definition did not adequately represent non-binary identities, to the best of our knowledge, there was no better instrument to guess gender for such a large pool of individuals.

Seniority

Reviewer seniority was estimated by using the number of years since their first publication record in the Scopus database. This information was retrieved through the Elsevier International Center for the Study of Research (ICSR Lab) computational platform. We used either the Scopus ID, the e-mail address or the full name plus country (in this order of preference) to find a unique matching profile in the Scopus database. We followed a conservative rule and reviewers without a profile in Scopus or with more than a single matching profile (*i.e.*, not being uniquely identifiable) were excluded from the analysis, whenever using seniority as a variable. By following *Squazzoni et al. (2021a)*, we assumed that first publications would correspond to the period in which reviewers were completing their MD or PhD. We then considered a cut-off of 18 years to identify junior vs. senior reviewers, *i.e.*, full professors.

RESULTS

Developmental score

Figure 2 shows that peer review reports submitted to social sciences and economics (SEE) journals showed the highest developmental standards compared to all areas of research. Table 10 shows that SSE reports had the highest scores in all developmental dimensions except for *Presentation*, for which they scored lower than reports from any other area of research. We used a Gamma Generalized Linear Model to analyze the relation with relevant covariates since the developmental score fits this family of distributions (as reported by Generalized Additive Models for Location, Scale and Shape (https://www.gamlss.com/)). Table 11 indicates that the differences in the developmental standards of peer review between areas of research were on average around 10%, with remarkable heterogeneity.

Except for SSE, journals with highest impact factors generally showed higher developmental standards of reports (see Fig. 3). It is interesting to note that the standards of reports in PS and LS did not seem to reflect impact factor hierarchies, as developmental scores were more stable across the first three quartiles than in any other research areas. Interestingly, in SSE journals with a higher impact factor did not show the highest report standards: journals listed among the second and third quartiles in the ranking of impact factor of economics and social science journals had relatively higher standards compared to high-ranked journals (see Fig. 3).

Table 10 Mean and standard deviation (in brackets) for each developmental score dimension per research area.

	PS	SSE	HMS	LS
Impact	0.473 (0.318)	0.662 (0.317)	0.516 (0.329)	0.521 (0.325)
Literature	0.377 (0.371)	0.509 (0.384)	0.328 (0.362)	0.358 (0.371)
Methods	0.527 (0.316)	0.585 (0.31)	0.43 (0.314)	0.479 (0.313)
Statistics	0.442 (0.329)	0.645 (0.329)	0.499 (0.338)	0.484 (0.335)
Conclusions	0.487 (0.302)	0.608 (0.303)	0.521 (0.306)	0.559 (0.307)
Limitations	0.369 (0.361)	0.608 (0.364)	0.437 (0.372)	0.405 (0.375)
Applicability	0.441 (0.361)	0.532 (0.359)	0.423 (0.366)	0.447 (0.369)
Presentation	0.42 (0.36)	0.314 (0.331)	0.322 (0.342)	0.407 (0.365)
Data	0.315 (0.364)	0.512 (0.39)	0.38 (0.377)	0.388 (0.376)
Writing	0.507 (0.31)	0.543 (0.305)	0.492 (0.314)	0.556 (0.313)

Table 11 Effect of research area and journal impact factor on the developmental score using a Gamma Generalized Linear Model with developmental score as response variable.

	Dependent variable:
	Developmental score
AreaHMS	−0.071*** (0.001)
AreaPS	−0.105*** (0.001)
AreaLS	−0.084*** (0.001)
IFQuartileQ2	0.033*** (0.002)
IFQuartileQ3	0.058*** (0.004)
IFQuartileQ4+NI	−0.025*** (0.006)
AreaHMS:IFQuartileQ2	−0.047*** (0.003)
AreaPS:IFQuartileQ2	−0.052*** (0.002)
AreaLS:IFQuartileQ2	−0.035*** (0.002)
AreaHMS:IFQuartileQ3	−0.158*** (0.004)
AreaPS:IFQuartileQ3	−0.069*** (0.004)
AreaLS:IFQuartileQ3	−0.052*** (0.004)
AreaHMS:IFQuartileQ4+NI	−0.060*** (0.007)
AreaPS:IFQuartileQ4+NI	−0.037*** (0.007)
AreaLS:IFQuartileQ4+NI	−0.027*** (0.007)
Constant	0.547*** (0.001)
Observations	1,331,247
Log Likelihood	196,834.500
Akaike Inf. Crit.	−393,636.900

Notes.
*$p < 0.1$.
**$p < 0.05$.
***$p < 0.01$.
Reference categories were: reports submitted to SSE journals listed in the first quartile of impact factor.

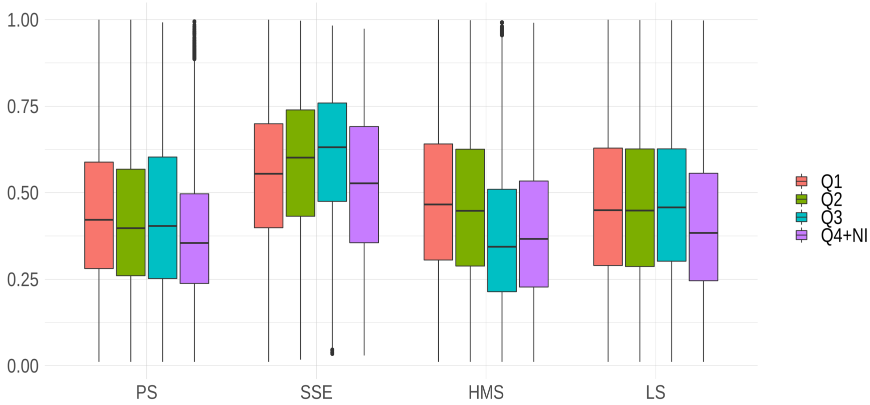

Figure 3 Interaction between journal prestige and research area. Note that due to the restricted number of cases in the sample and for the sake of readability, we included fourth quartile and not-indexed journals in the same category.

Full-size DOI: 10.7717/peerj.13539/fig-3

Table 12 Effect of report delivery time on the developmental score per research area using a Gamma Generalized Linear Model with developmental score as response variable.

	Dependent variable:			
	Developmental score			
	PS	**SSE**	**HMS**	**LS**
Report delivery time	0.002***	0.001***	0.002***	0.003***
	(0.00002)	(0.00003)	(0.00005)	(0.00004)
Constant	0.401***	0.525***	0.400***	0.414***
	(0.0004)	(0.001)	(0.001)	(0.001)
Observations	824,954	171,055	149,978	185,256
Log Likelihood	145,774.700	21,203.550	18,616.660	21,472.280
Akaike Inf. Crit.	−291,545.500	−42,403.110	−37,229.320	−42,940.560

Notes.
*$p < 0.1$.
**$p < 0.05$.
***$p < 0.01$.

However, the higher developmental score of reports seems to come at a price: in SSE journals, the median delivery time of reviewers is 24 days against 15 days for reviewers from HMS, 17 days in LS, and 19 days in PS journals. Table 12 shows a positive correlation between delivery time and developmental score of reports. Although various factors could influence the turn-round time of reports, including editorial standards of reminders, this would suggest a potential trade-off between the developmental content of reports and quick editorial decisions (*Sullivan et al., 2022*).

Reviewer characteristics

Here, we aimed to examine whether the developmental score of reports could reflect certain reviewer characteristics, such as gender and seniority. When considering reviewer gender,

Table 13 Effect of gender and seniority on the developmental score per area of research using a Gamma Generalized Linear Model with developmental score as response variable.

	Dependent variable:			
	Developmental score			
	PS	**SSE**	**HMS**	**LS**
Seniority 5 to 18 years	−0.050***	−0.064***	−0.039***	−0.033***
	(0.004)	(0.005)	(0.007)	(0.006)
Seniority > 18 years	−0.065***	−0.090***	−0.058***	−0.058***
	(0.004)	(0.005)	(0.007)	(0.006)
Gender Man	−0.018***	−0.087***	−0.089***	−0.054***
	(0.004)	(0.005)	(0.008)	(0.008)
Seniority 5 to 18 years: Gender Man	0.003	0.074***	0.026***	0.0001
	(0.004)	(0.005)	(0.009)	(0.008)
Seniority > 18 years: Gender Man	0.005	0.099***	0.020***	0.015***
	(0.004)	(0.006)	(0.009)	(0.008)
Constant	0.504***	0.630***	0.532***	0.538***
	(0.003)	(0.005)	(0.007)	(0.006)
Observations	762,864	156,575	138,933	171,641
Log Likelihood	129,470.100	19,139.370	17,814.220	19,073.790
Akaike Inf. Crit.	−258,928.200	−38,266.750	−35,616.450	−38,135.570

Notes.
*$p < 0.1$.
**$p < 0.05$.
***$p < 0.01$.
Reference categories were: women reviewers with < 5 years of seniority. Note that seniority was estimated by looking at the first publication of each reviewer indexed in Scopus.

we did not find any considerable effects on report standards. The only exception were reports submitted to SSE and HMS journals, where reports from women obtained scores approximately 8% higher than those submitted by men (see Table 13). This would confirm recent research reporting weak gender effects on reviewer attitudes, recommendations and writing styles in various research areas and journal contexts (*Bravo et al., 2019*; *Buljan et al., 2020*; *Bolek et al., 2022*).

When considering reviewer seniority (for detail on the measurement of seniority, see the Method Section), we found a difference of 8% between junior and senior reviewers in all research areas. Junior reviewers generally ensure comparatively highest developmental standards of reports (see Table 13). For instance, in SSE journals, reports from juniors scored around 10% higher than those submitted by seniors. While this could simply reflect the fact that seniors would be more concise in their reports or have less time for reviews (*Hochberg, 2010*; *Merrill, 2014*; *Bianchi et al., 2018*), the higher developmental scores of reports from junior scholars could also reflect reputation building strategies, *e.g.*, showing their diligence and reliability to journal editors in view of potential future submissions (*Mahmić-Kaknjo, Utrobičić & Marušić, 2021*).

Institutional and geographical factors

Here, we aimed to examine whether institutional or geographical factors could influence the developmental score of reports. This was to consider potential heterogeneity in practices and style of reviewing (*Publons, 2018*). Indeed, our results showed considerable variations of the developmental score when controlling for the institutional and geographical embeddedness of reviewers. Although with certain specificities due to research areas, results indicate that reviewers from Western Europe would have higher developmental standards compared to reviewers from other regions, except for reports submitted to HSM and LS journals, though with a very weak statistical difference (about 1%). Table 14 shows that reports submitted by reviewers from Asian regions would be less developmental (10–15% lower than reviewers from Europe).

Figure 4 shows the distribution of the developmental score per dimension and institutional and geographical origins of reviewers. The distribution suggests that report scores were generally higher for writing, conclusions, methods and impact, thus confirming research showing that reviewers would tend to concentrate more preferably on certain aspects of manuscripts (*Siler, Lee & Bero, 2015*; *Herber et al., 2020*; *Teplitskiy et al., 2018*; *Stephen, 2022*). Data and limitations showed lower scores, the latter also showing the greatest variation in the score distribution per region. More importantly, our results showed that reports from reviewers from Northern America scored higher on data and statistics than reports from reviewers from Western Europe and any other region. Note also that reports from reviewers from various regions greatly varied as to how they focused on the way the text of manuscripts was written and organized.

DISCUSSION AND CONCLUSIONS

Although academic journals have been recently threatened by the need for rapid dissemination of scientific information, their real hallmark is their capacity to maintain rigorous standards of peer review. This is key to ensure that scientific claims can be trusted by the public (*Kharasch et al., 2021*). This has been especially important during the recent pandemic and will also be so in the post-pandemic science (*Bauchner, Fontanarosa & Golub, 2020*; *Palayew et al., 2020*). However, this requires that each report submitted by reviewers meets the highest professional standards, which is also instrumental in maintaining the credibility and legitimacy of journals for authors who submit their manuscripts (*Pranić et al., 2021*).

Our research shows that standards of peer review are robust though with certain field-specific characteristics. The fact that developmental standards of peer review are higher in SSE journals would confirm the specificity of the historical institutional trajectory of peer review in these fields. As suggested by previous research (*Huisman & Smits, 2017*; *Merriman, 2020*), editorial standards of journals in these fields typically include double anonymized peer review and a tendency towards more constructive and elaborated reports. Furthermore, while the debate is open on the editorial standards of top journals in this area of research and their excessive prominence and concentration (*Card & DellaVigna, 2013*; *Teele & Thelen, 2017*; *Akbaritabar & Squazzoni, in press*), our findings would reveal that more specialized

Table 14 The effect of the geographical location of reviewers on the developmental score per area of research.

	Dependent variable:			
	Developmental score			
	PS	SSE	HMS	LS
Southern Asia	−0.133***	−0.081***	−0.105***	−0.170***
	(0.001)	(0.004)	(0.003)	(0.003)
Northern Europe	−0.018***	−0.026***	0.014***	0.005*
	(0.001)	(0.002)	(0.003)	(0.003)
Southern Europe	−0.036***	−0.032***	−0.036***	−0.071***
	(0.001)	(0.003)	(0.003)	(0.002)
Northern Africa	−0.131***	−0.158***	−0.113***	−0.148***
	(0.002)	(0.009)	(0.004)	(0.005)
Sub-Saharan Africa	−0.052***	−0.167***	−0.021***	−0.055***
	(0.003)	(0.006)	(0.007)	(0.006)
Latin America and the Caribbean	−0.057***	−0.083***	−0.038***	−0.080***
	(0.001)	(0.004)	(0.003)	(0.003)
Western Asia (Middle East)	−0.115***	−0.059***	−0.118***	−0.135***
	(0.001)	(0.003)	(0.003)	(0.003)
Australia and New Zealand	−0.041***	−0.028***	0.038 ***	0.011 ***
	(0.002)	(0.003)	(0.004)	(0.004)
Eastern Europe	−0.082***	−0.083***	−0.064***	−0.084***
	(0.001)	(0.005)	(0.004)	(0.003)
Northern America	−0.031***	−0.069***	0.001	−0.013***
	(0.001)	(0.002)	(0.002)	(0.002)
South-East Asia	−0.095***	−0.072***	−0.055***	−0.103***
	(0.002)	(0.005)	(0.005)	(0.004)
East Asia	−0.145***	−0.131***	−0.125***	−0.169***
	(0.001)	(0.002)	(0.002)	(0.002)
Constant	0.521***	0.617***	0.467***	0.527***
	(0.001)	(0.002)	(0.002)	(0.002)
Observations	821,213	169,984	148,745	183,842
Log Likelihood	167,148.900	23,405.270	21,644.070	28,294.330
Akaike Inf. Crit.	−334,271.800	−46,784.550	−43,262.150	−56,562.670

Notes.

Countries are classified according to ISO 3166 country codes, while their aggregation complies with the United Nation M49 standard. In case of Sub-Saharan Africa, more than the 50% of our observations included reviewers located in South Africa). We used a Gamma Generalized Linear Model with developmental score as response variable.

*$p < 0.1$.

**$p < 0.05$.

***$p < 0.01$.

The reference category were Western European reviewers.

or relatively newly established journals are more keen to adopt developmental peer review, with reviewers probably more encouraged to provide constructive and elaborate reports (*Merriman, 2020*). Furthermore, the fact that standards were more homogeneous across PS and LS journals, at least those listed among the first three quartile of impact factor, would suggest that in these fields, there are more consistent standards of evaluation.

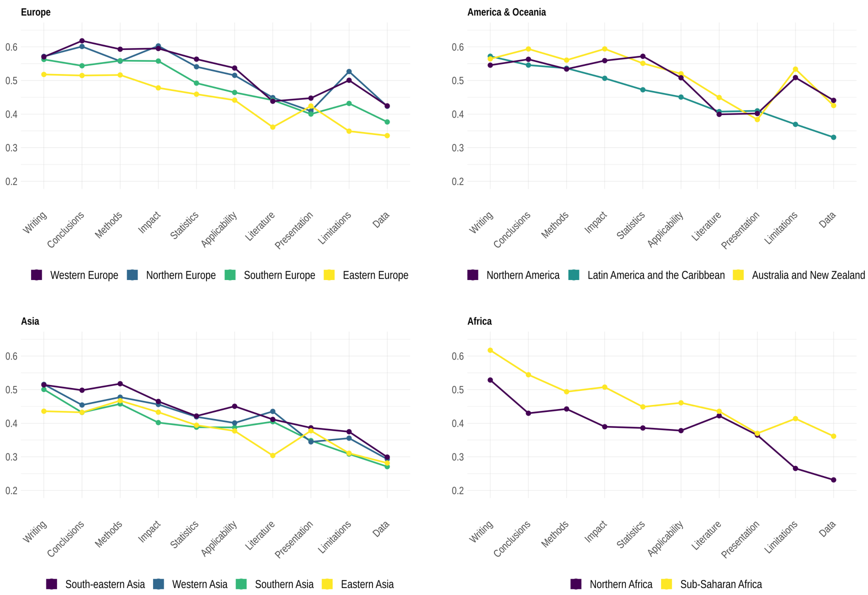

Figure 4 **Median values of each dimension of the developmental score (*i.e.*, cumulative distribution functions F_{D_i} in Materials and Methods) per geographical region.**
Full-size DOI: 10.7717/peerj.13539/fig-4

However, the price to be paid for developmental peer review seems to be a substantial delay in the process, which has always been subject to debate (*Bjrk & Solomon, 2013*). Our results suggest a clear trade-off between developmental peer review and delivery time. This means that, in principle, if reviewers would take more time to deliver their reports generally, this would result in a higher developmental content of reports. However, adding further time to reports in SSE journals would increase the developmental score of reports less than in other areas of research. For instance, we estimated that if reviewers in PS, HMS or LS journals would take ten days more than their median value for report delivery, their expected developmental score would on average increase 3%, thus not reaching the actual median developmental score of SSE reports. Given the recently established fast track options to speed up peer review during the pandemic, it would be interesting to study whether these time pressures have compromised the developmental standards reported here and in which research area (*Horbach, 2021*; *Squazzoni et al., 2021a*).

Our findings indicate that junior scholars are more developmental than more senior reviewers, as are women reviewers in certain fields, such as SSE and HMS, where women reviewers obtained scores slightly higher than men. This would confirm recent findings on relatively weak gender specificities in peer review in various contexts and research areas, including results from linguistic analysis of reports (*Bravo et al., 2019*; *Buljan et al., 2020*; *Squazzoni et al., 2021a*).

We also found evidence that standards reflect geographical and institutional conditions. The report standards are heterogeneous across world regions, while there is an increasing involvement of reviewers from Asian regions compared to less recent data from Publons' state of the art report (*Publons, 2018*). While standards could reflect certain language and cultural specificities and peer review has profound Western historical roots (*Lamont, 2009*), our findings suggest that efforts put forward by publishers and associations regarding higher involvement and inclusion of non-Western academics in journal peer review seem to paid off. However, we must improve on training initiatives and diversity policies to reinforce standards and establish widely shared practices of peer review.

Our findings call for reconsideration of various initiatives on peer review. First, it is important that whenever trying to assess the efficacy of intervention on peer review standards, we use multi-dimensional, context-specific measurements that expand our analysis beyond a few dimensions as in current research, *e.g.*, length of reports (*Publons, 2018*; *Bianchi, Grimaldo & Squazzoni, 2019*). For instance, previous research found that any intervention to improve peer review was relatively unsuccessful in improving the quality of reports (*Jefferson, Wager & Davidoff, 2002*; *Schroter et al., 2004*; *Schroter et al., 2006*; *Bruce, Chauvin & Trinquart, 2016*). Our findings suggest that these results could have been biased by not sufficiently rich, large-scale or systematic measurements of intervention outcomes, or in any case they could have been penalized by lack of appropriate, context-specific benchmarks. While experimental trials are key to assess interventions, measuring peer review only off-line, during specific designed interventions is costly and sometimes limited. Our study suggests that measuring peer review reports more regularly via natural language processing and other machine learning and data science techniques could be a viable alternative to assess internal editorial practices. However, this requires collaboration between publishers, journals and scholars in data sharing initiatives, which are unfortunately only rare (*Squazzoni et al., 2020*).

This given, our study also has certain limitations. First, we used gender guessing techniques, which did not adequately represent non-binary identities, and estimated the seniority of individuals by looking at the number of years since their first record in the Scopus database. However, to the best of our knowledge, there were no better instruments to guess gender and seniority for such a large pool of individuals. Second, our dataset includes only a restricted sample of reports from Elsevier journals in a short time-frame. Although Elsevier does have one of the largest journal portfolios of all publishers, expanding this analysis by including reports from journals from other publishers would be an important step forward. While creating a common database from different publishers is at the moment impossible, due to lack of a data sharing infrastructure solving legal and technical obstacles and creating opportunities for cooperation, a possible extension of our work would be to test our developmental score on available online repositories of peer review reports. Here, considering a longer time-frame could provide a dynamic picture of these standards and not only a cross-sectional comparison.

Furthermore, measuring peer review report standards by looking only at the text of reports separately from the context could provide a rather narrow view of peer review. For instance, each report is linked to others mutually associated with the same manuscript in that the quality of the process has a complex dimension. In this regard, unfortunately we could control neither for the possible effect of peer review guidelines at the journal level nor for the specific effect of varying peer review models adopted by journals. Although recent research suggests that the peer review model does not dramatically change the way reviewers write their reports (*Bravo et al., 2019*; *Buljan et al., 2020*), the fact that journals can vary greatly on the guidelines to their reviewers (*Seeber, 2020*) could be an interesting subject of investigation. Assessing the effect of these internal policies on the developmental content of reports systematically and comparatively would be indeed a major achievement.

Another point is the role of the context. Peer review is performed in a complex, hyper-competitive and hierarchical academic environment, with great variations in terms of areas of research and institutional contexts where competitive pressures and standards of cooperation greatly differ. In our study, we could not control for these confounding factors, including any author-editor-reviewer competitive/cooperative relationships, which could have important implications on the standards of reports (*Bravo et al., 2018*; *Teplitskiy et al., 2018*; *Dondio et al., 2019*).

Furthermore, while developmental peer review is deeply rooted in the institutional tradition of social sciences (*Lamont, 2009*; *Merriman, 2020*), in other areas of research and for specific type of journals, fast editorial decisions and rapid quality screening of manuscripts could be more relevant, regardless of the impact of exogenous factors such as the COVID pandemic. However, even editorial practices and journal guidelines could influence indirectly the development of manuscripts as authors could adapt their manuscript to potential requests and evaluation standards before submitting them to journals. This implies that drawing a straight line between quality screening and developmental function of peer review can be sometimes difficult. As correctly suggested by *Horbach & Halffman (2018)*, peer review is more than review reports and estimating its dimensions and properties calls for a complex set of factors and processes.

With all these caveats, we believe that concentrating on reports, making dimensions and measurements more transparent, identifying context-specific standards is also instrumental to enhance reviewer training initiatives. Given the higher involvement of non-Western regions and their importance in the changing demography of the scientific community, we must expand the traditional target and audiences of training initiatives, increase their diversity and inclusion, and ensure permanent initiatives rather than on-off programs (*Schroter et al., 2004*). Specifying the various functions of peer review and the required multi-dimensional competence, and establishing more informative and standardized journal guideline would also help to reduce the mismatch of expectations and practices (*Köhler et al., 2020*; *Seeber, 2020*).

ACKNOWLEDGEMENTS

We gratefully acknowledge the support on data extraction from the IT staff of Elsevier, specifically Ramsundhar Baskaravelu and his team. This work uses Scopus data provided by Elsevier through ICSR Lab (Elsevier International Center for Study of Research). We also thank Dave Santucci from Elsevier Scopus API team and Kristy James from ICSR for their support on data enrichment about authors and reviewers. We thank Maite Gandia for her preliminary work on analyzing the text of review reports and her help in the first steps of this piece of research. We thank Joan Marsh and Mario Maliki for interesting comments on a preliminary version of the manuscript. Usual caveats apply.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Daniel Garcia-Costa and Francisco Grimaldo are supported by the Spanish Ministry of Science, Innovation and Universities (MCIU), the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF) under project RTI2018-095820-B-I00. Flaminio Squazzoni is supported by a "Department of Excellence" grant from the Italian Ministry of Education, University and Research to the Department of Social and Political Sciences of the University of Milan, a grant from PRIN-MIUR (Progetti di Rilevante Interesse Nazionale –Italian Ministry of University and Research) (Grant Number: 20178TRM3F001 "14All") and a grant from the University of Milan (Grant Number: PSR2015-17 Transition Grant). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Spanish Ministry of Science, Innovation and Universities (MCIU).
The Spanish State Research Agency (AEI).
The European Regional Development Fund (ERDF): RTI2018-095820-B-I00.
Italian Ministry of Education, University and Research to the Department of Social and Political Sciences of the University of Milan.
Progetti di Rilevante Interesse Nazionale–Italian Ministry of University and Research: 20178TRM3F001 "14All".
University of Milan: PSR2015-17.

Competing Interests

Bahar Mehmani is an Elsevier employee. Elsevier provided data for this study through ICSR Lab (Elsevier International Center for Study of Research).

Author Contributions

- Daniel Garcia-Costa conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Flaminio Squazzoni conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Bahar Mehmani conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

- Francisco Grimaldo conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at Harvard Dataverse: Daniel Garcia-Costa,; Flaminio Squazzoni; Bahar Mehmani; Francisco Grimaldo, 2022, "Replication data for: Measuring the developmental function of peer review: A multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals", https://doi.org/10.7910/DVN/D96G2I, Harvard Dataverse, V1.

REFERENCES

Akbaritabar A, Squazzoni F. 2020. Gender patterns of publication in top sociological journals. *Science, Technology & Human Values* In press DOI 10.1177/0162243920941588.

Akbaritabar A, Stephen D, Squazzoni F. 2022. A study of referencing changes in preprint-publication pairs across multiple fields. *Journal of Informetrics* **16(2)**:101258 DOI 10.1016/j.joi.2022.101258.

Bauchner H, Fontanarosa PB, Golub RM. 2020. Editorial evaluation and peer review during a pandemic: how journals maintain standards. *JAMA* **324(5)**:453–454 DOI 10.1001/jama.2020.11764.

Bianchi F, Grimaldo F, Bravo G, Squazzoni F. 2018. The peer review game: an agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics* **116**:14011420 DOI 10.1007/s11192-018-2825-4.

Bianchi F, Grimaldo F, Squazzoni F. 2019. The F3-index. Valuing reviewers for scholarly journals. *Journal of Informetrics* **13(1)**:78–86 DOI 10.1016/j.joi.2018.11.007.

Bjrk B-C, Solomon D. 2013. The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics* **7(4)**:914–923 DOI 10.1016/j.joi.2013.09.001.

Bolek M, Bolek C, Shopovski J, Marolov D. 2022. The consistency of peer-reviewers: assessment of separate parts of the manuscripts vs final recommendations. *Accountability in Research* Epub ahead of print Jan 27 2022 DOI 10.1080/08989621.2022.2030719.

Bornmann L. 2011. Scientific peer review. *Annual Review of Information Science and Technology* **45(1)**:197–245 DOI 10.1002/aris.2011.1440450112.

Bravo G, Farjam M, Grimaldo Moreno F, Birukou A, Squazzoni F. 2018. Hidden connections: Network effects on editorial decisions in four computer science journals. *Journal of Informetrics* **12(1)**:101–112 DOI 10.1016/j.joi.2017.12.002.

Bravo G, Grimaldo F, López-Iñesta E, Mehmani B, Squazzoni F. 2019. The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications* **10(1)**:322 DOI 10.1038/s41467-018-08250-2.

PeerJ

Bruce R, Chauvin A, Trinquart LEA. 2016. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. *BMC Medicine* **14(85)**:1–6 DOI 10.1186/s12916-016-0631-5.

Buljan I, Garcia-Costa D, Grimaldo F, Squazzoni F, Marušić A. 2020. Meta-research: large-scale language analysis of peer review reports. *eLife* **9**:e53249 DOI 10.7554/eLife.53249.

Card D, DellaVigna S. 2013. Nine facts about top journals in economics. *Journal of Economic Literature* **51(1)**:144–61 DOI 10.1257/jel.51.1.144.

Casnici N, Grimaldo F, Gilbert N, Squazzoni F. 2017. Attitudes of referees in a multi-disciplinary journal: An empirical analysis. *Journal of the Association for Information Science and Technology* **68(7)**:1763–1771 DOI 10.1002/asi.23665.

Cowley SJ. 2015. How peer-review constrains cognition: on the frontline in the knowledge sector. *Frontiers in Psychology* **6**:1706 DOI 10.3389/fpsyg.2015.01706.

Deng Q, Hine MJ, Ji S, Sur S. 2017. Building an environmental sustainability dictionary for the IT industry. In: *Hawaii international conference on system sciences 2017*. Honolulu, Hawai: University of Hawai'i at Manoa.

Deng Q, Hine MJ, Ji S, Sur S. 2019. Inside the black box of dictionary building for text analytics: a design science approach. *Journal of International Technology and Information Management* **27**:7.

Dondio P, Casnici N, Grimaldo F, Gilbert N, Squazzoni F. 2019. The "invisible hand" of peer review: the implications of author-referee networks on peer review in a scholarly journal. *Journal of Informetrics* **13(2)**:708–716 DOI 10.1016/j.joi.2019.03.018.

Edwards MA, Siddhartha R. 2017. Academic research in the 21st century: maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science* **34(1)**:51–61 DOI 10.1089/ees.2016.0223.

Ghosal T, Kumar S, Bharti PK, Ekbal A. 2022. Peer review analyze: a novel benchmark resource for computational analysis of peer reviews. *PLOS ONE* **17**:1–29 DOI 10.1371/journal.pone.0259238.

Godbole S, Bhattacharya I, Gupta A, Verma A. 2010. Building re-usable dictionary repositories for real-world text mining. In: *Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10*. New York, NY, USA: Association for Computing Machinery, 1189–1198 DOI 10.1145/1871437.1871588.

Helmer M, Schottdorf M, Neef A, Battaglia D. 2017. Research: gender bias in scholarly peer review. *eLife* **6**:e21718 DOI 10.7554/eLife.21718.

Herber O, Bradbury-Jones C, Boling S, Combes S, Hirt J, Koop Y, Nyhagen E, Veldhuizen J. 2020. What feedback do reviewers give when reviewing qualitative manuscripts? A focused mapping review and synthesis. *BMC Medical Research Methodology* **20**:122 DOI 10.1186/s12874-020-01005-y.

Hochberg ME. 2010. Youth and the tragedy of the reviewer commons. *Ideas in Ecology and Evolution* **3**:8–10 DOI 10.1002/jwmg.763.

Horbach S, Halffman W. 2018. The changing forms and expectations of peer review. *Research Integrity and Peer Review* **3**:8 DOI 10.1186/s41073-018-0051-5.

Horbach SPJM. 2021. No time for that now! Qualitative changes in manuscript peer review during the Covid-19 pandemic. *Research Evaluation* **30(3)**:231–239 DOI 10.1093/reseval/rvaa037.

Huisman J, Smits J. 2017. Duration and quality of the peer review process: the author's perspective. *Scientometrics* **113**:633–650 DOI 10.1007/s11192-017-2310-5.

Jefferson T, Wager E, Davidoff F. 2002. Measuring the quality of editorial peer review. *JAMA* **287(21)**:2786–2790 DOI 10.1001/jama.287.21.2786.

Kharasch ED, Avram MJ, Clark JD, Davidson AJ, Houle TT, Levy JH, London MJ, Sessler DI, Vutskits L. 2021. Peer review matters: research quality and the public trust. *Anesthesiology* **134(1)**:1–6 DOI 10.1097/ALN.0000000000003608.

Köhler T, González-Morales MG, Banks GC, O'Boyle EH, Allen JA, Sinha R, Woo SE, Gulick L. MV. 2020. Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: introducing a competency framework for peer review. *Industrial and Organizational Psychology* **13(1)**:1–27 DOI 10.1017/iop.2019.121.

Lamont M. 2009. *How professors think inside the curious world of academic judgment.* Cambridge, MA: Harvard University Press.

Lewin AY. 2014. The peer-review process: the good, the bad, the ugly, and the extraordinary. *Management and Organization Review* **10(2)**:167–173 DOI 10.1017/S1740877600004095.

Mahmić-Kaknjo M, Utrobičić A, Marušić A. 2021. Motivations for performing scholarly prepublication peer review: a scoping review. *Accountability in Research* **28(5)**:297–329 DOI 10.1080/08989621.2020.1822170.

Merrill E. 2014. Reviewer overload and what can we do about it. *The Journal of Wildlife Management* **78(6)**:961–962 DOI 10.1002/jwmg.763.

Merriman B. 2020. Peer review as an evolving response to organizational constraint: evidence from sociology journals, 19522018. *The American Sociologist* **52**:341–366 DOI 10.1007/s12108-020-09473-x.

Mpouli S, Beigbeder M, Largeron C. 2020. Lexifield: a system for the automatic building of lexicons by semantic expansion of short word lists. *Knowledge and Information Systems* **62**:31813201 DOI 10.1007/s10115-020-01451-6.

Muresan S, Klavans J. 2002. A method for automatically building and evaluating dictionary resources. In: *Proceedings of the third international conference on language resources and evaluation (LREC'02). Las Palmas, Canary Islands - Spain*. Paris: European Language Resources Association (ELRA).

Palayew A, Norgaard O, Safreed-Harmon K, Andersen TH, Rasmussen LN, Lazarus JV. 2020. Pandemic publishing poses a new COVID-19 challenge. *Nature Human Behavior* **4**:666669 DOI 10.1038/s41562-020-0911-0.

Pranić SM, Malički M, Marušić SL, Mehmani B, Marušić A. 2021. Is the quality of reviews reflected in editors' and authors' satisfaction with peer review? A cross-sectional study in 12 journals across four research fields. *Learned Publishing* **34(2)**:187–197 DOI 10.1002/leap.1344.

Publons. 2018. Global state of peer review 2018. Technical report, clarivate analytics. DOI 10.14322/publons.GSPR2018.

Rice DB, Pham B, Presseau J, Tricco AC, Moher D. 2022. Characteristics of 'mega' peer-reviewers. *Research Integrity & Peer Review* **7**:1 DOI 10.1186/s41073-022-00121-1.

Rigby J, Cox D, Julian K. 2018. Journal peer review: a bar or bridge? An analysis of a papers revision history and turnaround time, and the effect on citation. *Scientometrics* **114(4)**:10871105 DOI 10.1007/s11192-017-2630-5.

Santamaría L, Mihaljević H. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* **4**:e156 DOI 10.7717/peerj-cs.156.

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* **328(7441)**:673 DOI 10.1136/bmj.38023.700775.AE.

Schroter S, Tite L, Hutchings A, Black N. 2006. Differences in review quality and recommendations for publication between peer reviewers suggested by authors or by editors. *JAMA* **295(3)**:314–317 DOI 10.1001/jama.295.3.314.

Seeber M. 2020. How do journals of different rank instruct peer reviewers? Reviewer guidelines in the field of management. *Scientometrics* **122**:13871405 DOI 10.1007/s11192-019-03343-1.

Siler K, Lee K, Bero L. 2015. Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences of the United States of America* **112(2)**:360–365 DOI 10.1073/pnas.1418218112.

Squazzoni F, Ahrweiler P, Barros T, Bianchi F, Birukou A, Blom H. JJ, Bravo G, Cowley S, Dignum V, Dondio P, Grimaldo F, Haire L, Hoyt J, Hurst P, Lammey R, MacCallum C, Marušić A, Mehmani B, Murray H, Nicholas D, Pedrazzi G, Puebla I, Rodgers P, Ross-Hellauer T, Seeber M, Shankar K, Van Rossum J, Willis M. 2020. Unlock ways to share data on peer review. *Nature* **578**:512–514 DOI 10.1038/d41586-020-00500-yz.

Squazzoni F, Bravo G, Farjam M, Marusic A, Mehmani B, Willis M, Birukou A, Dondio P, Grimaldo F. 2021a. Peer review and gender bias: a study on 145 scholarly journals. *Science Advances* **7(2)**:abd0299 DOI 10.1126/sciadv.abd0299.

Squazzoni F, Bravo G, Grimaldo F, García-Costa D, Farjam M, Mehmani B. 2021b. Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. a study on 2329 Elsevier journals. *PLOS ONE* **16(10)**:1–17 DOI 10.1371/journal.pone.0257919.

Squazzoni F, Grimaldo F, Marusic A. 2017. Publishing: journals could share peer-review data. *Nature* **546(352)**:352–352 DOI 10.1038/546352a.

Stephen D. 2022. Peer reviewers equally critique theory, method, and writing, with limited effect on the final content of accepted manuscripts. *Scientometrics* DOI 10.1007/s11192-022-04357-y.

Stockemer D. 2022. Introduction: the gendered distribution of authors and reviewers in major European political science journal. *European Political Science* DOI 10.1057/s41304-021-00357-3.

Sullivan P, Trapido E, Acquavella J, Gillum RF, Kirby RS, Kramer MR, Carmichael SL, Frankenfeld CL, Yeung E, Woodyatt C, Baral S. 2022. Editorial priorities and timeliness of editorial assessment and peer review during the COVID-19 pandemic. *Annals of Epidemiology* **69**:24–26 DOI 10.1016/j.annepidem.2022.01.003.

Superchi C, González J, Solá IEA. 2019. Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC Medical Research Methodology* **19**:48 DOI 10.1186/s12874-019-0688-x.

Superchi C, Hren D, Blanco D, Rius R, Recchioni A, Boutron I, González JA. 2020. Development of ARCADIA: a tool for assessing the quality of peer-review reports in biomedical research. *BMJ Open* **10(6)**:e035604 DOI 10.1136/bmjopen-2019-035604.

Teele DL, Thelen K. 2017. Gender in the journals: publication patterns in political science. *PS: Political Science & Politics* **50(2)**:433–447 DOI 10.1017/S1049096516002985.

Teplitskiy M, Acuna D, Elamrani-Raoult A, Krding K, Evans J. 2018. The sociology of scientific validity: how professional networks shape judgement in peer review. *Research Policy* **47(9)**:1825–1841 DOI 10.1016/j.respol.2018.06.014.

Thelwall M. 2022. Journal and disciplinary variations in academic open peer review anonymity, outcomes, and length. *Journal of Librarianship and Information Science* Epub ahead of print Mar 1 2022 DOI 10.1177/09610006221079345.

Van Rooyen S. 2001. The evaluation of peer-review quality. *Learned Publishing* **14(2)**:85–91 DOI 10.1087/095315101300059413.

Van Rooyen S, Black N, Godlee F. 1999. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology* **52(7)**:625–629 DOI 10.1016/S0895-4356(99)00047-5.

Wolfram D, Wang P, Abuzahra F. 2021. An exploration of referees' comments published in open peer review journals: the characteristics of review language and the association between review scrutiny and citations. *Research Evaluation* **30(3)**:314–322 DOI 10.1093/reseval/rvab005.

Apéndice E

Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals

PLOS ONE

RESEARCH ARTICLE

Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals

Flaminio Squazzoni[1]*, Giangiacomo Bravo[2,3], Francisco Grimaldo[4], Daniel García-Costa[4], Mike Farjam[5], Bahar Mehmani[6]

1 Department of Social and Political Sciences, University of Milan, Milan, Italy, **2** Centre for Data Intensive Sciences and Applications, Växjö, Sweden, **3** Department of Social Studies, Växjö, Sweden, **4** Department of Computer Science, University of Valencia, Burjassot, Spain, **5** European Studies, Centre for Languages and Literature, Lund University, Lund, Sweden, **6** STM Journals, Elsevier, Amsterdam, The Netherlands

* flaminio.squazzoni@unimi.it

Citation: Squazzoni F, Bravo G, Grimaldo F, García-Costa D, Farjam M, Mehmani B (2021) Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals. PLoS ONE 16(10): e0257919. https://doi.org/10.1371/journal.pone.0257919

Editor: Alberto Baccini, University of Siena, Italy, ITALY

Received: May 9, 2021

Accepted: September 13, 2021

Published: October 20, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0257919

Data Availability Statement: Data for findings replication are available at this link: https://doi.org/10.7910/DVN/S0T7Z5.

Abstract

During the early months of the COVID-19 pandemic, there was an unusually high submission rate of scholarly articles. Given that most academics were forced to work from home, the competing demands for familial duties may have penalized the scientific productivity of women. To test this hypothesis, we looked at submitted manuscripts and peer review activities for all Elsevier journals between February and May 2018-2020, including data on over 5 million authors and referees. Results showed that during the first wave of the pandemic, women submitted proportionally fewer manuscripts than men. This deficit was especially pronounced among more junior cohorts of women academics. The rate of the peer-review invitation acceptance showed a less pronounced gender pattern with women taking on a greater service responsibility for journals, except for health & medicine, the field where the impact of COVID-19 research has been more prominent. Our findings suggest that the first wave of the pandemic has created potentially cumulative advantages for men.

Introduction

The recent pandemic has spurred a flood of COVID-related research [1, 2]. Over 125,000 COVID-19–related papers were published in the first 10 months after the onset of the pandemic in 2020, of which more than 30,000 hosted by preprint servers [3]. The pandemic even increased the opportunities for publication in completely COVID-unrelated fields, such as ophthalmology [4]. Being a global, systemic challenge affecting nearly all the aspects of society, the pandemic has stimulated research on various health, economic, social and psychological factors [5], thus posing a challenge to journals called to handle an unprecedented volume of submissions at extraordinary speed [6].

However, from the onset of the pandemic, governments in many countries have enforced severe lockdown measures, requiring most academics to work from home. While academics are used to working at a distance and with flexible times, it is plausible that during the

Funding: FS is supported by a "Department of Excellence" grant from the Italian Ministry of Education, University and Research to the Department of Social and Political Sciences of the University of Milan and a Transition Grant from the University of Milan (PSR2015-17). FG and DG are partially supported by the Spanish Ministry of Science, Innovation and Universities (MCIU), the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF) under project RTI2018-095820-B-I00. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: BM is employee of Elsevier and organised the data sharing process.

pandemic competing demands from homeschooling, family obligations and other caring duties have affected the productivity of women and men differently [7, 8]. Indeed, home-schooling and elderly care responsibilities due to COVID-19 lock down regulations have imposed a major shift in family schedules and routines, probably cementing even more traditional gender roles [9, 10]. It has long been known that women drop out more frequently from academia due to difficulties in reconciling work and family life [11–13]. Given that gender inequality in family and work are connected, it is reasonable to hypothesize that the pandemic could have deepened the pre-existing gender inequalities in both realms [14].

For instance, a study in the U.S. showed that women with young children have reduced their working hours four to five times more than fathers during the pandemic [15]. A survey on 4,535 principal investigators in scientific projects in Europe and the U.S. indicated that women academics, those in the 'bench sciences' and, especially, scientists with young children, have experienced a substantial decline in research time [16]. A recent perspective analysis suggested that the effect of the pandemic was worsened by the closure of laboratories and the interruption of most field and observational studies due to restrictions in response to the COVID-19 pandemic, as well as the freeze of intramural research accounts and extra-mural funding sources to support the medical mission [17].

From the early onset of the pandemic, the impression that women were submitting fewer manuscripts to journals was confirmed by two studies using PubMed database and data on preprints to estimate the gender rate of authors posting or publishing COVID-19 related papers during the pandemic [1, 18, 19]. A more recent study on the author byline of 42898 PubMed indexed life science articles found that the percentage of articles on which men versus women were first authors widened by 14 percentage points during the pandemic [20].

However, these findings are still controversial. While a study on American Journal of Public Health confirmed that submissions were higher from men [21], other studies in specific fields reported no trace of gender inequality in the proportion of submissions [22]. For instance, a study on the impact of the pandemic on six journals published by the British Ecological Society (BES) found that the proportion of submissions authored by women during the COVID period of 2020 did not change relative to the same period in 2019 [23].

Unfortunately, these studies either considered only preprints or publications, without access to data to examine submissions to journals, or lacked cross-journal data in various fields, thus limiting evidence to only specific cases. Understanding whether the COVID-19 race for publications has possibly disproportionately benefited men requires accessing full individual data from various journals in a comparable time frame before and during the pandemic, so as to estimate the effect of the pandemic on individual scholars. Although research on preprints is important to estimate the academic response to the greater demand for research during the pandemic [1, 19], looking at manuscript submissions and peer review activities for journals in different research areas before and during the pandemic is key to estimate gendered gaps in time and effort investment for research by academics more precisely.

To fill this gap, we have established a confidential agreement with Elsevier publishing to access manuscript and peer review metadata from all their journals. These included individual records of authors and referees in a fully comparable monthly time frame—i.e., February-May 2018–2020, during the first wave of the pandemic in Asia, Europe and America (see Methods Section). Note that focusing on the early months of the pandemic was instrumental to estimate gender inequalities as most countries enforced similar lockdown measures, which were eventually eased during summer.

Given that our data came from manuscript submission systems, we had to re-purpose them for research by adding gender guessing algorithms and mobility data from Google to control for residential data, and completing them with Scopus data to estimate scholar's age and seniority. This allowed us to treat the pandemic as a 'quasi-experiment' and estimate its effect on academics' productivity at an individual level, by considering the seasonal rate of submissions in 2018–2020 in different research areas and residential countries. Unlike other studies, which looked at the gender proportion either of preprint or publication authors or submission authors [23], we examined the effect of the pandemic on each scholar active between 2018–2020 in these journal submission databases at the individual level. Furthermore, we included data on referees to understand whether women were penalized in their capacity to serve the community and influence the type of research performed during the pandemic.

Materials and methods

The dataset

Our dataset included complete information on manuscripts and reviews from 2329 Elsevier journals from January 2018 to May 2020 (see Table 1; S1 Table includes the total number of submissions in Feb-May 2018–2020). The sample included about 5 million academics listed as authors and/or referees. Data access required a confidential agreement to be signed on 12th May 2020 between Elsevier and each author of this study. The agreement was inspired by the PEERE protocol for data sharing and included anonymization, privacy, data management and security policies jointly determined by all partners [24].

For the sake of our analysis, we concentrated on the first wave of the COVID-19 pandemic, i.e., from February to May 2020 (more precisely, weeks 6–22, 2020). This allowed us to cover the large part of the outbreak during the first half of 2020, including the effect of restrictions on mobility in China and Asia in Feb 2020 and in Europe and United States later. Furthermore, Google COVID-19 Community Mobility Report used the first five weeks of 2020 as reference so that mobility data were only available starting from week 6, 2020 (see https://www.google.com/covid19/mobility/; accessed on 30 June 2020). On the other hand, few countries had any lockdown measures in place during January 2020. To ensure full comparability across years, including seasonality issues, we decided to limit our observations to the corresponding months of 2018 and 2019.

We used the e-mail (or the set of e-mails) associated to each user account in the underlying submission systems (i.e., Editorial Manager, Elsevier Editorial System and EVISE) to track academics across all journals and constructed an auto-generated anonymous unique identifier. We controlled for multiple e-mail addresses and this allowed us to circumvent the incompleteness of other alternative identifiers, which were either available only for a partial sub-sample of academics (e.g., ORCID) or not unique (e.g., ScopusID). Note that the same individuals may have been counted twice (or more) in the data reported in our analysis whenever submitting or reviewing to journals in different research areas.

Table 1. Overview of the main variables considered in the analysis by area of research.

	Health & Medicine	Life Sciences	Physical Sciences & Engineering	Social Sciences & Economics	Total
N. of journals	885	416	767	261	2329
Submissions (female)	1005590	653729	991304	128798	2779421
Submissions (male)	1816621	1063178	2967128	271821	6118748
Accepted reviews (female)	133989	104065	194062	43319	475435
Accepted reviews (male)	359600	239918	888022	104621	1592161
Declined reviews (female)	233484	211185	336275	53476	834420
Declined reviews (male)	527723	441786	1338245	109969	2417723

https://doi.org/10.1371/journal.pone.0257919.t001

To prevent de-anonymization of authors and referees, all submissions from countries with less than 20 authors/referees or with a number of authors that happens 5 times or less for the same journal were dropped from the dataset. This reduced our sample by 290082 submissions, i.e., about 6% of the observations. In addition to solving the privacy issue mentioned above, by removing observations from smaller countries we increased the robustness of the analysis, as the maximum likelihood estimation of random intercepts with few observations for each category may have caused convergence and over-fitting problems, thereby making it difficult to control possible statistical biases. Finally, these countries were also not covered by the Google COVID-19 Community Mobility Report and so should have been excluded in any case.

Gender guessing

Our procedure for gender guessing was based on a two-step disambiguation algorithm inspired by previous research [25–28] and already validated on several datasets of academics' names [29]. First, we queried the Python package gender-guesser about the first names and countries of origin, if any. Gender-guesser allowed us to minimize gender bias and achieve the lowest mis-classification rate (less than 3% for Benchmark 1 in [29]). For names classified by gender-guesser as 'mostly_male', 'mostly_female', 'andy' (androgynous) or 'unknown' (name not found), we used GenderAPI (see https://gender-api.com/), which ensures that the level of mis-classification is around 5% (see Table 4 in [29]) and has the highest coverage on multiple name origins (see Table 5 in [29]). This procedure allowed us to guess the gender of 94.5% of academics in our sample, 45.1% coming from gender-guesser and 49.2% from GenderAPI. The remaining 5.5% of academics were assigned an unknown gender. Note that this level of gender guessing is consistent with the non-classification rate for names of academics in previous research [29]. Note also that while we were aware that any gender binary definition did not adequately represent non-binary identities, to the best of our knowledge, there was no better instrument to estimate gender for such a large pool of individuals.

We checked the robustness of the analysis to variations of the gender guessing algorithm by estimating further models using a more restrictive version of the algorithms, which kept the rate of miss-classified names resolved by GenderAPI under 5% and required a minimum of 62 samples with at least 57% accuracy (see S9 and S10 Tables where the percentage of academics without a guessed gender increased to 28.5%).

Scholar's age

Scholars' age was estimated by using the number of years since their first record in the Scopus database. We followed a conservative rule and authors were identified by their Scopus IDs, e-mail addresses or the full name and country (in case a single profile was found). Authors without a profile in Scopus or not being uniquely identifiable were excluded from the analysis, whenever using age as a variable. Note that our aim here was not to estimate the age of each scholar precisely, which is impossible. We wanted to identify cohorts of academics to estimate the ones which most probably could have homeschooling and elderly care responsibilities. We assumed that first publications would correspond to the period in which academics were completing their MD or PhD period (i.e., estimated age around 25/30). We used this assumption to create the two cohorts mentioned in the text. The fact that our estimations could have mis-classified the actual age of scholars by some years (e.g., estimating someone being 40 instead of 43 years old in 2020) is irrelevant to the purpose of our study. Note that our classification could have underestimated the age of some authors who did not have any past formal training (e.g., a PhD title) and published their first paper only recently. While it is impossible to identify these cases in the database, our analysis using self-declared academic titles in Elsevier data could be seen as a supplementary check on these cases, as they could be ideally listed themselves as Mr. and Ms. etc. and so being controlled for in our analysis.

Indeed, for a robustness check, we used the self-declared academic title and degree in the Elsevier dataset. Note that the use of the title "Dr." could be different in certain communities and perhaps not allowing to clearly identify someone with a PhD title. On the other hand, the title "Prof." could be used more rarely among academic faculty members working in hospitals. However, the size of the sample and the large coverage of academics from different countries and areas of research could have reduced the effect of this possible bias on our outcomes.

COVID-19 related manuscripts

Elsevier data allowed us to distinguish COVID related and non-related manuscripts through an internal Boolean flag from the manuscript submission systems used by journals. A manuscript was considered COVID-19 related when the following condition was met by its keywords or abstract: ["covid-19" OR "covid 19" OR "covid19" OR "corona virus" OR "coronavirus" OR "corona-virus" OR "corona viruses" OR "coronaviruses" OR "corona-viruses" OR "orthocoronavirinae" OR "coronaviridae" OR "coronavirinae" OR "2019-ncov" OR "2019ncov" OR "2019 ncov" OR "hcov-19" OR "sars-cov" OR "sars cov" OR "severe acute respiratory syndrome" OR "sars-cov-2" OR "sars-cov2" OR "mers-cov" OR "mers cov" OR "middle east respiratory syndrome" OR "middle eastern respiratory syndrome" OR ("angiotensin-converting enzyme 2" AND "virus") OR ("ace2" AND "virus") OR "soluble ace2" OR ("angiotensin converting enzyme2" AND "virus") OR ("ards" AND "virus") OR "acute respiratory distress syndrome" OR ("sars" AND "virus") OR ("mers" AND "virus") OR ("wuhan" AND "virus")]. We used this taxonomy to track COVID-related manuscripts (i.e., manuscripts focusing on diseases caused by the the same family of viruses) before the start of the pandemic.

Data analysis

All analyses were performed using the *R* platform [30]. The statistical analysis was performed exploiting the high-performance computing facility of the Linnaeus University Centre for Data Intensive Sciences and Applications. If not explicitly otherwise mentioned in the text, standard test to check the model assumptions (homogeneity, normality of random effect, etc.) were performed for all models.

Results

The COVID-19 pandemic has caused an abnormal rate of journal submissions. Our data indicate that the number of manuscripts submitted to all Elsevier journals between February to May 2020 increased by 30% compared to the same period of the previous year (i.e., from 620,685 for February-May 2019 to 807,449 in 2020; the growth rate of submissions from 2018 to 2019 was 11%) (see S1 Table). Note that in health & medicine journals, this trend was even stronger with an increase of 63% (i.e., from 147,401 submissions from February-May 2019 to 240,587 in 2020). At the same time, the absolute numbers of accepted review invitations for all disciplines increased by 29%, from 1,847,256 in 2019 to 2,381,284 in 2020. In the case of health & medicine journals, the accepted invitations increased by 34% from 2019 to 2020 (i.e., 415,033 in 2019 against 554,895 in 2020) compared with an increase of 63% submissions.

Our analysis shows that while the number of manuscripts submitted to journals generally increased during the first wave of the pandemic, the number of manuscripts submitted by men was higher than those submitted by women (Fig 1A). The rate of accepted review invitations—i.e., the number of accepted invitations on the total number of invitations sent to potential

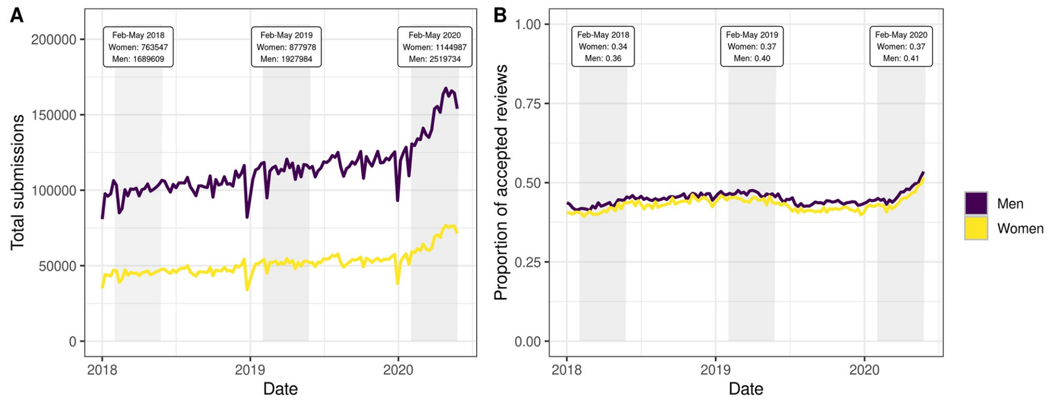

Fig 1. Total submissions (A) and proportion of accepted reviews (B) per week across the whole period covered by the dataset. The shaded areas indicate the February-May period of each year considered in the analysis. Note that in panel A co-authored submissions were reported multiple times depending on the number of co-authors. Each author or referee whose gender was not successfully guessed by our algorithm was excluded.

https://doi.org/10.1371/journal.pone.0257919.g001

referees—has been more constant around an average of ≈ 40% with women accepting slightly fewer invitations than men (Fig 1B).

In February-May 2018, 2019 and 2020, women submitted 2,779,421 manuscripts against 6,118,748 manuscripts submitted by men. Women agreed on performing 475,435 reviews while declining 834,420 invitations, with a proportion of 37% accepted invitations. Men accepted 1,592,161 review invitations while declining 2,417,723, with a similar acceptance proportion (40%) (see Table 1 for a summary of these descriptive statistics per research area).

The effect of the pandemic on submissions

We calculated a submission difference index for each author (Δ_S) as the number of new submissions in February-May 2020 minus the average number of submissions from the same author in the corresponding months of 2018 and 2019. We then estimated each scholar's age by using the number of years since their first record in the Scopus database. Given that students tend to complete their MD-PhD title when 25–30 years old [31, 32], we divided the sample in two age cohorts ($\leq$ or $>$ 20 years after receiving their title), and hypothesized that more junior cohorts of women would be most likely affected by homeschooling and elderly care responsibilities.

Results showed that the overall increase of submissions in 2020 led most authors to $\Delta_S \geq 0$. However, when considering differences in age and areas of research, we found that the Δ_S of men increased more than that of women, especially those in the more junior cohort mentioned above (Fig 2). This would suggest that women had at least comparatively fewer opportunities for research during the first wave of the pandemic.

To check for the significance of these effects, we estimated a mixed effects model using authors' gender and age to predict Δ_S (Table 2). In order to control for the fact that authors were based in countries with different university systems and contagion-prevention policies, we included random effects for countries in the model as a geographical control. Results indicated a statistically significant negative effect for women in all areas of research (Table 2). In addition, we found a consistent positive interaction effect between gender and age, with more senior cohorts of women less penalized than younger scholars.

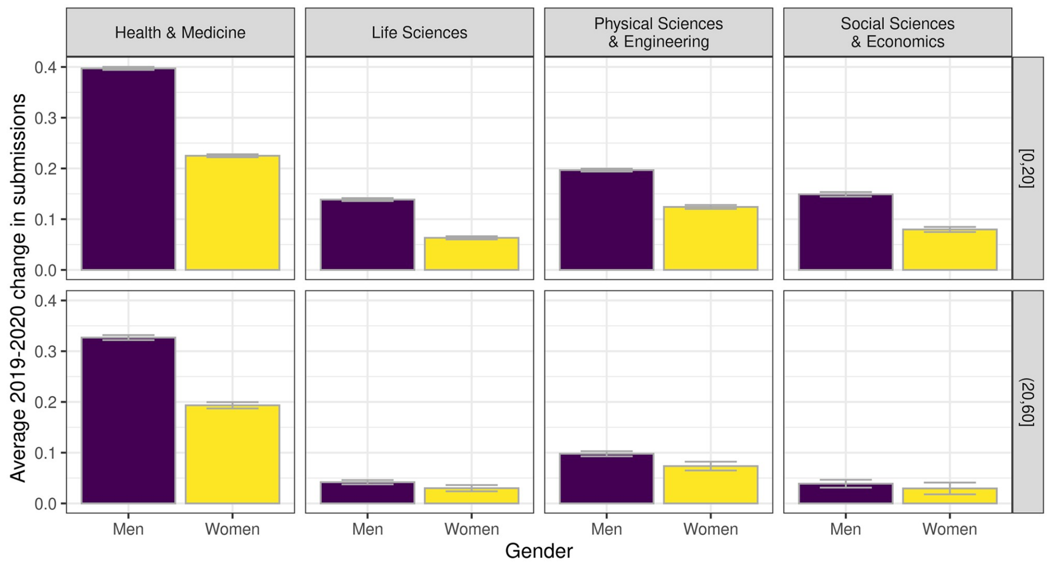

Fig 2. Average change in submissions by research area and age, the latter variable including authors in the first cohort (≤ 20 years from their first publication) in the first group with older authors in the second. Bars represent standard errors.

https://doi.org/10.1371/journal.pone.0257919.g002

As a robustness check of our age measurement, we estimated similar models using a measure of seniority based on the author's title (i.e., no title, Doctor or Professor) as recorded in Elsevier's database. Results confirmed previous findings. The deficit of women was pronounced especially in health & medicine, with a weakly significant effect in social sciences &

Table 2. Mixed effects models predicting February-May 2020 changes in the number of submissions per area of research.

	Health & Medicine	Life Sciences	Physical Sciences & Engineering	Social Sciences & Economics
Women	−0.164	−0.078	−0.097	−0.077
	(0.007)	(0.007)	(0.008)	(0.011)
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
Age	−0.001	−0.002	−0.003	−0.004
	(0.0002)	(0.0002)	(0.0002)	(0.0004)
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
Women×Age	0.001	0.002	0.002	0.002
	(0.0004)	(0.0004)	(0.001)	(0.001)
	$p = 0.022$	$p < 0.001$	$p < 0.001$	$p = 0.003$
Intercept	0.329	0.138	0.209	0.185
	(0.020)	(0.015)	(0.018)	(0.014)
	$p <0.001$	$p <0.001$	$p <0.001$	$p <0.001$
Observations	706126	480240	856454	152348
Log Likelihood	−1369462	−818103	−1816437	−245405

The baseline is represented by the average of corresponding months in 2018 and 2019. Random intercepts included for countries.

https://doi.org/10.1371/journal.pone.0257919.t002

Table 3. Mixed effects models predicting February-May 2020 changes in the number of submissions per area of research.

	Health & Medicine	Life Sciences	Physical Sciences & Engineering	Social Sciences & Economics
Women	−0.080	−0.0003	−0.016	−0.028
	(0.008)	(0.008)	(0.008)	(0.012)
	p <0.001	p = 0.973	p = 0.055	p = 0.016
Doctor	0.089	0.009	0.058	0.026
	(0.007)	(0.006)	(0.006)	(0.009)
	p <0.001	p = 0.126	p <0.001	p = 0.004
Professor	0.163	0.017	0.091	0.013
	(0.008)	(0.007)	(0.007)	(0.010)
	p <0.001	p = 0.020	p <0.001	p = 0.217
Women×Doctor	−0.027	−0.021	−0.016	0.005
	(0.009)	(0.009)	(0.010)	(0.014)
	p = 0.005	p = 0.022	p = 0.115	p = 0.708
Women×Professor	−0.065	−0.010	0.001	0.011
	(0.013)	(0.012)	(0.014)	(0.018)
	p <0.001	p = 0.405	p = 0.971	p = 0.552
Intercept	0.314	0.172	0.182	0.188
	(0.019)	(0.016)	(0.017)	(0.015)
	p <0.001	p <0.001	p <0.001	p <0.001
Observations	847892	571051	1026221	195157
Log Likelihood	−1635935	−984773	−2180291	−314846

The baseline is represented by the average of corresponding months in 2018 and 2019. Models include as predictor the title of each author with no title used as reference category. Random intercepts included for countries.

https://doi.org/10.1371/journal.pone.0257919.t003

economics and physical sciences (Table 3). Note that we considered any value of $0.05 < p < 0.005$ as being only weakly significant [33].

In order to test the hypothesis that these gender and age differences were a side-effect of the different anti-contagious measures adopted by various countries, we included in the model a proxy of how lockdown and social distancing measures, such as the closure of schools, could have affected academics in different countries. Following recent geographical research on the effect of contagion-prevention measures [34–36], we used Google's COVID-19 Community Mobility Report (see details in the supplementary materials), which tracks the amount of time spent by mobile-phone users in different places, including residential areas. Mobility reports are available at the country level (in some cases even at sub-national level) and are summarised in an index that calculates the different time rates spent by individuals in residential areas in a given day compared to the median value of January 2020.

We calculated the average values of the February-May 2020 period of the residential area index per country (see the map in S1 Fig) to control for the exposure of each scholar to the same mobility restrictions and lockdown measures. Unfortunately, certain countries (e.g., China and Iran) were not included in the mobility reports and so our analysis was performed on a restricted sample of academics (this caused a reduction of our observations from 16% to 32% depending on the area of research; see Table 4).

Results indicated a negative interaction between gender and time in residential areas when considering authors submitting manuscripts to health & medicine, and physical science & engineering journals. In addition, we found a significant or weakly significant and negative pure effect of gender in all areas of research (Table 4).

Table 4. Mixed effects models predicting February-May 2020 changes in the number of submissions per area of research.

	Health & Medicine	Life Sciences	Physical Sciences & Engineering	Social Sciences & Economics
Women	−0.056	−0.058	−0.052	−0.053
	(0.016)	(0.016)	(0.020)	(0.026)
	p = 0.001	p < 0.001	p = 0.010	p = 0.041
Age	−0.002	−0.002	−0.003	−0.003
	(0.0002)	(0.0002)	(0.0003)	(0.0004)
	p < 0.001	p < 0.001	p < 0.001	p = <0.001
Residential	0.018	0.005	0.003	0.003
	(0.004)	(0.003)	(0.004)	(0.003)
	p = <0.001	p = 0.147	p = 0.522	p = 0.279
Women×Age	0.002	0.002	0.002	0.002
	(0.0004)	(0.0004)	(0.001)	(0.001)
	p < 0.001	p < 0.001	p < 0.001	p = 0.008
Women×Residential	−0.010	−0.002	−0.004	−0.001
	(0.001)	(0.001)	(0.001)	(0.002)
	p < 0.001	p = 0.074	p = 0.005	p = 0.490
Intercept	0.096	0.090	0.181	0.138
	(0.052)	(0.044)	(0.053)	(0.041)
	p = 0.067	p = 0.041	p = 0.001	p = 0.001
Observations	587184	356988	577890	127263
Log Likelihood	−1098897.000	−581594.400	−1175353.000	−200281.500

The baseline is represented by the average of corresponding months in 2018 and 2019. Models include time in residential areas from Google's COVID-19 Community Mobility Report (see https://www.google.com/covid19/mobility/; accessed on 30 June 2020). Random intercepts included for countries.

https://doi.org/10.1371/journal.pone.0257919.t004

We then concentrated on 'COVID-related' manuscripts, i.e., manuscripts focusing on diseases caused by viruses of the *Coronaviridae* family (see Materials and methods). By using keywords similarity and internal classifications from Elsevier, we reconstructed the time trends of 'COVID-related' manuscripts submitted by academics to all Elsevier journals in the same period in 2018–2020, e.g., research on SARS-CoV-1. This also allowed us to focus on whether women doing research more directly relevant to COVID-19 were penalised during the pandemic.

Results confirmed that women submitted fewer COVID-19 related manuscripts in 2020 in health & medicine journals (Table 5). Note that we found non-significant or weakly-significant coefficients in other areas of research because of the relatively lower number of COVID-19 related manuscripts submitted to these journals (see S3 Table).

Table 5. Mixed effects models predicting February-May 2020 changes in the number of submissions of COVID-related manuscripts in health & medicine journals.

	Estimate	Std. Error	*t*	*p*
Intercept	1.421	0.037	38.007	<0.001
Women	-0.133	0.027	-4.967	<0.001
Age	0.003	0.001	3.304	0.001
Women×Age	-0.001	0.002	-0.857	0.392
Observations	51916			
Log Likelihood	-99295			

Random intercepts included for countries.

https://doi.org/10.1371/journal.pone.0257919.t005

We also performed similar analyses on other specific subsets of manuscripts or journals. For instance, we considered the type of submissions by concentrating only on manuscripts indexed as "research papers" (see S4 Table), on submissions to first quartile (Q1) journals in the 2020 Journal Citation Reports (see S2 and S5 Tables), on the gender of first authors (see S6 Table), and on manuscripts with only one author (see S7 Table). Results of these analyses were fully consistent with our main finding: women academics submitted fewer manuscripts, both research and non-research manuscripts (e.g., commentaries), either in Q1 or in journals with a lower impact factor, and also fewer manuscripts as first authors. This effect was significant for more junior cohorts of women except for manuscripts submitted as first authors, where women were penalized regardless of age. In the case of single-authored manuscripts, the intersection of gender and age was significant only for submissions to health & medicine journals, with gender having a negative effect on submissions in all research areas except in physical science journals.

The effect of the pandemic on academics' commitment to peer review

In order to measure the gender effect of the COVID-19 pandemic on academics' commitment to peer review, we calculated the proportion of review invitations accepted Δ_R for each invited referee in February-May 2020 compared to the corresponding period in 2019. This proportion excluded individuals who did not receive any invitations in February-May 2019 and 2020. To minimize missing values, we excluded the 2018 sample and restricted our analysis to February-May 2019 and 2020.

Besides an overall decline in the number of accepted invitations per individual, our results showed that the pandemic generally did not determine considerable gender differences by either research areas or scholar's age, the sole exceptions being peer review in health & medicine and physical science journals and the case of more senior referees (see Fig 3).

We then estimated two mixed-effect models per area of research, controlling for the time spent in residential areas. Results confirmed that the relative decline in Δ_R was more pronounced only for women in health & medicine (Table 6), although even in this case, we did not find any significant interaction with the time spent in residential areas (see S8 Table).

Discussion

The COVID-19 pandemic has generated unforeseen opportunities for research as a collective response of the academic community to the pandemic [1, 8]. While this will continue over the following months due to various factors, including the challenge of possible mutations of the virus, and the global, societal implications of the pandemic, the exceptional lockdown and social distancing measures introduced since the first wave of the pandemic in early 2020 could have created inequalities in this academic race due to the competing demands for homeschooling and other family care duties [9, 16, 37].

Our complete data on all Elsevier journals indicate that women submitted fewer manuscripts than men during the first wave of the pandemic in early 2020. This has been especially prominent in the research area where the academic production has been higher during the pandemic, i.e., health & medicine. This suggests that the pandemic could have exacerbated existing inequalities by imposing additional obstacles in terms of time and effort investment by women just as the demand for research was growing unprecedentedly.

Our findings suggest that more junior cohorts of women academics were penalized the most. If we consider our control for residential mobility, this could be possibly explained by a major shift in family schedules and routines caused by the pandemic due to interference of homeschooling and more intense family duties, which could have seen these cohorts of

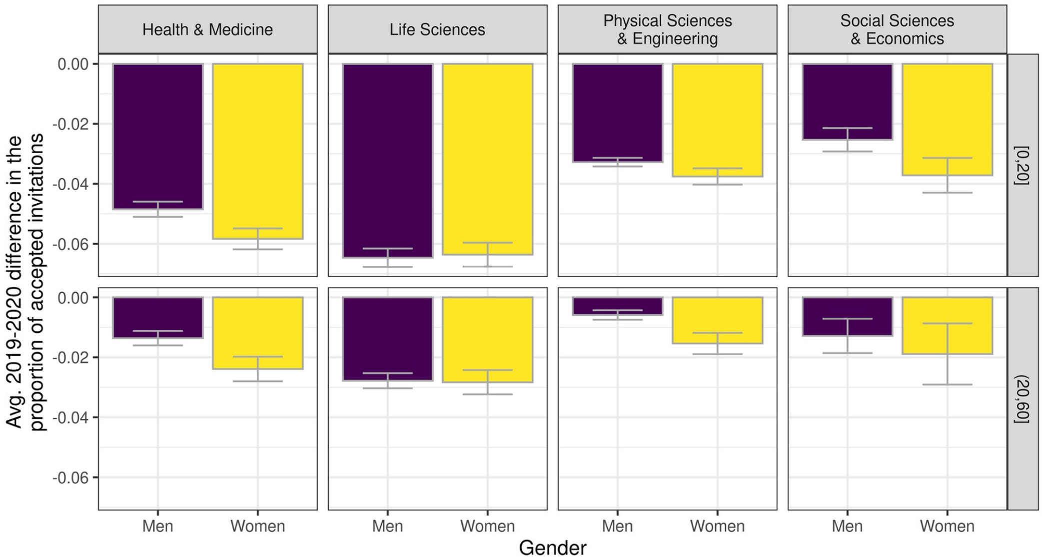

Fig 3. Average difference in the proportion of accepted invitations by areas of research and age, the latter variable including authors in the more junior cohort (≤ from their first publication) in the first group with more senior authors in the second. Bars represent standard errors.

https://doi.org/10.1371/journal.pone.0257919.g003

women on the front-line [15]. Note that these cohorts would probably include women without permanent academic positions, competing for tenure, promotion and grants.

While pressures on peer review are higher in this period, requiring also special arrangements by many journals—e.g., special fast-tracks—, our findings suggest a general decline of accepted invitations but without pronounced gender effects. On the one hand, our findings

Table 6. Mixed effects models predicting February-May 2020 changes in the proportion of accepted review invitations per area of research.

	Health & Medicine	Life Sciences	Physical Sciences & Engineering	Social Sciences & Economics
Women	−0.016	−0.001	−0.005	−0.016
	(0.007)	(0.008)	(0.005)	(0.012)
	p = 0.025	p = 0.914	p = 0.319	p = 0.173
Age	0.002	0.002	0.001	0.001
	(0.0002)	(0.0002)	(0.0001)	(0.0003)
	p < 0.001	p < 0.001	p < 0.001	p = 0.005
Women×Age	0.0003	0.0001	0.00002	0.0005
	(0.0003)	(0.0003)	(0.0002)	(0.001)
	p = 0.277	p = 0.648	p = 0.944	p = 0.500
Intercept	−0.068	−0.086	−0.052	−0.037
	(0.005)	(0.005)	(0.003)	(0.007)
	p < 0.001	p < 0.001	p < 0.001	p < 0.001
Observations	90902	78491	206426	29287
Log Likelihood	−55689	−49400	−123204	−19882

The baseline is represented by the average of the corresponding months in 2019. Random intercepts included for countries.

https://doi.org/10.1371/journal.pone.0257919.t006

indicate that women have taken on a greater service responsibility for journals and the community as referees at least comparatively comparable to men. At the same time, men have submitted more manuscripts, thus benefiting from the involvement of women as referees. On the other hand, women were less involved in peer review for health & medicine journals, the field where the impact of COVID-19 research has been more prominent. This would suggest that they were less capable of influencing the type of research that was published. This raises concern over the quality of peer review under increasing editorial pressures during the pandemic, which would require an entire follow up study [38].

This said, our study has certain limitations. Though we achieved an observation scale never achieved previously in this type of research, our sample was limited to only Elsevier journals. However, Elsevier does have one of the largest journal portfolios of all publishers, sufficiently covers all areas of research, and represents a balanced proportion of journals across research areas (see S2 Table). While a desirable extension would be to expand this analysis by including journals from other publishers, we must acknowledge that creating a common database with full data on submissions from different publishers is at the moment impossible due to lack of a data sharing infrastructure solving legal and technical obstacles and creating opportunities for cooperation [39].

Finally, as mentioned above, unfortunately Google mobility data were not available in certain countries and regions, e.g., China and Iran. Therefore, we could not include our lockdown proxy (extra time spent in residential areas) for all observations in the sample. This suggests to consider our models including mobility data more as a robustness check for our analysis. Note, however, that any other possible measurements of actual lockdown of our sampled academics, such as country-based dates when these measures were introduced, were intrinsically biased because individuals could anticipate these announcements by staying at home before their introduction and/or even after the specific dates when restrictions are removed.

Given that many submissions during the pandemic will eventually turn into publications and citations, and considering the importance of these latter for academic career and prestige, it is probable that the first wave of the pandemic that we have examined here could be seen as the *genealogy* of gender disparities that will have important short- and longer-term effects. Pandemics have always exacerbated existing inequalities [40]. Indeed, those who have already benefited from this COVID-19 research race may have better chances in the near future to receive prestigious grants and obtain tenures and promotion in prestigious institutions. Previous research on peer review and editorial processes at journals has shown that gender inequalities in the rate of submissions to journals is key to determine inequality of publications and recognition [41].

In conclusion, it is important that funding agencies and hiring and promotion committees at national and international levels reconsider their policies in these exceptional times. While voluntary disclosure of gender or gender quotas during journal submissions could lead to further biases [19], flagging, carefully pondering or even disregarding COVID-19 related publications and citations from applicants' assessment could be considered. Following the example of the Canadian Institutes of Health Research (CIHR), extending deadlines and supporting COVIGiven that the use of bibliometric indicators to assess applicants for funds and academic positions has been strongly criticized even in normal times [42], one of the most important lessons from the pandemic could be to follow multi-dimensional criteria in any academic assessment. This could include a COVID-19 impact statement where any candidate is required to explain the opportunities and constraints faced during the pandemic [43].

At the same time, improving career enhancement and retention by appropriate institutional interventions, such as promoting a more diverse, inclusive, and equitable working environment and embracing a family-friendly leadership policy in the management of labs and institutes, could help moderate the distortions caused by the pandemic [44]. These interventions could transform the pandemic in an unprecedented opportunity to reset certain established practices and reconsider how funders, institutes and universities could offer better support to academics who are more vulnerable to the effect of global crisis [45].

In this context, journals and publishers should increase their usual effort in internal assessment and monitoring with a special focus on the consequences of the pandemic on research [23, 46]. This study has paved the way for large-scale collaboration initiatives on data sharing between publishers and the scientific community [39] and could be used as a template to map the evolution of the pandemic science.

Supporting information

S1 Fig. Average increase in the time spent in residential areas by country. The change was calculated as different rate from the baseline given by median value during the first five weeks of 2020. Data from Google COVID-19 Community Mobility Report (see https://www.google.com/covid19/mobility/; accessed on 30 June 2020). White areas indicate missing data.
(TIF)

S1 Table. Total number of new submissions, review invitations, and accepted invitations per area of research in February-May 2020 and corresponding months of 2018 and 2019. Note that data reported here differ from those in Table 1 because: (i) several authors could have submitted the same manuscript to different journals, which was only counted once here; and (ii) submissions and reviews from academics whose gender was not guessed by our algorithm were included here but not in Table 1.
(PDF)

S2 Table. Proportion (%) of journals included in each quartile of the impact factor distribution by area of research. The quartiles were calculated using Journal Citation Reportsby Clarivate Analytics.
(PDF)

S3 Table. Mixed effects models predicting February-May 2020 changes in the number of submissions of Covid-related manuscripts per area of research. Random intercepts included for countries.
(PDF)

S4 Table. Mixed effects models predicting February-May 2020 changes in the number of submissions of research papers. The baseline is represented by the average of corresponding months in 2018 and 2019. Random intercepts included for countries.
(PDF)

S5 Table. Mixed effects models predicting February-May 2020 changes in the number of submissions of manuscripts submitted to Q1 journals. The baseline is represented by the average of corresponding months in 2018 and 2019. Random intercepts included for countries.
(PDF)

S6 Table. Mixed effects models predicting February-May 2020 changes in the number of submissions by first authors. The baseline is represented by the average of corresponding months in 2018 and 2019. Random intercepts included for countries.
(PDF)

S7 Table. Mixed effects models predicting February-May 2020 changes in the number of submissions by solo authors. The baseline is represented by the average of corresponding months in 2018 and 2019. Random intercepts included for countries.
(PDF)

S8 Table. Mixed effects models predicting February-May 2020 changes in the proportion of accepted review invitations per area of research. The baseline is represented by the average of the corresponding months in 2019. Models included time in residential areas from Google's COVID-19 Community Mobility Report (see https://www.google.com/covid19/mobility/; accessed on 30 June 2020). Random intercepts included for countries.
(PDF)

S9 Table. Mixed effects models predicting February-May 2020 changes in the number of submissions per area of research area. The baseline is represented by the average of corresponding months in 2018 and 2019. Random intercepts included for countries. Gender data based on the stricter version of the gender guessing algorithm.
(PDF)

S10 Table. Mixed effects models predicting February-May 2020 changes in the proportion of accepted review invitations per area of research. The baseline is represented by the average of the corresponding months in 2019. Random intercepts included for countries. Gender data based on the stricter version of the gender guessing algorithm.
(PDF)

Acknowledgments

We gratefully acknowledge the support on data extraction from the IT staff of Elsevier, specifically Ramsundhar Baskaravelu and his team. We also thank Dave Santucci from Elsevier Scopus API team and Kristy James from Elsevier International Center for Study of Research (ICSR) for their support on data enrichment about authors and reviewers. The statistical analysis was performed exploiting the high-performance computing facilities of the Linnaeus University Centre for Data Intensive Sciences and Applications.

Author Contributions

Conceptualization: Flaminio Squazzoni, Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa, Mike Farjam, Bahar Mehmani.

Data curation: Francisco Grimaldo, Daniel García-Costa, Bahar Mehmani.

Formal analysis: Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa, Mike Farjam.

Investigation: Flaminio Squazzoni.

Methodology: Flaminio Squazzoni, Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa.

Project administration: Flaminio Squazzoni, Bahar Mehmani.

Supervision: Flaminio Squazzoni, Francisco Grimaldo.

Validation: Francisco Grimaldo, Daniel García-Costa.

Visualization: Francisco Grimaldo, Daniel García-Costa.

Writing – original draft: Flaminio Squazzoni, Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa, Mike Farjam, Bahar Mehmani.

Writing – review & editing: Flaminio Squazzoni, Giangiacomo Bravo, Francisco Grimaldo, Daniel García-Costa, Mike Farjam, Bahar Mehmani.

References

1. Vincent-Lamarre P, Sugimoto CR, Larivière V. The decline of women's research production during the coronavirus pandemic. Nature Index. 2020;May 19.
2. Nowakowska J, Sobocińska J, Lewicki M, Żaneta Lemańska, Rzymski P. When science goes viral: The research response during three months of the COVID-19 outbreak. Biomedicine & Pharmacotherapy. 2020; 129:110451. https://doi.org/10.1016/j.biopha.2020.110451 PMID: 32603887
3. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. PLOS Biology. 2021; 19(4):1–28. https://doi.org/10.1371/journal.pbio.3000959 PMID: 33798194
4. Reitinger J, Jain SF, Suh D. Significant increase in non-COVID-19 related ophthalmology publications during the COVID-19 era: is this a new normal? Eye. 2021; 35:1041–1042. https://doi.org/10.1038/s41433-020-01220-3 PMID: 33051622
5. Aspachs O, Durante R, Graziano A, Mestres J, Reynal-Querol M, Montalvo JG. Tracking the impact of COVID-19 on economic inequality at high frequency. PLOS ONE. 2021; 16(3):1–14. https://doi.org/10.1371/journal.pone.0249121 PMID: 33788886
6. Palayew A, Norgaard O, Safreed-Harmon K, Andersen T, Helms R, Neimann L, et al. Pandemic publishing poses a new COVID-19 challenge. Nature Human Behavior. 2020; 4(4841):666–669. https://doi.org/10.1038/s41562-020-0911-0 PMID: 32576981
7. Collins C. Productivity in a pandemic. Science. 2020; 369(6504):603–603. https://doi.org/10.1126/science.abe1163 PMID: 32764040
8. Malisch JL, Harris BN, Sherrer SM, Lewis KA, Shepherd SL, McCarthy PC, et al. Opinion: In the wake of COVID-19, academia needs new solutions to ensure gender equity. Proceedings of the National Academy of Sciences. 2020; 117(27):15378–15381. https://doi.org/10.1073/pnas.2010636117 PMID: 32554503
9. Minello A. The pandemic and the female academic. Nature. 2020;April 17. https://doi.org/10.1038/d41586-020-01135-9 PMID: 32303729
10. Wenham C, Smith J, Sara E Davies, Feng Huiyun, Grépin KA, Harman S, et al. Women are most affected by pandemics—Lessons from past outbreaks. Nature. 2020; 583:194–198. https://doi.org/10.1038/d41586-020-02006-z PMID: 32641809
11. Greider CW, Sheltzer JM, Cantalupo NC, Copeland WB, Dasgupta N, Hopkins N, et al. Increasing gender diversity in the STEM research workforce. Science. 2019; 366(6466):692–695. https://doi.org/10.1126/science.aaz0649 PMID: 31699926
12. Cech EA, Blair-Loy M. The changing career trajectories of new parents in STEM. Proceedings of the National Academy of Sciences. 2019; 116(10):4182–4187. https://doi.org/10.1073/pnas.1810862116 PMID: 30782835
13. Day AE, Corbett P, Boyle J. Is there a gender gap in chemical sciences scholarly communication? Chemical Science. 2020; 11:2277–2301. https://doi.org/10.1039/c9sc04090k PMID: 32180933
14. Yavorsky JE, Qian Y, Sargent AC. The gendered pandemic: The implications of COVID-19 for work and family. Sociology Compass. 2021; 15(6):e12881. https://doi.org/10.1111/soc4.12881 PMID: 34230836
15. Collins C, Landivar LC, Ruppanner L, Scarborough WJ. COVID-19 and the Gender Gap in Work Hours. Gender, Work & Organization. 2020;In Press. https://doi.org/10.1111/gwao.12506 PMID: 32837019
16. Myers KR, Tham WY, Yin Y, Cohodes N, Thursby JG, Thursby MC, et al. Unequal effects of the COVID-19 pandemic on scientists. Nature Human Behaviour. 2020. https://doi.org/10.1038/s41562-020-0921-y PMID: 32669671
17. Carr RM, Lane-Fall MB, South E, Brady D, Momplaisir F, Guerra CE, et al. Academic careers and the COVID-19 pandemic: Reversing the tide. Science Translational Medicine. 2021; 13(584). https://doi.org/10.1126/scitranslmed.abe7189 PMID: 33692133
18. Andersen JP, Nielsen MW, Simone NL, Lewiss RE, Jagsi R. Meta-Research: COVID-19 medical papers have fewer women first authors than expected. eLife. 2020; 9:e58807. https://doi.org/10.7554/eLife.58807 PMID: 32538780

19. Pinho-Gomes AC, Peters S, Thompson K, Hockham C, Ripullone K, Woodward M, et al. Where are the women? Gender inequalities in COVID-19 research authorship. BMJ Global Health. 2020; 5(7): e002922. https://doi.org/10.1136/bmjgh-2020-002922 PMID: 32527733

20. Lerchenmüller C, Schmallenbach L, Jena AB, Lerchenmueller MJ. Longitudinal analyses of gender differences in first authorship publications related to COVID-19. BMJ Open. 2021; 11(4). https://doi.org/10.1136/bmjopen-2020-045176 PMID: 33820790

21. Bell ML, Fong KC. Gender Differences in First and Corresponding Authorship in Public Health Research Submissions During the COVID-19 Pandemic. American Journal of Public Health. 2021; 111(1):159–163. https://doi.org/10.2105/AJPH.2020.305975 PMID: 33211581

22. DeFilippis EM, Sinnenberg L, Mahmud N, Wood MJ, Hayes SN, Michos ED, et al. Gender Differences in Publication Authorship During COVID-19: A Bibliometric Analysis of High-Impact Cardiology Journals. Journal of the American Heart Association. 2021; 10(5):e019005. https://doi.org/10.1161/JAHA.120.019005 PMID: 33619980

23. Fox CW, Meyer J. The influence of the global COVID-19 pandemic on manuscript submissions and editor and reviewer performance at six ecology journals. Functional Ecology. 2021; 35(1):4–10. https://doi.org/10.1111/1365-2435.13734

24. Squazzoni F, Grimaldo F, Marušić A. Publishing: Journals could share peer-review data. Nature. 2017; 546. https://doi.org/10.1038/546352a PMID: 28617464

25. Buljan I, Garcia-Costa D, Grimaldo F, Squazzoni F, Marušić A. Meta-Research: Large-scale language analysis of peer review reports. eLife. 2020; 9:e53249. https://doi.org/10.7554/eLife.53249 PMID: 32678065

26. Bravo G, Grimaldo F, López-Iñesta E, Mehmani B, Squazzoni F. The effect of publishing peer review reports on referee behavior in five scholarly journals. Nature Communications. 2019; 10(1):322. https://doi.org/10.1038/s41467-018-08250-2 PMID: 30659186

27. Karimi F, Wagner C, Lemmerich F, Jadidi M, Strohmaier M. Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods. In: Proceedings of the 25th International Conference Companion on World Wide Web. WWW'16 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee; 2016. p. 53–54. Available from: https://doi.org/10.1145/2872518.2889385.

28. Helmer M, Schottdorf M, Neef A, Battaglia D. Research: Gender bias in scholarly peer review. eLife. 2017; 6:e21718. https://doi.org/10.7554/eLife.21718 PMID: 28322725

29. Santamaría L, Mihaljević H. Comparison and benchmark of name-to-gender inference services. PeerJ Computer Science. 2018; 4:e156. https://doi.org/10.7717/peerj-cs.156 PMID: 33816809

30. R Core Team. R: A Language and Environment for Statistical Computing; 2021. Available from: https://www.R-project.org/.

31. Andriole D, Whelan A, Jeffe D. Characteristics and career intentions of the emerging MD/PhD workforce. JAMA. 2008; 300. https://doi.org/10.1001/jama.300.10.1165 PMID: 18780845

32. dos Santos Rocha A, Scherlinger M, Ostermann L, Mehler DMA, Nadiradze A, Schulze F, et al. Characteristics and opinions of MD-PhD students and graduates from different European countries: a study from the European MD-PhD Association. Swiss Medical Weekly. 2020. https://doi.org/10.4414/smw.2020.20205 PMID: 32294222

33. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2017; 2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z

34. Gigliotti P, Martin EG. Predictors of State-Level Stay-at-Home Orders in the United States and Their Association With Mobility of Residents. Journal of Public Health Management and Practice. 2020; 26 (6):622–631. https://doi.org/10.1097/PHH.0000000000001236 PMID: 32969952

35. Saha J, Chouhan P. Lockdown and unlock for COVID-19 and its impact on residential mobility in India: an analysis of the COVID-19 Community Mobility Reports, 2020. International Journal of Infectious Diseases. 2020. https://doi.org/10.1016/j.ijid.2020.11.187 PMID: 33253865

36. Zhu D, Mishra SR, Han X, Santo K. Social distancing in Latin America during the COVID-19 pandemic: an analysis using the Stringency Index and Google Community Mobility Reports. Journal of Travel Medicine. 2020. https://doi.org/10.1093/jtm/taaa125 PMID: 32729931

37. Inno L, Rotundi A, Piccialli A. COVID-19 lockdown effects on gender inequality. Nature Astronomy. 2021; 4 (1114). https://doi.org/10.1038/s41550-020-01258-z

38. Bauchner H, Fontanarosa PB, Golub RM. Editorial Evaluation and Peer Review During a Pandemic: How Journals Maintain Standards. JAMA. 2020; 324(5):453–454. https://doi.org/10.1001/jama.2020.11764 PMID: 32589195

39. Squazzoni F, Ahrweiler P, Barros T, Bianchi F, Birukou A, Blom HJJ, et al. Unlock ways to share data on peer review. Nature. 2020; 578:512–514. https://doi.org/10.1038/d41586-020-00500-y PMID: 32099126

40. Perry BL, Aronson B, Pescosolido BA. Pandemic precarity: COVID-19 is exposing and exacerbating inequalities in the American heartland. Proceedings of the National Academy of Sciences. 2021; 118 (8). https://doi.org/10.1073/pnas.2020685118 PMID: 33547252

41. Squazzoni F, Bravo G, Farjam M, Marusic A, Mehmani B, Willis M, et al. Peer review and gender bias: A study on 145 scholarly journals. Science Advances. 2021; 7(2). https://doi.org/10.1126/sciadv.abd0299 PMID: 33523967

42. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. The Leiden Manifesto for research metrics. Nature. 2015; 520:429–431. https://doi.org/10.1038/520429a PMID: 25903611

43. Htun M. Tenure and promotion after the pandemic. Science. 2020; 368(6495):1075–1075. https://doi.org/10.1126/science.abc7469 PMID: 32499434

44. Kamerlin SCL, Wittung-Stafshede P. Female Faculty: Why So Few and Why Care? Chemistry—A European Journal. 2020. https://doi.org/10.1002/chem.202002522 PMID: 32583921

45. Gibson EM, Bennett FC, Gillespie SM, Güler AD, Gutmann DH, Halpern CH, et al. How Support of Early Career Researchers Can Reset Science in the Post-COVID19 World. Cell. 2020; 181(7):1445–1449. https://doi.org/10.1016/j.cell.2020.05.045 PMID: 32533917

46. Berenbaum MR. Speaking of gender bias. Proceedings of the National Academy of Sciences. 2019; 116(17):8086–8088. https://doi.org/10.1073/pnas.1904750116 PMID: 30967503

Bibliografía

A continuación se muestra la bibliografía utilizada para la elaboración de esta memoria de tesis doctoral, que no contiene las referencias bibliográficas de las contribuciones. La bibliografía de cada contribución puede consultarse en cada uno de los artículos.

Anaconda, I. (2022). State of data science 2022.

Atjonen, P. (2019). Peer review in the development of academic articles: Experiences of finnish authors in the educational sciences. *Learned Publishing*, 32(2):137–146.

Augsten, N. and Böhlen, M. H. (2014). *Token-Based Distances*, pages 25–59. Springer International Publishing, Cham.

Bianchi, F., García-Costa, D., Grimaldo, F., and Squazzoni, F. (2022). Measuring the effect of reviewers on manuscript change: A study on a sample of submissions to royal society journals (2006–2017). *Journal of Informetrics*, 16(3):101316.

Bianchi, F., Grimaldo, F., and Squazzoni, F. (2019). The f3-index. valuing reviewers for scholarly journals. *Journal of Informetrics*, 13(1):78–86.

Biega, A. J., Potash, P., Daumé, H., Diaz, F., and Finck, M. (2020). Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 399–408, New York, NY, USA. Association for Computing Machinery.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245.

Bravo, G., Farjam, M., Grimaldo, F., Birukou, A., and Squazzoni, F. (2018). Hidden connections: network effects on editorial decisions in four computer science journals. *Journal of Informetrics*, 12(1):101–112.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

Buljan, I., Garcia-Costa, D., Grimaldo, F., Squazzoni, F., and Marušić, A. (2020). Meta-research: Large-scale language analysis of peer review reports. *eLife*, 9:e53249.

Cook, H. V. and Jensen, L. J. (2019). *A Guide to Dictionary-Based Text Mining*, pages 73–89. Springer New York, New York, NY.

Coupé, T. (2013). Peer review versus citations – an analysis of best paper prizes. *Research Policy*, 42(1):295–301.

Cowley, S. J. (2015). How peer-review constrains cognition: on the frontline in the knowledge sector. *Frontiers in Psychology*, 6:1706.

Deng, Q., Hine, M. J., Ji, S., and Sur, S. (2017). Building an environmental sustainability dictionary for the it industry. In *Hawaii International Conference on System Sciences 2017*, Honolulu, Hawai. University of Hawaií at Manoa.

Edmonds, B., Gilbert, N., Ahrweiler, P., and Scharnhorst, A. (2011). Simulating the social processes of science. *Journal of Artificial Societies and Social Simulation*, 14(4):14.

Edwards, M. A. and Siddhartha, R. (2017). Academic research in the 21^{st} century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, 34(1):51–61.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabási, A.-L. (2018). Science of science. *Science*, 359(6379):eaao0185.

Garcia-Costa, D., Forte, A., Lòpez-Iñesta, E., Squazzoni, F., and Grimaldo, F. (2022a). Does peer review improve the statistical content of manuscripts? a study on 27 467 submissions to four journals. *Royal Society Open Science*, 9(9):210681.

Garcia-Costa, D., Squazzoni, F., Mehmani, B., and Grimaldo, F. (2022b). Measuring the developmental function of peer review: a multidimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. *PeerJ*.

García, S., Francisco, H., and Luengo, J. (2006). *Text Mining Application Programming*. Springer, New York.

Godbole, S., Bhattacharya, I., Gupta, A., and Verma, A. (2010). Building re-usable dictionary repositories for real-world text mining. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 1189–1198, New York, NY, USA. Association for Computing Machinery.

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., and Farkash, A. (2011). Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, 2(3):477–491.

Groves, T. (2010). Is open peer review the fairest system? yes. *BMJ*, 341.

Jo, T. (2019). *Text Mining. Concepts, Implementation, and Big Data Challenge*. Springer, New York.

Jones, K. S. (1994). *Natural Language Processing: A Historical Review*, pages 3–16. Springer Netherlands, Dordrecht.

Kassirer, J. P. and Campion, E. W. (1994). Peer Review: Crude and Understudied, but Indispensable. *JAMA*, 272(2):96–97.

Khan, K. (2010). Is open peer review the fairest system? no. *BMJ*, 341.

Kharasch, E. D., Avram, M. J., Clark, J. D., Davidson, A. J., Houle, T. T., Levy, J. H., London, M. J., Sessler, D. I., and Vutskits, L. (2021). Peer review matters: Research quality and the public trust. *Anesthesiology*, 134(1):1–6.

Köhler, T., González-Morales, M. G., Banks, G. C., O'Boyle, E. H., Allen, J. A., Sinha, R., Woo, S. E., and Gulick, L. M. V. (2020). Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: Introducing a competency framework for peer review. *Industrial and Organizational Psychology*, 13(1):1–27.

Lee, C. J. and Moher, D. (2017). Promote scientific integrity via journal peer review data. *Science*, 357(6348):256–257.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Luengo, J., García-Gil, D., Ramírez-gallego, S., García, S., and Francisco, H. (2006). *Big Data Preprocessing*. Springer, New York.

Maimon, O. and Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. Springer, New York.

Martín, E. (2016). How double-blind peer review works and what it takes to be a good referee. *Current Sociology*, 64(5):691–698.

Merriman, B. (2020). Peer review as an evolving response to organizational constraint: Evidence from sociology journals, 1952–2018. *The American Sociologist*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mpouli, S., Beigbeder, M., and Largeron, C. (2020). Lexifield: a system for the automatic building of lexicons by semantic expansion of short word lists. *Knowledge and Information Systems*, 62:3181–3201.

Muresan, S. and Klavans, J. (2002). A method for automatically building and evaluating dictionary resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Murthy, S., Abu Bakar, A., Abdul Rahim, F., and Ramli, R. (2019). A comparative study of data anonymization techniques. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 306–309.

Piel, G. (1986). The social process of science. *Science*, 231(4735):201–201.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Rigby, J., Cox, D., and Julian, K. (2018). Journal peer review: a bar or bridge? an analysis of a paper's revision history and turnaround time, and the effect on citation. *Scientometrics*, 114(4):1087–1105.

Ross-Hellauer, T. (2017). What is open peer review? a systematic review. *F1000Research*, 6:588.

Santamaría, L. and Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156.

Seeber, M. (2020). How do journals of different rank instruct peer reviewers? Reviewer guidelines in the field of management. *Scientometrics*, 122:1387–140.

Squazzoni, F., Ahrweiler, P., Barros, T., Bianchi, F., Birukou, A., Blom, H. J. J., Bravo, G., Cowley, S., Dignum, V., Dondio, P., Grimaldo, F., Haire, L., Hoyt, J., Hurst, P., Lammey, R., MacCallum, C., Marušić, A., Mehmani, B., Murray, H., Nicholas, D., Pedrazzi, G., Puebla, I., Rodgers, P., Ross-Hellauer, T., Seeber, M., Shankar, K., Van Rossum, J., and Willis, M. (2020). Unlock ways to share data on peer review. *Nature*, 578:512–514.

Squazzoni, F., Bravo, G., Grimaldo, F., García-Costa, D., Farjam, M., and Mehmani, B. (2021). Gender gap in journal submissions and peer review during the first wave of the covid-19 pandemic. a study on 2329 elsevier journals. *PLoS ONE*, 16(10).

Squazzoni, F., Brezis, E., and Marušić, A. (2017a). Scientometrics of peer review. *Scientometrics*, 113(1):501.

Squazzoni, F., Grimaldo, F., and Marusic, A. (2017b). Publishing: Journals could share peer-review data. *Nature*, 546(352).

Superchi, C., Hren, D., Blanco, D., Rius, R., Recchioni, A., Boutron, I., and González, J. A. (2020). Development of arcadia: a tool for assessing the quality of peer-review reports in biomedical research. *BMJ Open*, 10(6).

Taher Pilehvar, M. and Camacho-Collados, J. (2021). *Embeddings in Natural Language Processing*. Springer Cham, Switzerland.

Wilkinson, J. (2017). Tracking global trends in open peer review.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences.