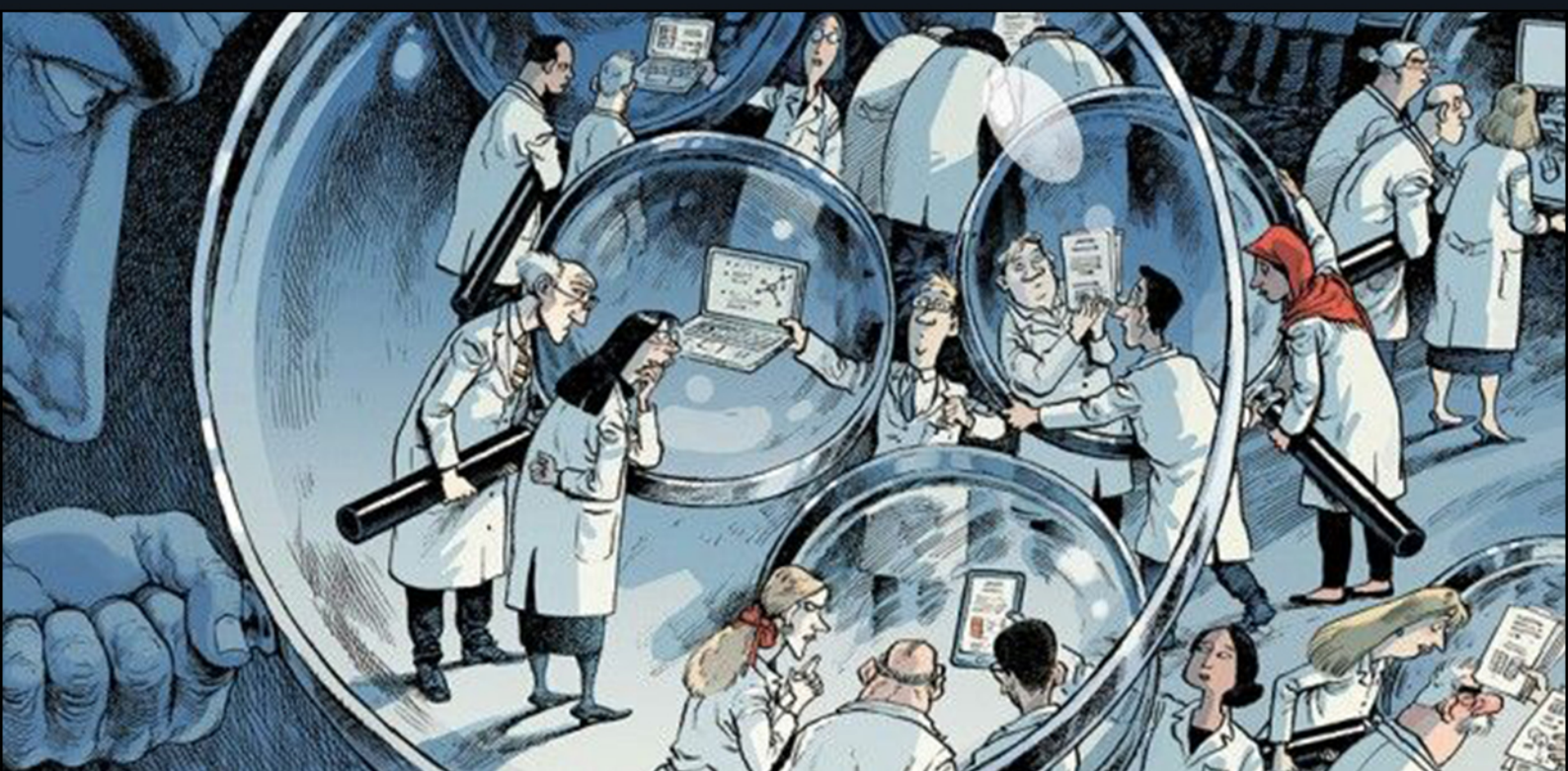


MEASURING REVIEW REPORT QUALITY IN HEALTH RESEARCH

DEVELOPMENT AND VALIDATION OF ARCADIA

CECILIA SUPERCHI



DOCTORAL THESIS 2020

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH

UNIVERSITAT POLITÈCNICA DE CATALUNYA



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH



MEASURING REVIEW REPORT QUALITY IN HEALTH RESEARCH

Development and validation of ARCADIA

Doctoral thesis by:

Cecilia Superchi

Thesis director:

José Antonio González, PhD

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya, Spain

Thesis co-directors:

Darko Hren, PhD

Faculty of Humanities and Social Science
University of Split, Croatia

Isabelle Boutron, MD, PhD

Centre d'épidémiologie Clinique
Hôpital Hôtel-Dieu, France

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya, Barcelona-Tech
Barcelona, Spain 2020



Université de Paris, CRESS, INSERM, INRA
Paris, France 2020

Thesis submitted to obtain the title of Doctor by the Universitat Politècnica de Catalunya.
Department of Statistics and Operations Research
Barcelona, December 2020

All articles included in the dissertation are open-access articles distributed in accordance with the Creative Commons Attribution 4.0 International License.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 676207.

© Illustration by David Parkins used in the cover page. The illustration was included in the published article entitled "Make peer review scientific" by D. Rennie (Nature, 2016).

“But even though Quality cannot be defined, you know what Quality is!”

— Robert M. Pirsig, *Zen and the Art of Motorcycle Maintenance* (1974)

ACKNOWLEDGMENTS

I began this journey in March 2017 when the MiRoR project was already started. At that time, I felt extremely excited, but also quite scared, to start this PhD programme, mainly because I did not know what was expected of me. However, just after a few months, I realized that this programme was not only a significant professional and academic experience but also an important human and life adventure, which allowed me to meet extraordinary people.

I would like to express my sincere gratitude to my three supervisors which each of them expertly guided me through this journey teaching me diverse academic expertise and, more important, human qualities. Thanks to José Antonio for his kindness, humanity, and sense of humor. Muchas gracias por estar presente cuando más lo necesitaba. Thanks to Darko for his deep soul and philosophical, and also artistic, approach to facing life. Thanks to Isabelle for her dedication, professionalism, and passion for science. Each of them has been fundamental for the result of this PhD project.

During this journey, I had also the singular pleasure of meeting and working with those researchers whose names I used to read in those articles that I keep in a computer folder titled “Relevant literature”. Thanks to the colossal and inimitable Doug Altman, who sadly passed away just after my secondment at Centre for Statistics in Medicine (CSM) in Oxford, for his simplicity and revolutionary mind. Thanks to the experienced Gary Collins for his passion and enthusiasm for carrying out the extraordinary work of EQUATOR Network. Thanks to Sara Schroter for her intelligence, expertise, and humility.

Thanks to the entire MiRoR Network for being an important source of inspiration for me. In particular, I would like to express my gratitude to my mentor Alessandro Recchioni for his advice and agreeable chats around Barcelona, Split, and Paris. Thanks also to the other MiRoR

students (Linda, Maria, Alice, Evi, Van, Camila, Lorenzo, Thang, Mel, Mona, Anna, Cristopher, Keti, and David) for sharing with me this programme and learning experience.

This journey has been much easier and funny thanks to my *compañeros* and *compañeras de viaje*. Thanks to Marta for being not only a good colleague and a flatmate but also a real friend. Thanks to David for sharing with me not only all-conference trips but also every day of this journey with energy and enthusiasm. Thanks to Yovi for our chats, coffee breaks and laughs at the gym. Thanks to Fran for her comfortable words and enormous kindness.

I would like also to thank Sonia for being of great support and help for all the administrative issues I had over these years. Thanks to Erik, Jordi, Klaus, Lupe, Roser, and Toni for being always present when I needed it. Thanks to all of you for making the Department of Statistics and Operations Research a pleasant and nice working environment.

Thanks to my *Little women* who have been always next to me although the geographical distances. Thanks to Francesca for her creativity, *Peter Pan's* soul, and for being my friend since over 20 years. Thanks to Federica for her loyalty, healthy craziness, and endless chats. Thanks to Malda for her perseverance and sweetness. Thanks to Elena for being just 'La Mora' with her congeniality. Thanks to Camilla for listening to me in one of the darkest moments of this long journey.

Thanks also to Jack for understanding me better than anyone else thanks to our common life choices. Thanks to Trocci and his family for their positivity and joy of life. Thanks to Alessandro and Pier for sharing with me two unforgettable trips in South America and Greece. Thanks also to all *Da ufo* for making me feel like I have never left Parma although I spent the last eight years of my life around Europe.

Thanks to 'THE HOUSE' and all its friends for the *cervezas* (watching the Barça matches), parties and laughs in good company in Plaça Osca. Thanks also to all people who I met during this journey and they made me feel at home in Barcelona, Oxford, Paris, and Split.

Grazie a miei due fratelli Filippo e Alberto perché, nonostante la vita in alcuni momenti ci abbia allontanato, so che ci siete stati e ci sarete sempre per me.

Il grazie più sentito va ai miei genitori, Anna e Gabriele per avermi incoraggiata, supportata e ascoltata SEMPRE. Grazie per avermi aiutato a prendere il volo e avermi dimostrato, in tutti questi anni lontana da casa, che per qualsiasi necessità siete sempre pronti a riaccogliermi. Grazie per tutte le volte che mi siete venuti a prendere in aeroporto, per tutte le volte che mi siete venuti a trovare ad Oxford, Barcelona, Parigi e Spalato e per tutti i pacchi spediti in giro per l'Europa. Grazie per farmi sentire amata ogni giorno.

Questo viaggio mi ha anche permesso di incontrare (o meglio rincontrare) Nicolò. Grazie per avermi sorriso ogni giorno di questi due anni insieme. Grazie anche per essere non solo un fantastico e presente partner ma anche un divertissimo compagno di viaggi e balli. Spero che la vita ci regali altrettante avventure insieme.

My deepest appreciation goes to all of you. Thank you! Merci! Gracias! Hvala! Grazie!

Cecilia

ABSTRACT

Editorial peer review is the gateway to scientific publication. It was established to ensure that research papers were vetted by independent experts before they are published. Despite the importance of this process, its impact is still considered suboptimal and it needs to be improved. For this purpose, we need appropriate outcome measures, particularly a validated tool that clearly defines the quality of peer review reports. The final aim of the present PhD project was to develop and validate a new tool for assessing the quality of peer review reports in biomedical research.

As the starting point for the development of a new tool, we performed a systematic review aimed to identify and describe the existing tools used to assess peer review report quality in biomedical research. We identified a total number of 24 tools: 23 scales and 1 checklist. None of the tools reported a definition of 'quality'. Only one described the scale development and 10 provided measures of validity and reliability. We classified the quality components of the 18 tools with more than one item into 9 main quality domains and 11 subdomains.

Secondly, we formed a steering committee composed of five members with diverse expertise, which defined the quality of peer review reports. We then conducted an online survey intended for biomedical editors and authors to 1) determine if participants endorsed the proposed definition of peer review report quality; 2) identify the most important items to include in the tool, and 3) identify any missing items. Based on the participants' qualitative and quantitative answers, the steering committee modified the initially proposed definition of peer review report quality, reviewed all items, and ultimately, drafted and refined the final version of the tool.

The ARCADIA (Assessment of Review reports with a Checklist Available to eDItors and Authors) tool was finally developed. The tool is a checklist that includes 14 items encompassed in 5 domains. Each item should be ticked as 'Yes' or 'No'. However, an item could also be

assessed as 'Not applicable' (NA) depending on the reviewer's expertise, type of study, type of biomedical journal, availability of study data, materials and protocol.

Finally, we tested the tool and evaluated its acceptability, reliability, and validity. ARCADIA was validated by a heterogeneous sample of both biomedical editors and authors using a sample of peer review reports from two different biomedical journals (i.e., The BMJ and BMJ Open). Field-testing demonstrated that the psychometric properties of ARCADIA are not entirely satisfactory. Results from the validation study should be used to inform a new version of the ARCADIA tool, which should be also validated in a real-editorial setting using peer review reports associated with manuscripts with different study designs and from different types of journals.

This thesis reports the development and validation of ARCADIA, a new tool for assessing the quality of peer review reports in biomedical research. ARCADIA constitutes the first tool that has been systematically developed to assess the quality of peer review reports and its validation is based on a large and diverse sample of biomedical editors and authors. This tool could be used regularly by editors to evaluate the reviewers' work, and also as an outcome when evaluating interventions to improve the peer review process.

RESUMEN

La revisión editorial por pares es la puerta de entrada a la publicación científica. Se estableció para garantizar que los artículos de investigación fueran examinados por expertos independientes antes de su publicación. A pesar de la importancia de este proceso, su impacto aún se considera subóptimo y debe mejorarse. Para este propósito, necesitamos outcomes apropiados, particularmente una herramienta validada que defina claramente la calidad de los informes de revisión por pares. El objetivo final del presente proyecto de doctorado fue desarrollar y validar una nueva herramienta para evaluar la calidad de los informes de revisión por pares en la investigación biomédica.

Como punto de partida para el desarrollo de una nueva herramienta, realizamos una revisión sistemática destinada a identificar y describir las herramientas existentes que se utilizan para evaluar la calidad de los informes de revisión por pares en la investigación biomédica. Identificamos un total de 24 herramientas: 23 escalas y 1 lista de comprobación. Ninguna de las herramientas incorporó una definición de 'calidad'. Solo una describió el desarrollo de la escala y 10 proporcionaron medidas de validez y confiabilidad. Clasificamos los componentes de calidad de las 18 herramientas con más de un elemento en 9 dominios de calidad principales y 11 subdominios.

En segundo lugar, formamos un comité directivo compuesto por cinco miembros con experiencia diversa que definieron la calidad de los informes de revisión por pares. Luego, realizamos una encuesta online destinada a editores y autores biomédicos para 1) determinar si los participantes respaldaban la definición propuesta de calidad del informe de revisión por pares; 2) identificar los ítems más importantes para incluir en la herramienta; y 3) identificar cualquier ítem faltante. Sobre la base de las respuestas cualitativas y cuantitativas de los participantes,

el

comité directivo modificó la definición propuesta inicialmente de calidad del informe de revisión por pares, revisó todos los ítems y, por último, redactó y perfeccionó la versión final de la herramienta.

Seguidamente se desarrolló la herramienta ARCADIA (Assessment of Review reports with a Checklist Available to eDItors and Authors). La herramienta es una lista de comprobación que incluye 14 elementos englobados en 5 dominios. Cada elemento debe marcarse como ‘Sí’ o ‘No’. Sin embargo, un elemento también podría evaluarse como ‘No aplicable’ (NA) según la experiencia del revisor, el tipo de estudio, el tipo de revista biomédica, la disponibilidad de los datos del estudio, los materiales y el protocolo.

Finalmente, probamos la herramienta y evaluamos su aceptabilidad, fiabilidad y validez. ARCADIA fue validada por una muestra heterogénea de editores y autores biomédicos utilizando informes de revisión por pares de dos revistas biomédicas diferentes (es decir, The BMJ y BMJ Open). Las pruebas de campo demostraron que las propiedades psicométricas de ARCADIA no son del todo satisfactorias. Los resultados del estudio de validación deben usarse para impulsar una nueva versión de la herramienta ARCADIA, que también debe validarse en un entorno editorial real utilizando informes de revisión por pares asociados con manuscritos con diferentes diseños de estudio y de diferentes tipos de revistas.

Esta tesis informa sobre el desarrollo y validación de ARCADIA, una nueva herramienta para evaluar la calidad de los informes de revisión por pares en la investigación biomédica. ARCADIA constituye la primera herramienta que se ha desarrollado sistemáticamente para evaluar la calidad de los informes de revisión por pares, validada con una amplia y diversa muestra de editores y autores biomédicos. Los editores podrían utilizar esta herramienta con regularidad para evaluar el trabajo de los revisores y también como variable respuesta al evaluar intervenciones para mejorar el proceso de revisión por pares.

CONTENTS

| | |
|--|------------|
| LIST OF FIGURES | III |
| LIST OF TABLES | IV |
| ABBREVIATIONS | V |
| CHAPTER 1. INTRODUCTION | 1 |
| RESEARCH ON RESEARCH..... | 2 |
| THE MiROR PROJECT..... | 3 |
| THE PEER REVIEW PROCESS..... | 3 |
| HOW DOES THE EDITORIAL PEER REVIEW WORK? | 4 |
| OPEN PEER REVIEW | 5 |
| THE HISTORY OF THE PEER REVIEW PROCESS..... | 6 |
| ACTORS OF THE PEER REVIEW PROCESS..... | 7 |
| JOURNAL EDITORS | 7 |
| PEER REVIEWERS | 8 |
| CRITICISMS OF THE PEER REVIEW PROCESS..... | 9 |
| GOAL AND THESIS STRUCTURE | 11 |
| CHAPTER 2. TOOLS USED TO ASSESS THE QUALITY OF PEER REVIEW REPORTS: A METHODOLOGICAL SYSTEMATIC REVIEW | 13 |
| BACKGROUND..... | 14 |
| METHODS | 14 |
| STUDY DESIGN..... | 14 |
| INFORMATION SOURCES AND SEARCH STRATEGY..... | 14 |
| ELIGIBILITY CRITERIA | 15 |
| STUDY SELECTION | 15 |
| DATA EXTRACTION..... | 16 |
| RESULTS..... | 18 |
| STUDY SELECTION AND GENERAL CHARACTERISTICS OF REPORTS | 18 |
| GENERAL CHARACTERISTICS OF THE TOOLS | 19 |
| QUALITY COMPONENTS OF THE PEER REVIEW REPORTS CONSIDERED IN THE TOOLS WITH MORE THAN ONE ITEM | 26 |
| CLUSTERING ANALYSIS AMONG TOOLS..... | 29 |
| DISCUSSION | 31 |
| CONCLUSIONS..... | 32 |
| CHAPTER 3. THE DEVELOPMENT OF ARCADIA: A TOOL FOR ASSESSING THE QUALITY OF PEER REVIEW REPORTS IN BIOMEDICAL RESEARCH | 33 |
| BACKGROUND..... | 34 |
| METHODS | 34 |
| STEERING COMMITTEE | 34 |
| DEFINING THE TOOL'S OBJECTIVE..... | 35 |
| GENERATING THE ITEMS | 35 |
| SURVEY | 35 |
| SURVEY QUESTIONNAIRE | 36 |
| PARTICIPANTS AND RECRUITMENT STRATEGY..... | 37 |
| DATA ANALYSIS | 39 |
| SELECTING ITEMS | 40 |
| RESULTS..... | 40 |
| PARTICIPANTS | 40 |

| | |
|---|------------|
| DEFINITION OF PEER REVIEW REPORT QUALITY | 41 |
| QUANTITATIVE RESULTS | 42 |
| QUALITATIVE RESULTS | 46 |
| STEERING COMMITTEE MEETING | 50 |
| THE ARCADIA TOOL..... | 54 |
| DISCUSSION | 62 |
| CONCLUSIONS..... | 62 |
| | |
| CHAPTER 4. PSYCOMETRIC TESTING OF THE ARCADIA TOOL | 63 |
| BACKGROUND..... | 64 |
| METHODS | 64 |
| SCORING OF THE ARCADIA TOOL | 64 |
| ASSESSMENT OF THE PSYCOMETRIC PROPERTIES | 65 |
| PHASE I. ACCEPTABILITY, INTERNAL CONSISTENCY, INTER-RATER AND TEST-RETEST RELIABILITY..... | 67 |
| PHASE II. PRACTICABILITY, INTER-RATER RELIABILITY AND CONSTRUCT VALIDITY | 70 |
| RESULTS..... | 77 |
| PHASE I. ACCEPTABILITY, INTERNAL CONSISTENCY, INTER-RATER AND TEST-RETEST RELIABILITY..... | 77 |
| PHASE II. PRACTICABILITY, INTER-RATER RELIABILITY AND CONSTRUCT VALIDITY | 83 |
| DISCUSSION | 94 |
| CONCLUSIONS..... | 94 |
| | |
| CHAPTER 5. GENERAL DISCUSSION | 95 |
| | |
| ARTICLES | 105 |
| | |
| OTHER PUBLICATIONS | 106 |
| | |
| SCIENTIFIC PORTFOLIO..... | 107 |
| | |
| REFERENCES..... | 110 |
| | |
| APPENDIX | 121 |
| APPENDIX 1. SEARCH STRATEGIES | 122 |
| APPENDIX 2. EXCLUDED STUDIES AND REASONS FOR EXCLUSION | 124 |
| APPENDIX 3. INCLUDED STUDIES | 127 |
| APPENDIX 4. CLASSIFICATION OF PEER REVIEW REPORT QUALITY COMPONENTS | 130 |
| APPENDIX 5. SURVEY QUESTIONNAIRE | 135 |
| APPENDIX 6. INVITATION EMAIL | 149 |
| APPENDIX 7. TOP 30-BIOMEDICAL JOURNALS WITH THE HIGHEST IMPACT FACTORS..... | 151 |
| APPENDIX 8. COMPLETE PARTICIPANTS CHARACTERISTICS..... | 152 |
| APPENDIX 9. SURVEY QUESTIONNAIRE | 154 |
| APPENDIX 10. INVITATION EMAIL FOR BIOMEDICAL EDITORS AND AUTHORS WHO AGREED TO TAKE PART IN THE VALIDATION STUDY | 156 |
| APPENDIX 11. PARTICIPANTS´ FEEDBACK ON ARCADIA | 157 |
| APPENDIX 12. CODEBOOK OF PARTICIPANTS´ COMMENTS (N=32)..... | 160 |

LIST OF FIGURES

| | |
|---|----|
| FIG. 1 BASIC PUBLISHING WORKFLOW | 4 |
| FIG. 2. STUDY SELECTION FLOW DIAGRAM | 19 |
| FIG. 3. FREQUENCY OF QUALITY DOMAINS AND SUBDOMAINS..... | 29 |
| FIG. 4. HIERARCHICAL CLUSTERING OF TOOLS BASED ON THE NINE QUALITY DOMAINS..... | 30 |
| FIG. 5. SHINY APPLICATION WELCOME PAGE | 42 |
| FIG. 6. THE 20 ITEMS RATED BY EDITORS AND AUTHORS..... | 43 |
| FIG. 7. PCA PLOT (PC1 vs PC2)..... | 45 |
| FIG. 8. PCA PLOT (PC2 vs PC3 | 45 |
| FIG. 9. FLOWCHART OF ITEMS | 51 |
| FIG. 10. SURVEY WELCOME PAGE | 71 |
| FIG. 11. COMPARISON OF THE ASSESSMENT OF EACH ARCADIA ITEM BETWEEN THE TWO RATERS ON THE SAMPLE OF 162 PEER REVIEW REPORTS..... | 77 |
| FIG. 12. DISTRIBUTION OF THE ARCADIA OVERALL QUALITY SCORES | 78 |
| FIG. 13. COMPARISON OF BEFORE-AFTER ARCADIA OVERALL SCORES BETWEEN RATERS ON A SUBSAMPLE OF 30 PEER REVIEW REPORTS | 82 |
| FIG. 14. ASSESSMENT OF EACH ARCADIA ITEM BY SURVEY PARTICIPANTS | 84 |
| FIG. 15. COMPARISON OF ARCADIA DOMAIN MEAN SCORES AND OVERALL SCORE WITH A SUBJECTIVE SCALE BETWEEN RATERS..... | 86 |
| FIG. 16. SURVEY PARTICIPANTS' FEEDBACK ON ARCADIA | 87 |
| FIG. 17. PRINCIPAL COMPONENT ANALYSIS (PCA) OF VARIABLES (I.E., ARCADIA DOMAIN SCORES AND OVERALL QUALITY SCORE WITH A SUBJECTIVE SCALE)..... | 91 |
| FIG. 18. PRINCIPAL COMPONENT ANALYSIS (PCA) OF INDIVIDUALS (I.E., PEER REVIEW REPORTS) | 91 |
| FIG. 19. CORRESPONDENCE ANALYSIS OF VARIABLES (I.E. THREE RESPONSE CATEGORIES - 'YES', 'NO' AND 'NA' - FOR EACH ARCADIA ITEMS) | 93 |

LIST OF TABLES

| | |
|---|----|
| TABLE 1. DEFINITION OF TERMS USED IN THE PRESENT STUDY | 15 |
| TABLE 2. EXAMPLES OF DEFINITION OF SCORING SYSTEM INSTRUCTIONS | 18 |
| TABLE 3. MAIN CHARACTERISTICS OF THE INCLUDED TOOLS | 21 |
| TABLE 4. DESCRIPTIVE CHARACTERISTICS OF TOOLS USED TO ASSESS THE QUALITY OF A PEER REVIEW REPORT | 22 |
| TABLE 5. EXPLANATIONS AND EXAMPLES OF QUALITY DOMAINS AND SUBDOMAINS | 27 |
| TABLE 6. THE 20 ITEMS TO ASSESS PEER REVIEW REPORT QUALITY INCLUDED IN THE SURVEY | 36 |
| TABLE 7. SURVEY PARTICIPANTS' CHARACTERISTICS | 41 |
| TABLE 8. ITEMS LOADINGS | 46 |
| TABLE 9. SURVEY PARTICIPANTS' COMMENTS ON THE IMPORTANCE AND/OR WORDING OF THE 20 ITEMS TO ASSESS PEER REVIEW REPORT QUALITY | 47 |
| TABLE 10. NEW ITEMS SUGGESTED BY SURVEY PARTICIPANTS | 50 |
| TABLE 11. THE ARCADIA TOOL | 56 |
| TABLE 12. ARCADIA ITEMS EXPLANATION | 57 |
| TABLE 13. PSYCHOMETRIC TESTS USED IN THE PRESENT STUDY | 66 |
| TABLE 14. METHOD USED TO ASSIGN PEER REVIEW REPORTS TO SURVEY PARTICIPANTS | 73 |
| TABLE 15. DEFINITION OF QUALITY MEASURES USED IN THE PRESENT STUDY | 76 |
| TABLE 16. ENDORSEMENT FREQUENCIES FOR EACH ARCADIA ITEM | 79 |
| TABLE 17. INTERNAL CONSISTENCY OF ARCADIA TESTED ON 162 PEER REVIEW REPORTS ASSESSED BY 2 RATERS | 80 |
| TABLE 18. EFFECT OF 'RATER', 'REPORT' AND 'WORDS' ON THE ASSESSMENT OF PEER REVIEW REPORT QUALITY IN PHASE I | 81 |
| TABLE 19. INTER-RATER RELIABILITY TESTED ON 162 PEER REVIEW REPORTS ASSESSED BY 2 RATERS | 81 |
| TABLE 20. TEST-RETEST RELIABILITY TESTED ON A SUBSAMPLE OF 30 PEER REVIEW REPORTS ASSESSED BY 2 RATERS | 83 |
| TABLE 21. PARTICIPANTS' CHARACTERISTICS | 85 |
| TABLE 22. EFFECT OF 'RATER', 'REPORT' AND 'WORDS' ON THE ASSESSMENT OF PEER REVIEW REPORT QUALITY IN PHASE II | 89 |
| TABLE 23. INTER-RATER RELIABILITY TESTED ON 18 PEER REVIEW REPORTS ASSESSED BY 2 RATERS | 90 |
| TABLE 24. FACTOR LOADINGS | 92 |

ABBREVIATIONS

ARCADIA: Assessment of Review reports with a Checklist Available to eDitors and Authors

EASE: European Association of Science Editors

ICC: Intra-class correlation

ICMJE: International Committee of Medical Journal Editors

LMM: Linear mixed-effect model

MCA: Multiple Correspondence Analysis

MiRoR: Methods in Research on Research

MJA: Medical Journal of Australia

NEJM: England Journal of Medicine

OPR: Open peer review

PC: Principal Component

PCA: Principal Component Analysis

PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis

RCT: Randomized controlled trial

RoB: Risk of Bias

RoR: Research on Research

RQI: Review Quality Instrument

CHAPTER 1. INTRODUCTION

Research on Research

Research on Research (RoR), also known as meta-research and metascience, is an evolving discipline aimed to improve the quality of scientific research and reduce research waste. It is the study of research itself, covering themes from methods (i.e., how research is performed) to incentives (i.e., how research is rewarded) (1). This discipline requires a multidisciplinary approach, by combining scientists with different expertise, to “*study, promote and defend robust science*” (2).

Evidence shows that medical research is deeply flawed (3). Over the last years, the term ‘reproducibility crisis’ has been frequently used to illustrate the failure of researchers to replicate another scientist’s research and even their own. The main factors which contribute to irreproducible research have been identified into selective reporting and pressure to publish (4).

In 2009, Iain Chalmers and Paul Glasziou showed that 85% of invested effort and resources in biomedicine are wasted because of studies that are redundant, flawed in their design, never published, or poorly reported (5). About ten years later, they affirmed that “*research waste is still a scandal*” and more work is needed to improve how clinical research is conducted and reported (6).

In addition to implying a tremendous waste of resources (7), low-quality biomedical research explicitly affects public and patients’ lives. Health practitioners, consumers, public health professionals, policymakers, and health and research funding bodies rely on evidence from biomedical research to make informed health-related decisions (8). To tackle this alarming problem, the Methods in Research on Research (MiRoR) project was launched in March 2016.

The MiRoR Project

MiRoR was a joint doctoral training programme in the field of clinical research funded by Marie Skłodowska-Curie Actions (9). Its objective was to train a future generation of scientists in Research on Research and to develop creative solutions to transform clinical research practice and increase its value. The MiRoR consortium consisted of seven research teams from six different European countries, six non-academic partners, and six academic partners¹.

Furthermore, the MiRoR project involved 15 early-stage researchers conducting their PhD projects tackling different steps of the clinical research (i.e., planning, conduct, reporting, and peer review). The present PhD work is one of the three MiRoR projects contributing to the research on peer review.

The peer review process

According to the Oxford English Dictionary, the peer review process is defined as the “*evaluation of scientific, academic, or professional work by others working in the same field*” (10). It is the pivotal process of all science, from grant assignment to academic promotion (11).

Particularly, the present research focuses on the editorial peer review process, which was established to ensure that research papers are assessed by independent experts before they are published. It is a longstanding and established process aimed at providing a fair decision-making mechanism and improving the quality of a submitted manuscript (12).

¹ *Research teams*: Université de Paris, Academic Medical Centre (AMC) of the University of Amsterdam, Universitat Politècnica de Catalunya, Centre National de la Recherche Scientifique (CNRS), University of Ghent, University of Liverpool and University of Split.

Non-academic partners: European Clinical Research Infrastructure Network (ECRIN); Cochrane; National Institute for Health and Care Excellence (NICE); The British Medical Journal (The BMJ), BioMed Central (BMC) and Sideview.

Academic partners: EQUATOR network, Centre for Evidence-Based Medicine (CEBM) at the University of Oxford, University of Exeter Medical School, Université Paris Saclay, Meta-Research Innovation Center at the Stanford University (METRICS) and Ottawa Hospital Research Institute (OHRI).

How does the editorial peer review work?

Once a manuscript is ready to be submitted, the authors send it to a journal. In a basic and traditional peer review process workflow, the journal editor assesses if the manuscript meets the criteria for submission and it is in line with the journal scope. If it does, some peer reviewers, usually two referees, are selected by the editor and they are invited to review the submitted manuscript, providing their comments, and make a recommendation (i.e., accepted/major revisions/minor revisions or rejected manuscript) if it is required. Finally, the editor takes a decision on the manuscript's outcome based on these recommendations/feedback and also own judgment (Figure 1).

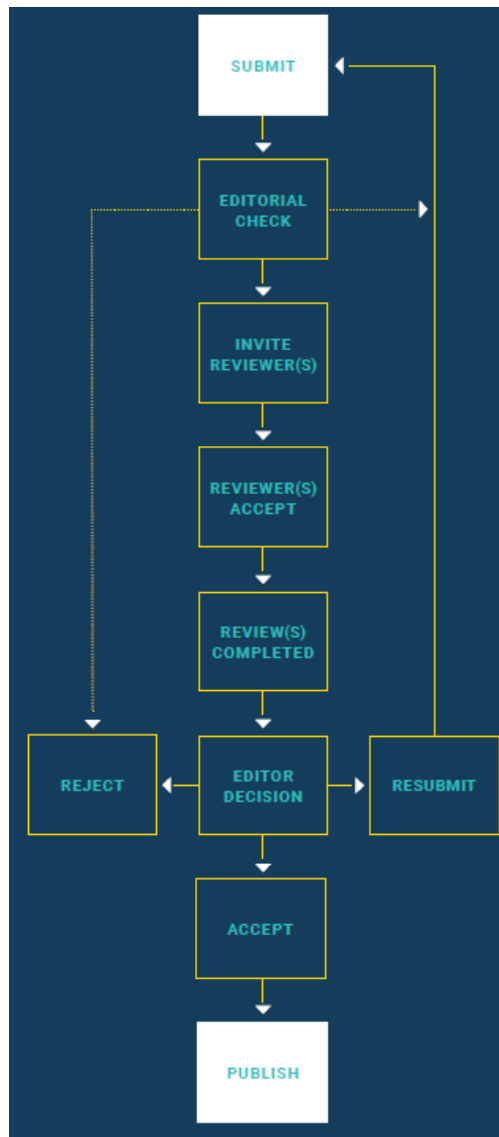


Fig. 1 Basic publishing workflow (13)

There are three peer review models commonly used among biomedical journals: 1) single-blind, 2) double-blind, 3) open-identities. In the single-blind model, the peer reviewers know the authors' identities, while the authors do not know who reviewed their manuscript (e.g., New England Journal of Medicine [NEJM]). In the double-blind model, peer reviewers and authors' identities are both blinded (e.g., Medical Journal of Australia [MJA]). Finally, peer reviewers and authors know each other identities in the open peer review model (e.g., The BMJ), and often peer review reports are published along with the manuscript.

Open peer review

In the Open Science Era, a new model of peer review has emerged. Open peer review (OPR) is defined as an “*umbrella term*” encompassing different characteristics of the peer review in relation to the Open Science goal (e.g., open identities, open reports, or open participation) (14). It is considered a way to increase transparency in research and make the peer review increasingly “*a scientific discourse rather than a summary judgement*” (15).

A few randomized controlled trials (RCTs) have been conducted to evaluate the impact of open peer review interventions, such as revealing peer reviewer identities to other peer reviewers (16,17) or to the authors of the manuscript under review (18–20), compared to the anonymized procedure. A meta-analysis combining these studies found that the quality of peer review reports, measured by scales such as the Review Quality Instrument (RQI) or editor routine quality rating scales, increases (SD=0.14, 95% CI=0.05-0.24) using an open peer review approach.

However, a recent study (2019) showed that publishing peer review reports does not affect the willingness to review, type of recommendations, and turnaround time of reviewers, at least when their anonymity is guaranteed (21). Moreover, results from a survey of 3062 authors, reviewers, and editors found that the great majority of participants (60.3%) are in favour of

using OPR as a mainstream practice, but more than half of them (52%) agreed or strongly agreed that reviewers are less likely to review for journals publishing their peer review reports (22). It is therefore needed to conduct further research to better understand the level of transparency and openness that journals could adopt in the Open Science Era (21).

The history of the peer review process

The use of editorial peer review stretched back to the eighteenth century (23). Its provenience has been commonly attributed to the foundation of the Philosophical Transactions in 1665 (24). A first example of “*society’s editorial policy and objectives*” was stated in the first volume of “Medical Essays and Observations” published in 1731 by the Royal Society of Edinburgh (23). The editorial peer review was institutionalized and became more common just after the Second War World (25).

An important analysis looking at the merits and defects in the editorial peer review in health research has been provided by Stephen Lock, long time editor of The BMJ, with the publication of the book “A difficult balance: Editorial peer review in medicine” in 1985 (26). In 1989 the first international Peer Review Congress was held in Chicago and for that occasion, Drummond Rennie, chair of the advisory board of the congress, affirmed:

“there are scarcely any bars to eventual publication. There seems to be no study too fragmented, no hypothesis too trivial, no literature citation too biased or too egotistical, no design too warped, no methodology too bungled, no presentation of results too inaccurate, too obscure, and too contradictory, no analysis too self-serving, no argument too circular, no conclusions too trifling or too unjustified, and no grammar and syntax too offensive for a paper to end up it print” (27)

About 30 years later, that statement is still considered true (28), and therefore new studies into this process are strongly encouraged for the upcoming congress (29) planned for September 2022.

Actors of the peer review process

Journal editors and peer reviewers, defined as “*custodians of high-quality science*” (30), are central actors of this process.

Journal editors

Editors are responsible for deciding which articles publish in their journals. They are asked for ensuring that what they disseminate is of the highest quality possible, following specific ethical principles, as those highlighted by the Declaration of Helsinki in 2013 (31), and recommendations (e.g., International Committee of Medical Journal Editors [ICMJE] recommendations). However, it has been shown that most biomedical editors work on a voluntary basis (3), as an extra job to their academic and/or clinical work, and operate mostly without formal training. The consequences of deciding to publish low-quality manuscripts impact future research and health-related decisions (32).

Moreover, it was found that no consensus existed on the competencies required for a biomedical editor (32) and therefore, a minimum set of core competencies was developed by Moher et al. as “*a baseline of the knowledge, skills and characteristics needed*” for biomedical editors (33). The authors identified 14 key core competencies divided into three major areas: 1) Editors’ qualities and skills, 2) Publication ethics and research integrity and, 3) Editorial principles and processes. The endorsement of this minimum set of core competencies among biomedical journals is encouraged by the authors to increase and guarantee a higher quality of scientific publications.

Peer reviewers

Peer reviewers are requested to write a review report evaluating the submitted manuscript. A peer review report helps authors improve the quality of their manuscripts, and it also helps editors make an informed decision about the outcome of the manuscript. Evidence shows that there is a need to improve the quality of peer review reports (34,35).

Peer reviewers, most volunteers, rarely receive formal training (11), and their capacity to detect errors (36,37), identify deficiencies in reporting (38) and spin (39) has been found lacking. Moreover, it has been shown that they sometimes require changes, such as additional analyses that were not pre-defined in the trial protocol, which might be considered inappropriate (40).

Chauvin et al. identified 36 tasks expected of peer reviewers when evaluating a RCT report. The authors then surveyed peer reviewers to classify the importance of each task relative to the other tasks (41). They found that the tasks for peer reviewers which were considered the most important by the survey participants (i.e., to evaluate the methodology, statistics, and results) were not congruent with the tasks required by editors in their guidelines. While key core competencies have been developed for biomedical editors, no consensus exists on which tasks and roles are needed for peer reviewers of biomedical journals.

A scoping review (2019) found that a large number of tasks and roles are expected to be performed by peer reviewers (42). According to a recent qualitative study (2019), editors' perspectives on the roles and tasks of peer reviewers are very influenced by their journal's context and characteristics, "*including also financial and human resources and journal reputation and prestige*". However, most of the editors agreed on the expected technical tasks of peer reviewers related to the scientific aspects of a manuscript (43).

As recommended by David Moher and Doug Altman in 2015, key core competencies for reviewers are strongly needed "*to help to improve the medical research literature*" (3). More

research is therefore needed to identify key core competencies for peer reviewers in biomedical research.

Criticisms of the peer review process

Despite the long history and employment of the peer review process, its impact is still considered a topic of controversy (11,44–47).

It has been shown that the research on peer review greatly increased from 2005, however only through small-scale research projects (48). Little evidence is available on the efficacy of this process, while several studies show its flaws and weaknesses. Particularly, it has been shown that it is a slow process that is often biased and easily abused. Although peer review is mostly performed as a voluntary service by researchers, it is also considered an expensive business (11). It has been estimated that the value of voluntary peer review services provided per year is around £ 1.9 billion, with 15 million hours wasted through redundancy in the reject-resubmit cycle each year (49).

In 2007 a Cochrane review on the efficacy of editorial peer review was published and little evidence was found to support the use of editorial peer review as a mechanism to ensure the quality of biomedical research (35). A more recent systematic review (2016) investigated the impact of interventions to improve the quality of peer review, such as providing training to reviewers (50,51) or adding a statistical peer reviewer (52), and it showed a lack of RCTs assessing those interventions (34). Moreover, a vignette survey (2019) shows that there is a gap between the study designs for those interventions preferred by experts (e.g., cluster RCTs and interrupted series analysis) and the designs actually utilized (53). The authors stated that well-performed trials are strongly needed to assess interventions to improve the peer review process.

However, before starting to evaluate these interventions, it is essential to clarify the outcomes (such as, for example, the quality of peer review reports) and outcome measures, which should be used in well-performed trials (34). A validated tool is direly needed to clearly define the quality of a peer review report in biomedical research. This tool could be used regularly by editors to evaluate the reviewers' work, and also as an outcome when evaluating interventions to improve the peer review process.

Goal and thesis structure

The final aim of the present PhD project was to develop and validate a new tool for assessing the quality of peer review reports in biomedical research. The present work is organized as follows:

In chapter 2, we carry out a systematic review to identify and describe the existing tools used to assess peer review report quality in biomedical research. This research was published in the *BMC Medical Research Methodology* (54).

In chapter 3, we report the development of ARCADIA, a new tool for assessing peer review report quality. Firstly, we formed a steering committee composed of five members with different expertise, which defined the quality of peer review reports. Secondly, we conducted an online survey intended for biomedical editors and authors to 1) determine if participants endorsed the proposed definition of peer review report quality; 2) identify the most important items to include in the tool, and 3) identify any missing items. Finally, based on the participants' qualitative and quantitative answers, the steering committee modified the initially proposed definition of peer review report quality, reviewed all items and, ultimately, drafted and refined ARCADIA. This research was published in *BMJ Open* (55).

In chapter 4, we evaluate the psychometric properties of the newly developed tool. We are planning to submit this research to a peer-reviewed journal in the upcoming months.

Finally, in chapter 5, we discuss the overall results of the present PhD project and describe new lines of research.

Chapters 2 and 3 are based on published papers. In each of these chapters, we made some changes, compared to the original published reports, by reducing the 'Background' and 'Discussion' sections, which would be redundant in the context of this thesis, and including

additional information in the ‘Methods’ and ‘Results’ sections. Chapter 5 is partially based on the ‘Discussion’ sections of both published papers. All the data supporting the conclusions of each published study is publicly available in the Zenodo repository in the MiRoR community (<https://zenodo.org/communities/miror/?page=1&size=20>).

CHAPTER 2. TOOLS USED TO ASSESS THE QUALITY OF PEER REVIEW REPORTS: A METHODOLOGICAL SYSTEMATIC REVIEW

This chapter is based on the following published research paper:

Superchi C, González JA, Solà I, Cobo E, Hren D, Boutron I. Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC Med Res Methodol.* 2019;19(1):48. doi:10.1186/s12874-019-0688-x

The dataset and R codes supporting the conclusions of the present study are available in the Zenodo repository in the Methods in Research on Research community doi:10.5281/zenodo.3608685

BACKGROUND

A validated tool that clearly defines peer review report quality in biomedical research is greatly needed. This will allow researchers to have a structured instrument to evaluate the impact of interventions aimed at improving the peer review process in well-performed trials. Such a tool could also be regularly used by editors to evaluate the work of peer reviewers. Herein, as starting point for the development of a new tool, we identify and describe existing tools that assess the quality of peer review reports in biomedical research.

METHODS

Study design

We conducted a methodological systematic review and followed the standard Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines (56). The quality of peer review reports is an outcome that in the long term is related to clinical relevance and patient care. However, the protocol was not registered in PROSPERO, as this review does not contain direct health-related outcomes (57).

Information sources and search strategy

We searched PubMed, EMBASE (via Ovid) and The Cochrane Methodology Register (via The Cochrane Library) from their inception to October 27, 2017 as well as Google® (search date: October 20, 2017) for all reports describing a tool to assess the quality of a peer review report in biomedical research. Search strategies were refined in collaboration with an expert methodologist (IS) and are presented in the Appendix 1. We hand-searched the citation lists of included papers and consulted a senior editor with expertise in editorial policies and peer review processes to further identify relevant reports.

Eligibility criteria

We included all reports describing a tool to assess the quality of a peer review report. Sanderson and colleagues defined a tool as ‘any structured instrument aimed at aiding the user to assess the quality [...]’ (58). Building on this definition, we defined a quality tool as any structured or unstructured instrument assisting the user to assess the quality of peer review report (for definitions see Table 1). We restricted inclusion to the English language.

Table 1. Definition of terms used in the present study

| |
|--|
| Structured quality tool: scale or checklist including more than one item aimed at guiding the user to assess the overall quality of a peer review report. |
| Unstructured quality tool: scale or checklist including only one item inquiring the overall quality of a peer review report. |
| Items: elements of a scale or checklist representing a component of peer review report quality. Items in a scale could or could not have an attached numerical score. If there is no attached score, these items provide the evaluator with a guidance to assess the overall quality of a peer review report. |
| Overall quality score in a scale is measured as: Sum of scores: score obtained by summing all scores for each item present in a scale. Mean of scores: score obtained by dividing the sum of scores for each item with the total number of items included in the tool. Single score: score obtained in those scales based on a single item. Summary score: score obtained in those scales with more than one item deriving from a question inquiring the overall quality of peer review report. |

Study selection

We exported the references retrieved from the search into the reference manager Endnote X7 (Clarivate Analytics, Philadelphia, United States), which was subsequently used to remove duplicates. We reviewed all records manually to verify and remove duplicates that had not been previously detected. A reviewer (CS) screened all titles and abstracts of the retrieved citations. A second reviewer (JAG) carried out quality control on a 25% random sample obtained using the statistical software R 3.3.3 (59). We obtained and independently examined the full-text

copies of potentially eligible reports for further assessment. In the case of disagreement, consensus was determined by a discussion or by involving a third reviewer (DH). We reported the result of this process through a PRISMA flowchart (56). When several tools were reported in the same article, they were included as separate tools. When a tool was reported in more than one article, we extracted data from all related reports.

Data extraction

General characteristics of tools

We designed a data extraction form using Google® Forms and extracted the general characteristics of the tools. We determined whether the tool was scale or checklist. We defined a tool as a scale when it included a numeric or nominal overall quality score while we considered it as a checklist when an overall quality score was not present. We recorded the total number of items (for definitions see Table 1). For scales with more than 1 item we extracted how items were weighted, how the overall score was calculated, and the scoring range. Moreover, we checked whether the scoring instructions were adequately defined, partially defined, or not defined according to the subjective judgement of two reviewers (CS and JAG) (an example of the definition for scoring instructions is shown in Table 2). Finally, we extracted all information related to the development, validation, and assessment of the tool's reliability and if the concept of quality was defined.

Two reviewers (CS and JAG) piloted and refined the data extraction form on a random 5% sample of extracted articles. Full data extraction was conducted by two reviewers (CS and JAG) working independently for all included articles. In the case of disagreement, consensus was obtained by discussion or by involving a third reviewer (DH). Authors of the reports were contacted in cases where we needed further clarification of the tool.

Quality components of the peer review report considered in the tools

We followed the systematic multi-step approach recently described by Gentles (60), which is based on a constant comparative method of analysis developed within the Grounded Theory approach (61). Initially, a researcher (CS) extracted all items included in the tools and for each item identified a ‘key concept’ representing a quality component of peer review reports. Next, two researchers (CS and DH) organized the key concepts into a domain-specific matrix (analogous to the topic-specific matrices described by Gentles). Initially, the matrix consisted of domains for peer review report quality, followed by items representative of each domain and references to literature sources that items were extracted from. As the analysis progressed, subdomains were created and the final version of the matrix included domains, subdomains, items and references.

Furthermore, we calculated the proportions of domains based on the number of items included in each domain for each tool. According to the proportions obtained, we created a domain profile for each tool. Then, we calculated the matrix of Euclidean distances between the domain profiles. These distances were used to perform the hierarchical, complete-linkage clustering analysis, which provided us with a tree structure that we represent in a chart. Through this graphical summary, we were able to identify domain similarities among the different tools, which helped us draw our analytical conclusions. The calculations and graphical representations were obtained using the statistical software R 3.3.3 (59).

Table 2. Examples of definition of scoring system instructions

| Scoring system instructions | | |
|---|--|---|
| Defined | Partially defined | Not defined |
| <p>5 (Exceptional) = The rare outstanding critique that is comprehensive, objective, and insightful. Evaluates purpose of the study, study design, scientific validity, and conclusions by numbering questions and constructive suggestions to be addressed by the author. Includes comments to the editor about whether this is something new and important and useful to our readers.</p> <p>4 (Very good) = Excellent review indicating that the paper was carefully evaluated. Helpful comments to the author and editor with well-documented reasons for decision.</p> <p>3 (Good) = Useful type of very satisfactory review. Analysis not as well organized, documented, or as complete as above but is reasonable, with adequate comments for the authors.</p> <p>2 (Below average) = Very brief, superficial evaluation. Reasons for the decision not explained and comments to authors not helpful.</p> <p>1 (Unacceptable) = Such a poor review that consideration should be given to not sending further papers to this reviewer. Reasons could include evidence of bias, unfair, faulty reasoning, or evaluation (totally disagrees with the opinion of other reviewers and editor) and comments to author either absent, inappropriate, or inadequate to explain how the paper was rated.</p> <p>(Landkroon 2006) (62)</p> | <p>1 (Poor) = Does not follow reviewer guideline structure or preferred formatting in providing comments; unfavourable timeliness.</p> <p>2 (Acceptable) = Comments are somewhat helpful; review meets timeline.</p> <p>3 (Reliable) = Thorough and helpful comments; timely submission.</p> <p>4 (Excellent) = Very strong and detailed comments; review was submitted early or on time; comments enhance the manuscript's merit and relevance in the field.</p> <p>(Rajesh 2013) (63)</p> | <p>1 = poor; 2 = fair; 3 = good; 4 = excellent</p> <p>(Friedam1995) (64)</p> |

RESULTS

Study selection and general characteristics of reports

The screening process is summarized in a flow diagram (Figure 2). Of the 4312 records retrieved, we finally included 46 reports: 39 research articles; 3 editorials; 2 information guides; 1 was a letter to the editor and 1 study was available only as an abstract (excluded studies are listed in Appendix 2; included studies are listed in Appendix 3).

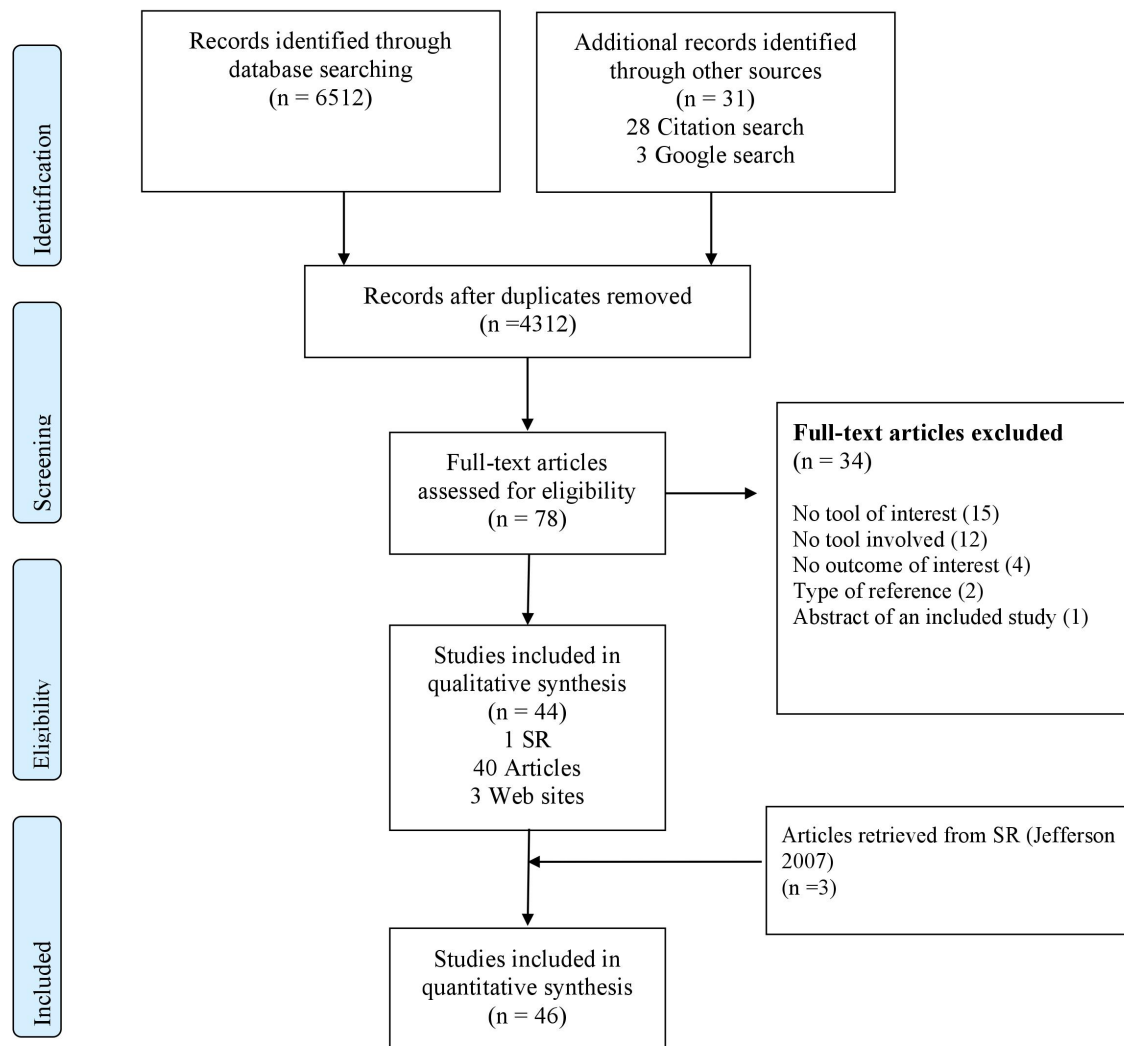


Fig. 1. Study selection flow diagram

General characteristics of the tools

In the 46 reports, we identified 24 tools, including 23 scales and 1 checklist. The tools were developed from 1985 to 2017. Four tools had from 2 to 4 versions (64–67). Five tools were used as an outcome in a RCT (18,65,67–69). Table 3 lists the general characteristics of the identified tools. Table 4 presents a more complete descriptive summary of the tools' characteristics, including types and measures of validity and reliability.

Six scales consisted of a single item enquiring into the overall quality of the peer review report, all of them based on directly asking users to score the overall quality (63,64,67,70–72). These tools assessed the quality of a peer review report by using: 1) a 4 or 5 Likert point scale (n=4); 2) as ‘good’, ‘fair’ and ‘poor’ (n=1); and 3) a restricted scale from 80 to 100 (n=1). Seventeen scales and one checklist had several items ranging in number from 4 to 26. Of these, 10 used the same weight for each item (18,65,66,69,73–78). The overall quality score was the sum of the score for each item (n=3); the mean of the score of the items (n=6); or the summary score (n=11) (for definitions see Table 1). Three scales reported more than one way to assess the overall quality (65,66,78). The scoring system instructions were not defined in 67% of the tools.

None of the tools reported the definition of peer review report quality, and only one described the tool development (79). The first version of this tool was designed by a development group composed of four researchers and three editors. It was based on a tool used in an earlier study and that had been developed by reviewing the literature and interviewing editors. Successively, the tool was modified by rewording some questions after some group discussions and a guideline for using the tool was drawn up.

Only 3 tools assessed and reported a validation process (79–81). The assessed types of validity included face validity, content validity, construct validity, and preliminary criterion validity. Face and content validity could involve either a sole editor and author or a group of researchers and editors. Construct validity was assessed with multiple regression analysis using discriminant criteria (reviewer characteristics such as age, sex, and country of residence) and convergent criteria (training in epidemiology and/or statistics); or the overall assessment of the peer review report by authors and an assessment of (n=4-8) specific components of the peer review report by editors or authors. Preliminary criterion was assessed by comparing grades obtained by an editor to those obtained by an editor-in-chief using an earlier version of the tool. Reliability was assessed in 9 tools (18,62,67,68,71,78,79,81,82); all reported inter-rater

reliability and 2 also reported test-retest reliability. One tool reported the internal consistency measured with the Cronbach's alpha (79).

Table 3. Main characteristics of the included tools

| Characteristics of tools | N (%) |
|--|-----------------------|
| Type of tool: | |
| Scale | 23 (96%) |
| Checklist | 1 (4%) |
| Number of items: | |
| 1 | 6 (25%) |
| >1 | 18 (75%) |
| Weight of items ^a: | |
| Same weight | 10 (42%) |
| Different weight | 2 (8%) |
| User defined weight | 1 (4%) |
| Not applicable | 11 (46%) ^a |
| Score System Instruction: | |
| Defined | 5 (21%) |
| Partially defined | 3 (12%) |
| Not defined | 16 (67%) |
| Tool development: | |
| Reported | 1 (4%) |
| Not reported | 23 (96%) |
| Overall quality assessment ^b | |
| Single score | 6 (22%) |
| Summary score | 11 (41%) |
| Mean score | 6 (22%) |
| Sum score | 3 (11%) |
| Not reported | 1 (4%) |

^a Item weight is not applicable for scale with a single item (n=6), checklist (n=1) and for scale including more than one item without a numerical score attached but presenting only a summary score (n=4)

^b The total number is different because three tools presented more than one way to assess the overall quality and the checklist did not provide an overall score

Table 4. Descriptive characteristics of tools used to assess the quality of a peer review report

| Journal or Company Name ^a | First Author, Year | Format | Quality defined ^b | Overall quality assessment | Items (n) | Items weights ^c | Scoring range ^d | Scoring system instruction ^e | Scale/ Checklist Development ^f | Validity ^g | Reliability ^h | Internal consistency | RCTs ⁱ |
|---|------------------------------|--------|------------------------------|----------------------------|-----------|----------------------------|----------------------------|---|---|-----------------------|---|----------------------|-------------------|
| Advances in Nursing Science; Issues in Mental Health Nursing; The Journal of Holistic Nursing | Shattell 2010 (75) | Scale | N | Summary Score | 6 | S | 1-10 | N | NR | NR | NR | NR | 0 |
| American Journal of Roentgenology | Friedman 1995 (64) | Scale | N | Single Score | 1 | NA | 1-4 | N | NR | NR | NR | NR | 0 |
| American Journal of Roentgenology | Kliwer 2005 (83) | Scale | N | Summary Score | 4 | NA | 1-4 | N | NR | NR | NR | NR | 0 |
| American Journal of Roentgenology | Rajesh 2013 (63) | Scale | N | Single Score | 1 | NA | 1-4 | P | NR | NR | NR | NR | 0 |
| American Journal of Roentgenology | Berquist 2017 (84) | Scale | N | Summary Score | 4 | NA | 0-4 | Y | NR | NR | NR | NR | 0 |
| Annals of Emergency Medicine | Callaham 1998 (67) | Scale | N | Single Score | 1 | NA | 1-5 | N | NR | NR | Inter-Rater (ICC=0.44, 0.24, 0.12) ¹ | NR | 2 ^m |
| Annals of Emergency Medicine | Callaham 2002 (68,85) | Scale | N | Summary Score | 6 | NA | 1-5 | N | NR | NR | Inter-Rater (ICC=0.44, 0.24, 0.12) ¹ | NR | 1 |
| Annals of Emergency Medicine; Annals of Internal Medicine; | Justice 1998 (77) | Scale | N | Summary Score | 4 | S | 1-5 | N | NR | NR | NR | NR | 0 |

| | | | | | | | | | | | | | | |
|---|---|-----------|---|------------------------------|---|----|------------------------|---|----|---|--|--|--|---|
| JAMA; Obstetrics & Gynecology and Ophthalmology | | | | | | | | | | | | | | |
| British Journal of General Practice | Moore 2014 (70) | Scale | N | Single Score | 1 | NA | A-E | Y | NR | NR | NR | | | 0 |
| British Medical Journal | Black 1998 (RQI 3.2) (65,79) | Scale | N | Summary Score Mean | 7 | S | 1-5 | N | Y | Face (N=20) Content (N=20) Construct | Test-Retest (Kw=1.00) Inter-Rater (Kw=0.83) | Internal Consistency (Cronbach's alpha= 0.84) | | 5 |
| British Medical Journal | Van Rooyen 1999 (RQI 4) (18) | Scale | N | Mean ⁿ | 8 | S | 1-5 | N | NR | NR | Inter-Rater (Kw=0.38- 0.67) ^o | | | 2 |
| Chinese Journal of Tuberculosis and Respiratory Diseases | Yang 2009 (86) | Checklist | N | NA | 5 | NA | NA | N | NR | NR | NR | | | 0 |
| Journal of Clinical Investigation | Stossel 1985 (72) | Scale | N | Single Score | 1 | NA | Good- Fair- Poor | Y | NR | NR | NR | | | 0 |
| Journal of General Internal Medicine | McNutt 1990 (69,80) | Scale | N | Summary Score | 9 | S | 1-5 | N | NR | Construct | NR | | | 1 |
| Journal of Vascular Interventional Radiology | Feurer 1994 (81) | Scale | N | Sum | 7 | D | 0-14 | N | NR | Content (N=2) Preliminary Criterion (N=2) | Inter-Rater (ICC=0.84) | | | 0 |

| | | | | | | | | | | | | |
|--|---|-------|---|-----------------------------|----|----------------------|-------|---|----|------------------|--|---|
| | | | | | | | | | | (Kendall = 0.94) | | |
| NA | Review quality collector (RQC) 2012 (87) | Scale | N | Mean | 4 | User-defined weights | 0-100 | N | NR | NR | NR | 0 |
| Nursing Research | Henly 2009 (66) | Scale | N | Mean (CAS, GAS scale) | 15 | S | 1-5 | P | NR | NR | Inter-Rater (ICC=0.79) _p | 0 |
| | | | | Summary Score (OAS scale) | | | 1-5 | | | | | |
| | | | | Summary Score (GRQ scale) | | | 0-100 | | | | | |
| Nursing Research | Henly 2010 (78) | Scale | N | Mean (CAS,GAR, SARNR scale) | 26 | S | 1-5 | P | NR | NR | Inter-Rater (ICC=0.75) ^p | 0 |
| | | | | Summary Score (GRQ scale) | | | 0-100 | | | | | |
| Obstetrics & Gynecology, Dutch Journal of Medicine | Landkroon 2006 (62) | Scale | N | Summary Score | 5 | NA | 1-5 | Y | NR | NR | Test-Retest (ICC =0.66-0.88) Inter-Rater (ICC = 0.62) | 0 |
| Pakistan Journal of Medical | Jawaid 2006 (76) | Scale | N | NR ^q | 5 | S | 1-5 | N | NR | NR | NR | 0 |

| | | | | | | | | | | | | |
|---------------------------------------|--|-------|---|--------------|---|----|------------------------|---|----|----|--|---|
| Sciences | | | | | | | | | | | | |
| Peerage of science | Peerage Essay Quality (PEQ) 2011 (73) | Scale | N | Mean | 3 | S | 1-5 | N | NR | NR | NR | 0 |
| Publons Academy | Review Rating and Feedback Form 2016 (74) | Scale | N | Sum | 4 | S | 0-3 (Full score: 0-12) | N | NR | NR | NR | 0 |
| The Journal of Bone and Joint Surgery | Thompson 2016 (71) | Scale | N | Single Score | 1 | NA | 80-100 | Y | NR | NR | Inter-Rater (ICC =-4.5 to 0.99) ^r | 0 |
| The National Medical Journal of India | Das Sinha 1999 (16) | Scale | N | Sum | 5 | D | 0-100 | N | NR | NR | NR | 0 |

^a Name of journal or company/organization where the tool was used to assess the quality of their peer review reports

^b The quality of a peer review report is not clearly defined in any reports

^c NA= not applicable. S= same weight for each item; D= different weight for each item

^d NA= not applicable

^e Y= yes defined; P= partially defined; N = not defined

^{f, g, h} NR=not reported

ⁱ Number of RCTs where the tool was used as outcome criteria

^l The ICC was 0.44 for reviewers, 0.24 for editors, and 0.12 for manuscripts

^m One article consists of two studies. First study is not a RCT while the second one is a RCT (88)

ⁿ The overall quality is based on the mean of the first seven items (the item about the tone of the review was not included)

^o The inter-rater reliability was measured with weighted K for item from 1 to 7 for two editors' independent assessments

^p The tool includes more than one scale. We reported inter-rater reliability only for General Review Quality (GRQ) scale

^q Not reported. Although the authors reported that the reviewers were rated as excellent, good and average based on the quality of the reviews, it is not reported how they assessed the overall quality of peer review reports

Quality components of the peer review reports considered in the tools with more than one item

We extracted 132 items included in the 18 tools. One item asking for the percentage of co-reviews the reviewer had graded was not included in the classification because it represented a method of measuring reviewer's performance and not a component of peer review report quality.

We organized the key concepts from each item into 'topic-specific matrices' (Appendix 4), identifying nine main domains and 11 subdomains: 1) relevance of study (n=9); 2) originality of the study (n=5); 3) interpretation of study results (n=6); 4) strengths and weaknesses of the study (n=12) (general, methods and statistical methods); 5) presentation and organization of the manuscript (n= 8); 6) structure of the reviewer's comments (n=4); 7) characteristics of reviewer's comments (n=14) (clarity, constructiveness, detail/thoroughness, fairness, knowledgeability, tone); 8) timeliness of the review report (n=7); and 9) usefulness of the review report (n=10) (decision making and manuscript improvement). The total number of tools corresponding to each domain and subdomain is shown in Figure 3. An explanation and example of all domains and subdomains is provided in Table 5. Some domains and subdomains were considered in most tools, such as whether the reviewers' comments were *detailed/thorough* (n=11) and *constructive* (n=9), whether the reviewers' comments were on the *relevance of the study* (n=9) and if the peer review report was *useful for manuscript improvement* (n=9). However, other items were rarely considered, such as whether the reviewer made comments on the *statistical methods* (n=1).

Table 5. Explanations and Examples of quality domains and subdomains

| N | Domains | Subdomains | Explanations and Examples |
|---|---|---------------------|--|
| 1 | Relevance of the study | | <p><u>Explanation:</u> Items inquiring if the reviewer has discussed in the peer review report the importance of the research question and usefulness of the study.</p> <p><u>Example:</u> ‘Did the reviewer give appropriate attention to the importance of the question?’</p> |
| 2 | Originality of the study | | <p><u>Explanation:</u> Items inquiring if the reviewer has commented in the peer review report on the originality of the manuscript.</p> <p><u>Example:</u> ‘Did the reviewer discuss the originality of the paper?’</p> |
| 3 | Interpretation of the study results | | <p><u>Explanation:</u> Items inquiring if the reviewer has commented in the peer review report on how authors interpreted and discussed the results of the study.</p> <p><u>Example:</u> ‘The reviewer commented accurately and productively on the quality of the author’s interpretation of the data, including acknowledgment of the data’s limitations.’</p> |
| 4 | Strengths and weaknesses of the study | General | <p><u>Explanation:</u> Items inquiring if the reviewer has identified and commented in the peer review report on the general strong and weak points of the study.</p> <p><u>Example:</u> ‘How well it identified the study’s strengths and weaknesses?’</p> |
| | | Methods | <p><u>Explanation:</u> Items inquiring if the reviewer has identified and commented in the peer review report on the strong and weak points specifically related to study’s methods</p> <p><u>Example:</u> ‘Did the reviewer clearly identify strengths and weaknesses in the study’s methods?’</p> |
| | | Statistical methods | <p><u>Explanation:</u> Items inquiring if the reviewer has identified and commented in the peer review report on the strong and weak points specifically related to study’s statistical methods</p> <p><u>Example:</u> ‘Confidence intervals/p-values/overall fit’</p> |
| 5 | Presentation and organization of the manuscript | | <p><u>Explanation:</u> Items inquiring if the reviewer has made comments in the peer review report on the data presentation such as tables and figures and on the organization of the manuscript such as writing communication.</p> <p><u>Example:</u> ‘Are there any constructive suggestions on improvement of a. writing; b. data presentation and c. interpretation’</p> |
| 6 | Structure of reviewer’s comments | | <p><u>Explanation:</u> Items inquiring if the reviewer has made in the peer review report organized and structured comments.</p> <p><u>Example:</u> ‘Concise well-organized comments to the editor’</p> |

7 Characteristics of reviewer's comments

| | |
|---------------------|--|
| Clarity | <p><u>Explanation:</u> Items inquiring if the reviewer has provided in the peer review report clear and easily to read comments.</p> <p><u>Example:</u> 'How clear was this review? The review was easily read and interpreted by the editor and authors.'</p> |
| Constructiveness | <p><u>Explanation:</u> Items inquiring if the reviewer has provided in the peer review report helpful, relevant and realistic comments.</p> <p><u>Example:</u> 'Were the reviewer's comments constructive?'</p> |
| Detail/Thoroughness | <p><u>Explanation:</u> Items inquiring if the reviewer has provided in the peer review report detailed and thorough comments supplying appropriate evidence.</p> <p><u>Example:</u> 'Detail of commentary'</p> |
| Fairness | <p><u>Explanation:</u> Items inquiring if the reviewer has provided in the peer review report balanced and objective comments.</p> <p><u>Example:</u> 'Balanced/fair'</p> |
| Knowledgeability | <p><u>Explanation:</u> Items inquiring if the reviewer has showed in the peer review report to know and understand correctly the content of the manuscript.</p> <p><u>Example:</u> 'Knowledge of the manuscript's content area.'</p> |
| Tone | <p><u>Explanation:</u> Items inquiring if the reviewer has used a courteous tone in the peer review report.</p> <p><u>Example:</u> 'Overall tone of the reviewers was also assessed as harsh or courteous.'</p> |

8 Timeliness of the review report

Explanation: Items inquiring if the reviewer has completed the peer review report on time.

Example: 'Punctuality of the review'

9 Usefulness of the review report

Decision making

Explanation: Items inquiring if the reviewer has provided a peer review report useful to make a decision about the acceptance, revision or rejection of a manuscript

Example: 'The reviewer provided the editor with the proper context and perspective to make a decision about acceptance or revision of the manuscript.'

Manuscript improvement

Explanation: Items inquiring if the reviewer has provided useful suggestions in the peer review report to improve the manuscript.

Example: 'This aspect is solely interested in how well the review aids the authors for improving their work and/or writing. Whether the review makes a good judgment regarding acceptance of the submission plays no role here whatsoever.'

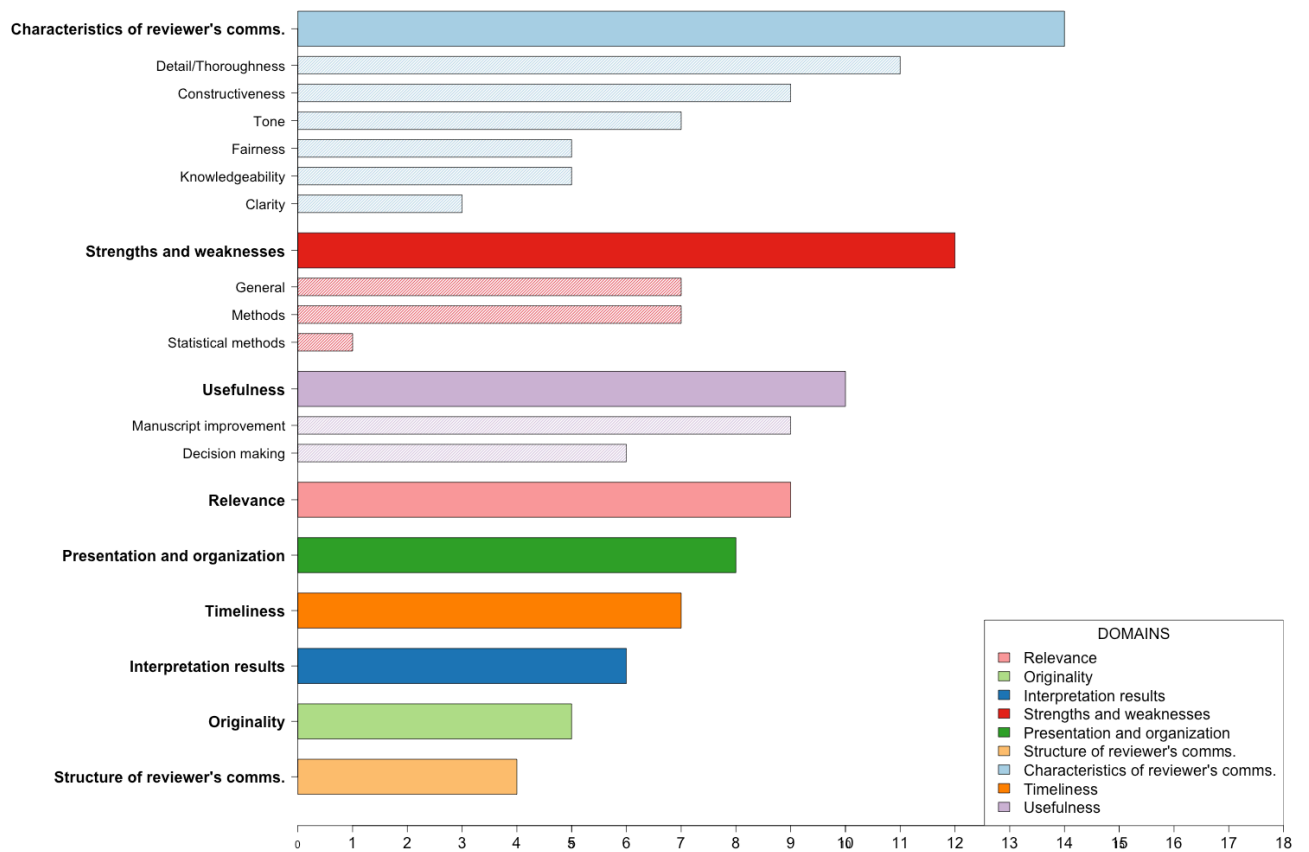


Fig. 2. Frequency of quality domains and subdomains

The abbreviation 'comms.' in the figure was used for 'comments'.

Clustering analysis among tools

We created a domain profile for each tool. For example, the tool developed by Justice et al consisted of 5 items (77). We classified three items under the domain '*Characteristics of the reviewer's comments*', one under '*Timeliness of the review report*' and one under '*Usefulness of the review report*'. According to the aforementioned classification, the domain profile (represented by proportions of domains) for this tool was 0.6:0.2:0.2 for the incorporating domains and 0 for the remaining ones. The hierarchical clustering used the matrix of Euclidean distances among domain profiles, which led to five main clusters (Figure 4).

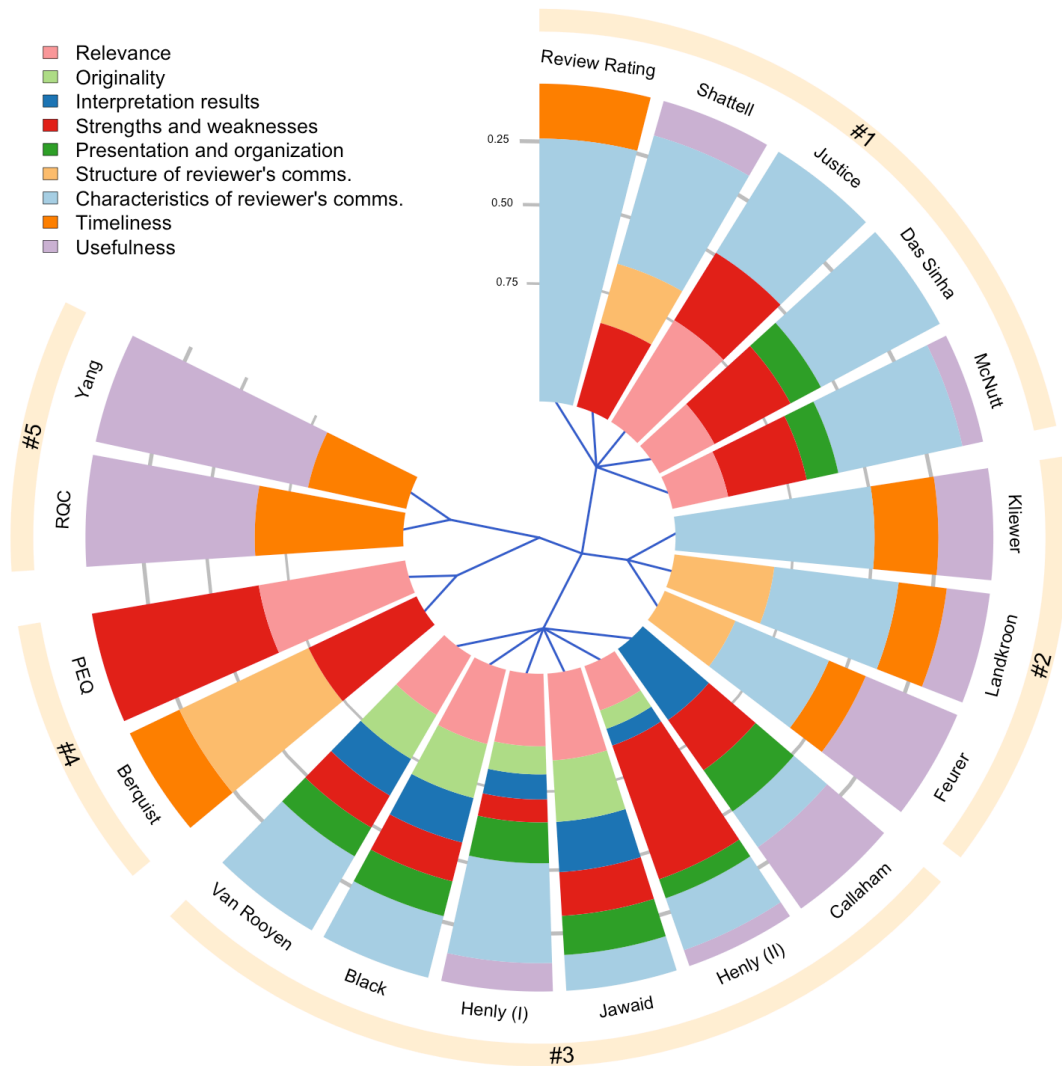


Fig. 3. Hierarchical clustering of tools based on the nine quality domains

The figure shows which quality domains are present in each tool. A slice of the chart represents a tool, and each slice is divided into sectors, indicating quality domains (in different colours). The area of each sector corresponds to the proportion of each domain within the tool. For instance, the “Review Rating” tool consists of two domains: *Timeliness*, meaning that 25% of all its items are encompassed in this domain, and *Characteristics of reviewer’s comments* occupying the remaining 75%. The blue lines starting from the centre of the chart define how the tools are divided into the five clusters. Clusters #1, #2 and #3 are sub-nodes of a major node grouping all three, meaning that the tools in these clusters have a similar domain profile compared to the tools in clusters #4 and #5. The abbreviation ‘comms.’ in the figure was used for ‘comments’.

The first cluster consisted of 5 tools developed from 1990 to 2016. All tools included at least one item in the *characteristics of the reviewer’s comments* domain, representing at least 50% of each domain profile. In the second cluster, there were 3 tools developed from 1994 to 2006. These tools were characterized to incorporate at least one item in the *usefulness* and *timeliness* domains. The third cluster included 6 tools that had been developed from 1998 to 2010 and

exhibited the most heterogeneous mix of domains. These tools were distinct from the rest because they encompassed items related to *interpretation of the study results* and *originality of the study*. Moreover, the third cluster included two tools with different versions and variations. The first, second, and third cluster were linked together in the hierarchical tree that presented tools with at least one quality component grouped in the domain *characteristics of the reviewer's comments*. In the fourth cluster, there are 2 tools developed from 2011 to 2017 that consist of at least one component in the *strengths and weaknesses* domain. Finally, the fifth cluster included 2 tools developed from 2009 to 2012 and which consisted of the same 2 domains. The fourth and fifth clusters were separated from the rest in the hierarchical tree that presented tools with only a few domains.

DISCUSSION

To the best of our knowledge, this is the first comprehensive review that has systematically identified tools used in biomedical research for assessing the quality of peer review reports. We have identified 24 tools from both the medical literature and an internet search: 23 scales and 1 checklist. Since a definition of overall quality was not provided, these tools consisted exclusively of a subjective quality assessment by the evaluators.

The present study has some limitations. Although we implemented a comprehensive search strategy for reports by following the guidance for conducting methodological reviews (60), we cannot exclude a possibility that some tools were not identified. Moreover, we limited the eligibility criteria to reports published only in English. Finally, although the number of eligible records we identified through Google® was very limited, it is possible that we introduced selection bias due to a (re)search bubble effect (89).

CONCLUSIONS

A variety of tools exist for assessing the quality of a peer review report. However, the development and validation process of those tools is not clearly reported and the concepts evaluated by these tools vary widely. The results from this study and from further investigations will inform the development of a new tool for assessing the quality of peer review reports in biomedical research.

**CHAPTER 3. THE DEVELOPMENT OF ARCADIA: A TOOL FOR
ASSESSING THE QUALITY OF PEER REVIEW REPORTS IN BIOMEDICAL
RESEARCH**

This chapter is based on the following published research paper:

Superchi C, Hren D, Blanco D, Rius R, Recchioni A, Boutron I, González JA. Development of ARCADIA: a tool for assessing the quality of peer-review reports in biomedical research. *BMJ Open* 2020;0:e035604. doi:10.1136/bmjopen-2019-035604

The anonymised survey data and codebook supporting the conclusions of the present study are available in the Zenodo repository in the Methods in Research on Research community. doi: 10.5281/zenodo.3997118

BACKGROUND

Evidence shows that there is a need to improve the quality of peer review reports in biomedical research (34,35). Tools for assessing the quality of peer review reports have been proposed, of which we have conducted a systematic review and identified 24 tools: 23 scales and 1 checklist (54). However, none reported any definition of peer review report quality, only one described the scale development, and 10 provided measures of reliability and validity. Further, the development and validation process resulted from a small consensus of people, and the concepts evaluated by these tools were quite heterogeneous. In the present study, we report on the development of a new tool to assess quality of peer review reports in biomedical research.

METHODS

The study was approved by the Research Committee of the Governing Council of the Universitat Politècnica de Catalunya, Barcelona Tech, Spain (Reference: EC 02, Date: 02/05/2018).

We followed the methods for instrument development suggested by Streiner and Norman (90) consisting of three main steps: 1) definition of the aim of the tool; 2) generating items and 3) selecting items.

Steering committee

We formed a steering committee of five members, whose expertise include clinical epidemiology, biostatistics, social science and editorial peer review². The steering committee

²The steering committee was composed by Cecilia Superchi, Darko Hren (PhD), Alessandro Recchioni, Isabelle Boutron (MD, PhD) and José Antonio Gonzalez (PhD). Darko Hren is an Assistant Professor in the Faculty of Humanities and Social Sciences at the University of Split, Croatia. Alessandro Recchioni is a Senior Editor of BMC Medicine. José Antonio Gonzalez is a Tenured Associate Professor in the Department of Statistics and Operations Research at the Universitat Politècnica de Catalunya, Spain. Isabelle Boutron is a Full Professor in Epidemiology at the University of Paris, France. All the steering committee members have a deep knowledge of peer review.

agreed on how to define peer review report quality; they agreed on the survey questionnaire based on the results of a previous systematic review (54); they interpreted the results of the survey; and they agreed on the final version of the tool.

Defining the tool's objective

The tool aims to assess the quality of peer review reports in biomedical research. We defined the quality of a peer review report as “the extent to which a peer review report helps editors make a fair decision and authors improve the quality of the submitted manuscript”.

Generating the items

A systematic review allowed identifying 24 tools aimed at assessing the quality of peer review reports (54). We extracted 132 items from these tools. All redundant items were merged and we included all items that met our definition of peer review report quality. Overall, 20 items were identified for assessing peer review report quality (Table 6).

Survey

We conducted an online survey of editors and authors in order to: 1) determine if they endorsed the proposed definition of peer review report quality; 2) identify the most important items to include in the final tool; and 3) identify any new items that should be included.

Table 6. The 20 items to assess peer review report quality included in the survey

| Labels | Items to assess PR report quality |
|-------------------------------------|---|
| Relevance | The reviewer comments on the relevance of the study |
| Originality | The reviewer comments on the originality of the study |
| Interpretation results | The reviewer comments on the interpretation of study results |
| Strengths and weaknesses (general) | The reviewer comments on the general strengths and weaknesses of the study |
| Strengths and weaknesses (methods) | The reviewer comments on the strengths and weaknesses of the study methods |
| Statistical methods | The reviewer comments on the appropriateness of the statistical methods |
| Methodological quality | The reviewer comments on the methodological quality (internal validity) of the study |
| Applicability and external validity | The reviewer comments on the applicability and external validity of the study results |
| Presentation and organization | The reviewer comments on the presentation and organization of the manuscript |
| Adherence to reporting guideline | The reviewer comments on the adherence of the manuscript to the reporting guideline |
| Structure of reviewer's comments. | The reviewer's comments are structured and organized |
| Clarity | The reviewer's comments are clear and easy to read |
| Constructiveness | The reviewer's comments are constructive |
| Detail/Thoroughness | The reviewer's comments are detailed and thorough |
| Objectivity | The reviewer's comments are objective |
| Fairness | The reviewer's comments are fair |
| Support by evidence | The reviewer's comments are evidence based |
| Knowledgeability | The reviewer knows and understands correctly the content of the manuscript |
| Tone | The reviewer uses a courteous tone |
| Timeliness | The reviewer completes the peer review report on time |

Survey questionnaire

The questionnaire was constructed using the online survey software SurveyMonkey (91). This software allowed us to send the invitation email to participants and track how many people

opened the invitation email, clicked through to the survey, responded to the survey, or opted out of the survey. This information was tied to the invitation email, not to the data entered by the survey participants.

It was structured into four main parts and included both open and multiple-choice questions. First, the participants were asked to agree (“yes/no/partially”) on the definition we provided for peer review report quality. They were also invited to add any comments or ideas on how to improve the definition. Second, they were asked to rate the importance of the 20 items for assessing the quality of peer review reports we identified. Their responses were based on a 1–5 Likert scale (1 being not important and 5 very important). In particular, we asked the participants if the item should be included in a tool for assessing the quality of peer review reports. Moreover, they were invited to comment on the importance and wording of each item. In order to eliminate the question order effect, the items appeared in random order for each respondent. Third, the participants were invited to suggest any additional items missing that they considered important for assessing the quality of peer review reports. Finally, the questionnaire included nine demographic questions related to sex, age, education level, job title, referring institution and job experience as biomedical editor and/or author. We developed two versions of the questionnaire because biomedical editors and authors were recruited differently, despite the fact that some of them could play both roles (see Appendix 5). The two versions were structured in the same way; they only differed in some questions related to the demographic characteristics. The questionnaire was piloted among six experienced scientific editors and authors, followed by a subsequent revision based on their feedback.

Participants and recruitment strategy

We targeted biomedical editors and authors using a purposive sampling approach to recruit a heterogeneous sample of information-rich cases (92).

Biomedical editors

By means of standardized email (see Appendix 6), we invited two groups of editors to participate in the survey: 586 biomedical editors from 43 journals in the BMJ Publishing group; and 478 editors from 235 journals identified in a previous cross-sectional bibliometric study (93). The survey was also distributed to 27 editors from 48 journals in BMC (part of Springer Nature), using internal email, and to members of the European Association of Science Editors (EASE) through their newsletter. In the invitation email and newsletter, the editors were encouraged to forward the survey to colleagues who might be interested in issues related to peer review. This recruitment strategy, known as snowballing, allowed us to identify “information-rich key informants” (92). On the first page of the survey, participants were informed that the collected data would be anonymous, and they were further asked if they would agree to share their de-identified data in an open access repository. Two reminder emails were sent at 2-week intervals to non-respondents. Finally, the survey was promoted on Twitter and on the EASE blog (94) and Methods in Research on Research (MiRoR) (95) website.

Authors

Searching the top 30 biomedical journals with the highest Journal Impact Factors (see Appendix 7), we identified 4396 corresponding authors of articles that reported original research and which were published in MEDLINE between February 1 and October 31 2018. We used the R package easyPubMed to extract the email contacts (96). The corresponding authors received a standardized email that explained the purpose of the study and included a link to the survey (see Appendix 6). The first page of the survey informed participants that the data were collected anonymously and also asked if they would agree to share their de-identified data in an open access repository. Two reminder emails were sent at 2-week intervals to non-respondents.

Data analysis

We described the demographic data in terms of frequencies and percentages. The importance of the 20 items to assess peer review report quality was described in means and proportions of editors or authors who rated the importance of the items from 1 to 5. The items were also sorted according to the mean ranking of all participants and either editors or authors. We also calculated Pearson correlations among items. The calculations and graphical representations were all obtained using the statistical software R 3.5.0 (59).

Principal component analysis of quantitative data

We conducted a principal component analysis (PCA) to examine item redundancy among the 20 items to assess peer review report included in the survey. PCA is a multivariate statistical technique used to reduce the number of variables in a dataset to a smaller number of dimensions (97). The new dimensions (or *principal components*) are mutually independent and are determined by choosing the directions that explain the most variation in the data. The first principal component (PC1) accounts for the largest possible variance in the data, and each succeeding PC accounts for decreasing amounts of the remaining. This exploratory analysis helps reveal simple underlying structures in complex datasets. We performed PCA using the R package FactoMineR (98).

Inductive content analysis of qualitative data

We used a general inductive approach for qualitative data analysis. In particular, we followed the five steps of inductive analysis proposed by David R. Thomas: 1) Preparation of raw data files; 2) Close reading of text; 3) Creation of codes; 4) Overlapping coding and uncoded text; 5) Continuing revision and refinement of themes system (99). In the third phase, two investigators (CS and DB) created independently the initial codes from the responses of the first 100 participants for each open-ended question. In order to ensure consistency and credibility, the initial codes were discussed with a third investigator (DH) and a codebook was developed and

was used for analysing the remaining responses. In case new codes were successively created from the remaining responses, the emerging codes were added to the codebook and applied to entire dataset. Two investigators (CS and DH) reviewed and refined the codebook and further clustered the codes into major themes. We used the software NVivo V.12 for data management and analysis (100).

Selecting items

The steering committee reviewed all items and, ultimately, drafted and refined the final version of the tool. Based on the participants' qualitative and quantitative answers, redundant items were combined, existing items were modified and/or expanded on, and new items proposed by survey participants were added.

RESULTS

Participants

Between November 7 2018 and February 4 2019, 198 biomedical editors and 248 authors completed the survey. Participants were mainly male (263/399, 65.9%) with a PhD degree (225/399, 56.4%), and their ages were equally distributed across ranges (mean=50.3, SD=13). They were mainly located in Europe (219/389, 56.3%) and North America (118/389, 30.3%). More than half of the editors had editorial work experience of more than 5 years (91/165, 55.2%), while over one-third of the authors had work experience of more than 20 years (84/224, 37.5%) (see Table 7 and Appendix 8).

Table 7. Survey participants' characteristics

| Characteristics | Editors N=198 | Authors N=248 | Total N=446 |
|---|------------------|------------------|----------------|
| Gender | N=169 | N=230 | N=399 |
| Woman | 46 (27.2%) | 83 (36.1%) | 129 (32.3%) |
| Man | 121 (71.6%) | 142 (61.7%) | 263 (65.9%) |
| Other | 2 (1.2%) | 5 (2.2%) | 7 (1.8%) |
| Age | N=156 | N=220 | N=376 |
| <40 | 32 (20.5%) | 71 (32.3%) | 103 (27.4%) |
| 41-50 | 29 (18.6%) | 59 (26.8%) | 88 (23.4%) |
| 51-60 | 52 (33.3%) | 37 (16.8%) | 89 (23.7%) |
| >60 | 43 (27.6%) | 53 (24.1%) | 96 (25.5%) |
| Education | N=169 | N=230 | N=399 |
| Bachelor Degree | 4 (2.4%) | 3 (1.3%) | 7 (1.8%) |
| Master Degree | 11 (6.5%) | 20 (8.7%) | 31 (7.8%) |
| PhD | 107 (63.3%) | 118 (51.3%) | 225 (56.4%) |
| M.D. or equivalent | 34 (20.1%) | 76 (33.0%) | 110 (27.6%) |
| Prefer not to answer | 2 (1.2%) | 1 (0.4%) | 3 (0.8%) |
| Other | 11 (6.5%) | 12 (5.2%) | 23 (5.8%) |
| Location journal/institution | N=165 | N=224 | N=389 |
| Europe | 132 (80.0%) | 87 (38.8%) | 219 (56.3%) |
| North America | 23 (13.9%) | 95 (42.4%) | 118 (30.3%) |
| South America | 2 (1.2%) | 5 (2.2%) | 7 (1.8%) |
| Africa | 1 (0.6%) | 1 (0.4%) | 2 (0.5%) |
| Asia | 3 (1.8%) | 11 (4.9%) | 14 (3.6%) |
| Australia | 4 (2.4%) | 25 (11.2%) | 29 (7.5%) |
| Number of years of work experience | N=165 | N=224 | N=389 |
| <5 years | 74 (44.8%) | 36 (16.1%) | 110 (28.3%) |
| 6-10 years | 46 (27.9%) | 51 (22.8%) | 97 (24.9%) |
| 11-15 years | 27 (16.4%) | 34 (15.2%) | 61 (15.7%) |
| 16-20 years | 7 (4.2%) | 19 (8.5%) | 26 (6.7%) |
| >20 years | 11 (6.7%) | 84 (37.5%) | 95 (24.4%) |

Definition of peer review report quality

Overall 84% (362/431) participants, precisely 85% (160/188) editors and 83% (202/243) authors, agreed on the definition of peer review report quality that we provided in the survey. The definition was slightly modified to take into account participants' comments (the complete codebook is available in the Zenodo repository). The quality of a peer review report is now defined 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”.

Quantitative results

We created a web application that is publicly available at <https://www-eio.upc.edu/redir/ReportQuality> (Figure 5). Through the application, the readers can easily access and explore the quantitative results of the survey.

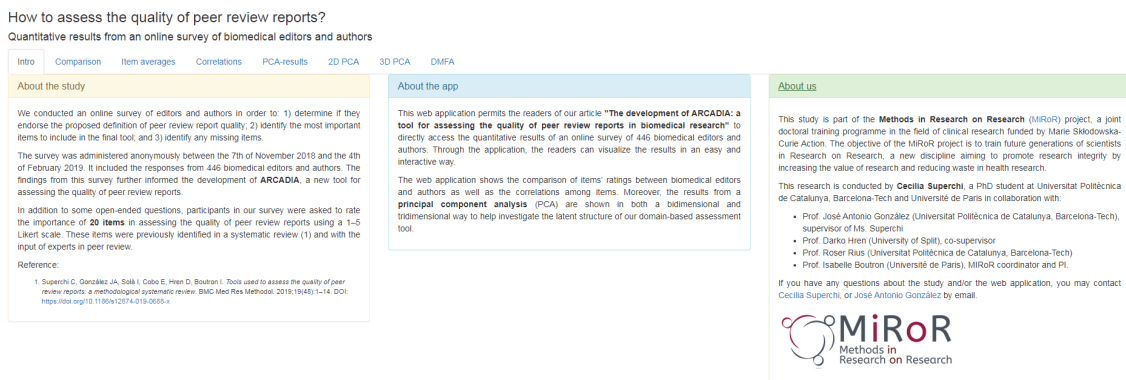


Fig. 4. Shiny application welcome page

Rating the importance of items

Figure 6 shows the 20 items rated by editors and authors, listed in order of all participants' mean ranking. The items were generally highly rated, with a mean score ranging from 3.38 (SD=1.13) to 4.60 (SD=0.69). All the items were scored 4 or 5 by >50% of the participants. The three items rated as the most important were: 1) *Knowledgeability*; 2) *Methodological quality*; and 3) *Fairness*. The three least important items were: 1) *Originality*, 2) *Presentation and organization*; and 3) *Adherence to reporting guideline*.

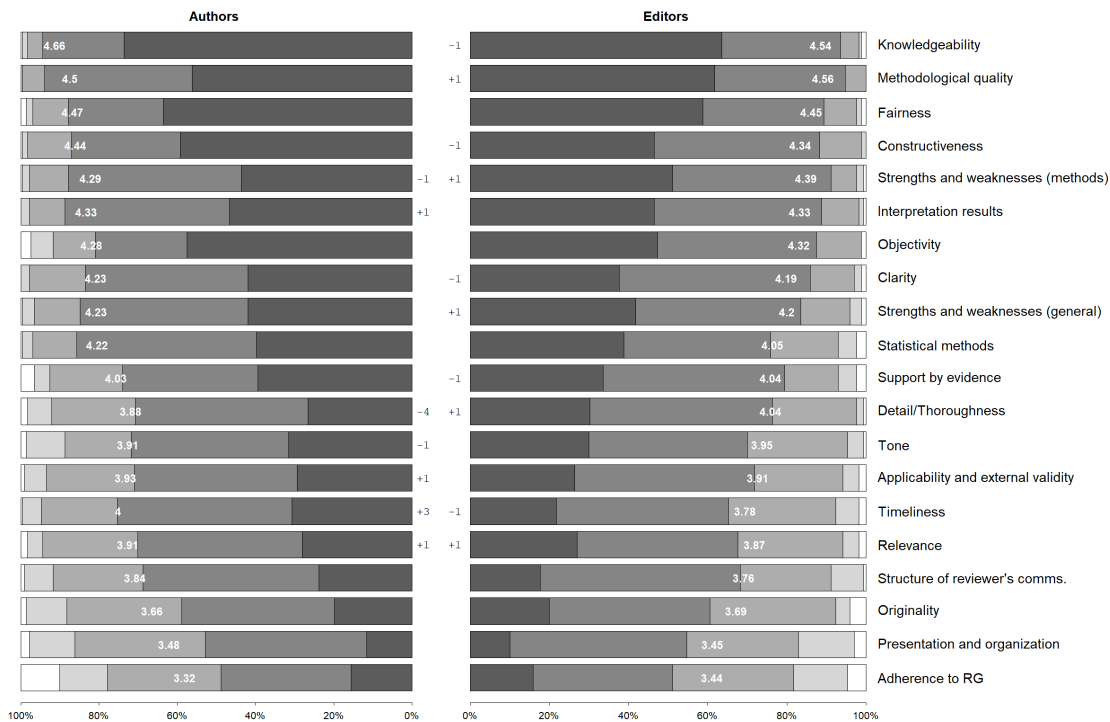


Fig. 5. The 20 items rated by editors and authors

The 20 items scored from 1 (in white) to 5 (in darker grey) are ordered according to the mean raking taking into account all survey participants. The mean score of each item according to either authors or editors' raking is reported within each bar. The numbers next to some authors or editors' bars (in the centre of the figure) represent the rank position of the corresponding item according to the separate authors' or editors' raking. The abbreviations 'comms.' and 'RG' in the figure were used for 'comments' and 'reporting guidelines', respectively.

A peer review report aims to help authors improve their submitted manuscripts and assist editors in taking editorial decisions. Due to this dual objective, we compared editors' and authors' mean scores in order to investigate whether any difference is found in their perceptions regarding the importance of the 20 items that assess peer review report quality. We found little discrepancy in the mean scores between biomedical editors and authors, with only two items indicating any difference: 1) *Timeliness* and 2) *Detail/Thoroughness*. The *Timeliness* of the peer review report was considered more important to authors than to editors (respectively, in the 12th and 16th rank positions). Meanwhile, editors rated the *Detail/Thoroughness* of the reviewer's comments higher than did authors (respectively, in the 11th and 16th rank positions).

Correlations among items

Overall, we found relatively weak positive correlations among items. The largest positive correlations were found between *Relevance* and *Originality*, and between *Fairness* and *Objectivity* ($r = 0.55$ and 0.43 , respectively).

Principal Component Analysis

We described the first three principal components because among all, these dimensions better explained the data variability, even though accounted for relatively small percentages of variances. The first principal component (PC1) accounted for 22.1% of data variability (Fig. 7). The next two dimensions (PC2 and PC3) accounted for 38.5% of the cumulative variability and contributed gradually, that is, they increased at only small increments (Fig. 8).

PC1 was positively correlated to all items (or variables), and it showed correlations higher than 0.4—which is the figure commonly used as a threshold reference for factor loadings—for 16 out of 20 items (Table 8). Particularly, PC1 defined a ‘size effect’ because it separated participants according to a magnitude: editors and authors who generally rated the items higher were situated on the right of the factorial plan, while those who rated them lower were on the left side. The second principal component (PC2) explained the 8.83% of the data variation. PC2 clearly differentiated the items into three main groups: items related to the *form*, items related to the *content* and those related to the *reviewer’s expertise and scientific rigour* of a peer review report³. Finally, the third principal component (PC3) captured the 7.59% of the data variability. PC3 was positively correlated with items related to the importance of the study such as

³Items related to the form are: Clarity, Constructiveness, Presentation and organization, Structure of reviewer’s comments, Tone, Timeliness,

Items related to the content are: Relevance, Originality, Interpretation results, Strengths and weaknesses (general), Strengths and weaknesses (methods), Applicability and external validity.

Items related to the reviewer’s expertise and scientific rigour of the PR report: Adherence to RG, Detail/Thoroughness, Fairness, Knowledgeability, Methodological quality, Objectivity. Statistical methods, Support by evidence.

Relevance and *Originality* and inversely with items about characteristics of the reviewer's comments such as the *Knowledgeability* and *Objectivity*.

These results illustrate that the data variance was not concentrated in a few components but distributed across all of them; hence, reducing the number of items is not recommended, since this would imply an important loss of data information.

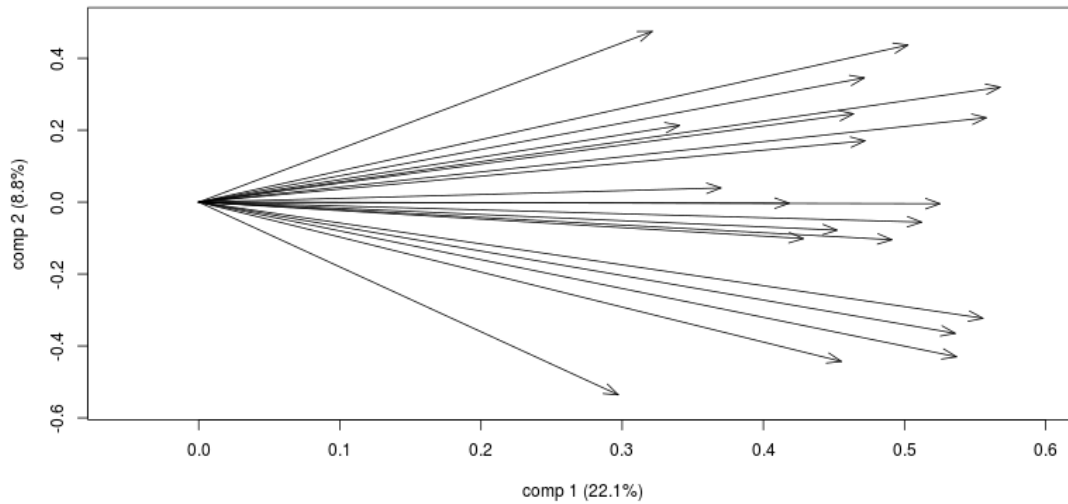


Fig. 6. PCA plot (PC1 vs PC2)

The interactive figure including the name of the variables is available at <https://www-eio.upc.edu/redir/ReportQuality>

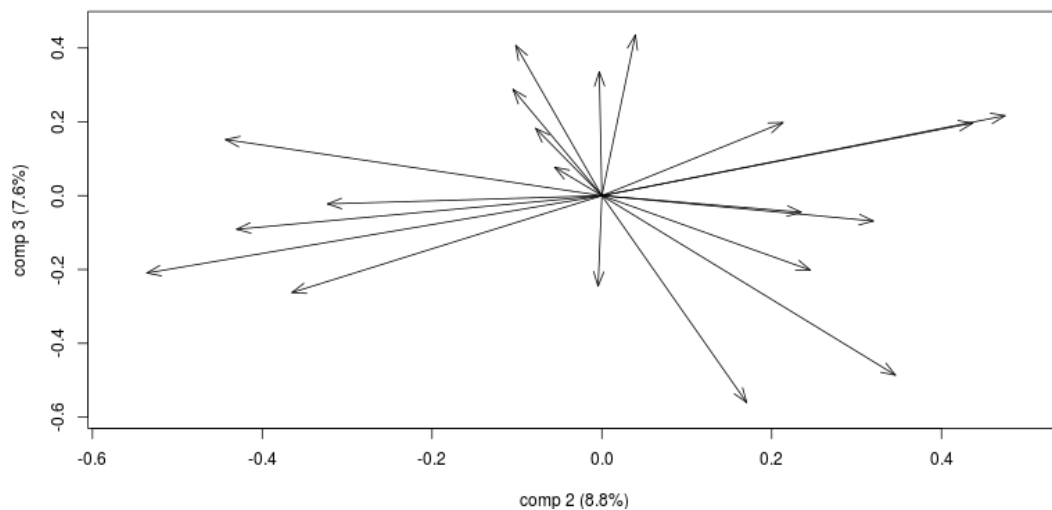


Fig. 7. PCA plot (PC2 vs PC3)

The interactive figure including the name of the variables is available at <https://www-eio.upc.edu/redir/ReportQuality>

Table 8. Items loadings

| Items | PC1 | PC2 | PC3 |
|-------------------------------------|-------|--------|--------|
| Strengths and weaknesses (general) | 0.568 | <±0.4 | <±0.4 |
| Applicability and external validity | 0.558 | <±0.4 | <±0.4 |
| Tone | 0.556 | <±0.4 | <±0.4 |
| Clarity | 0.537 | -0.43 | <±0.4 |
| Structure of reviewer's comments. | 0.536 | <±0.4 | <±0.4 |
| Presentation and organization | 0.525 | <±0.4 | <±0.4 |
| Adherence to reporting guidelines | 0.512 | <±0.4 | <±0.4 |
| Strengths and weaknesses (methods) | 0.502 | 0.437 | <±0.4 |
| Fairness | 0.491 | <±0.4 | <±0.4 |
| Relevance | 0.472 | <±0.4 | -0.56 |
| Originality | 0.472 | <±0.4 | -0.486 |
| Interpretation results | 0.464 | <±0.4 | <±0.4 |
| Constructiveness | 0.455 | -0.443 | <±0.4 |
| Detail/Thoroughness | 0.452 | <±0.4 | <±0.4 |
| Objectivity | 0.428 | <±0.4 | 0.407 |
| Support by evidence | 0.418 | <±0.4 | <±0.4 |
| Knowledgeability | <±0.4 | <±0.4 | 0.435 |
| Statistical methods | <±0.4 | <±0.4 | <±0.4 |
| Methodological quality | <±0.4 | 0.475 | <±0.4 |
| Timeliness | <±0.4 | -0.536 | <±0.4 |

Qualitative results

Comments on importance and/or wording of items

Out of 446 survey participants, 267 (59.9%) made at least one comment on the importance and/or wording of the items. Based on the initial coding of the comments, we were able to identify eight general themes that they addressed: Peer reviewer; Wording; Importance; Dependency; Responsibility; Item; Structure and content; and Improvement. Table 9 reports the eight themes together with their definition and the most frequent codes ($n > 5$), with example quotes (the complete codebook is available in the Zenodo repository).

Table 9. Survey participants' comments on the importance and/or wording of the 20 items to assess peer review report quality

| Themes | Definition | Codes | Examples |
|---------------|---|---|--|
| Dependencies | Theme including codes on how the importance of an item depends on different factors (e.g., type of study, paper quality, type of journal, etc.) | Dependency on the type of study (n=34) | <i>Depends on type of study. For systematic reviews of course fundamental. For other studies this will be more and more important for easier comparisons between studies and for quality improvement. It makes our work easier if the authors also compliance also improve</i> |
| | | Dependency on the paper quality (n=20) | <i>This depends on the quality of the manuscript. Sometimes the quality is so low that a reviewer can highlight one or two major methodological flaws which are sufficient to reject.</i> |
| | | Dependency on the type of journal (n=19) | <i>This depends on the journal's criteria</i> |
| | | Dependency on the author's claim and impact of the study (n=7) | <i>This depends on the claims made</i> |
| Importance | Theme including codes on the importance (or not) of an item. | Importance of the item (n=43) | <i>This is absolutely key to the interpretation of the study. Unfortunately most reviewers, in my field, do not fully understand current (and correct) methods.</i> |
| | | Importance of replication and conformation study (n=18) | <i>Not always important to be original study as some are trying to duplicate findings from previous studies.</i> |
| | | Importance of perceptions, opinions and experience (n=14) | <i>But some comments will inevitably be opinion, regarding emphasis, values, writing style</i> |
| | | Importance of a high quality review rather than on time review (n=13) | <i>Better to have a late high quality report than a moderate quality report on time.</i> |
| Improvements | Theme including codes on how an item is useful for both authors and editors in the peer review process. | Useful for authors and editors (n=21) | <i>It's important to make it easy for the editor and authors to understand the review, and for authors to respond.</i> |
| | | Improving the manuscript (n=9) | <i>Important when it will help improve the quality of the communication. Not necessary when it</i> |

| | | | |
|-----------------------|---|---|---|
| | | | <i>flows well.</i> |
| | | Avoiding exaggeration and misinterpretation (n=8) | <i>This is an area where the reviewer may have a valuable role in tempering an author's enthusiasm, hubris or bias.</i> |
| Item | Theme including codes on the characteristics of an item. | Related to other item (n=43) | <i>Yes, but it is confusing to separate this from the general strength and weaknesses. The question should be if the reviewer thinks that the message can (potentially) answer the research question.</i> |
| | | Subjective item (n=22) | <i>Too subjective! What is relevant to one person of field could be totally not-relevant to another</i> |
| | | Requirement (n=9) | <i>But it's an ethical requirement, and helps improve everyone's experience.</i> |
| Reviewer | Theme including codes on the expertise and characteristics of a peer reviewer. | Reviewer's expertise (n=148) | <i>Some reviewers know about methods and some about content. It would be ideal to always have both, but that is often not the case.</i> |
| | | Impossibility to be totally objective (n=35) | <i>100% objectivity doesn't exist</i> |
| | | Reviewer as an extra unpaid job (n=10) | <i>For the most part, reviews are done on a voluntary basis</i> |
| Responsibility | Theme including codes on the editor and/or author's responsibility to assess an item. | Editor's responsibility (n=48) | <i>In my experience this is usually picked up by the Editors and Associate Editors rather than the reviewers.</i> |
| | | Joint responsibility (n=24) | <i>I think this is the role of the editors as well as the reviewers.</i> |
| | | Author's responsibility (n=6) | <i>Authors should already be doing this</i> |
| Structure and content | Theme including codes on the structure and content of a peer review report. | Straight to the critical points (n=14) | <i>Sometimes a succinct review is still helpful, if it cuts straight to the critical points. For example, if it is clear that a manuscript has major flaws, then a review that points out those flaws clearly and dispassionately would be very helpful. It would not necessarily need to delve into the finer details.</i> |
| | | Unnecessary to provide evidence to each comment | <i>I don't think reviewers need to cite something</i> |

| | | | |
|---------|---|--------------------------------------|--|
| | | (n=10) | <i>for every point that they make.</i> |
| | | Declaration of COI (n=8) | <i>Peer reviewers should disclose COI.</i> |
| | | Standard structure of a review (n=7) | <i>I would suggest providing a template to reviewers.</i> |
| | | Not necessary for all reviews (n=6) | <i>Reviews come in all lengths and vary in detail. It is helpful to have some reviewers provide detailed information but not necessary that all do so.</i> |
| Wording | Theme including codes on how to improve the wording of an item. | Wording of the item (n=110) | <i>Rather than "The reviewer's comments are evidence-based" I would suggest that the category should be: "The reviewer distinguishes between comments that are supported by evidence (and provides suitable citations) and those based on opinion or experience"</i> |

New items

Participants suggested 13 items that were not included in the initial list of items. These items are listed in Table 10 (the complete codebook is available in the Zenodo repository).

Table 10. New items suggested by survey participants

| New items | Example |
|---|---|
| 1. Adherence to ethical guidelines | <i>“Comment on the study's adherence to ethical guidelines”</i> |
| 2. Author's contribution and acknowledgements | <i>“Clearly articulate the role of every team member, and their contribution to the study. For evidence syntheses, require librarian involvement and give them authorship, the same with statisticians. Everyone in the team, without whose knowledge the study would not be possible, sound, or complete, should be acknowledged.”</i> |
| 3. Data availability and software | <i>“Referees check the data availability and if new software actually works”</i> |
| 4. Disclosure of COI | <i>“Conflict of interests could be included”</i> |
| 5. Data sharing statements | <i>“Reviewers should ensure data sharing statements are included”</i> |
| 6. Study protocol | <i>“Whether a protocol was lodged in publication or on an independent site e.g., OSF and whether it matches the paper and if not, if reporting of deviations is transparent.”</i> |
| 7. Addressing study aims | <i>“I think the ‘does this study address its stated aims’ issue that I raised in my earlier responses is very important”</i> |
| 8. Study introduction | <i>“If the in introduction leads to the research question”</i> |
| 9. Study limitations | <i>“Whether limitations are acknowledged”</i> |
| 10. Study conclusion | <i>“And finally if the conclusion answers the research question.”</i> |
| 11. Theoretical framework | <i>“Logic of the theoretical framework”</i> |
| 12. Relevant literature | <i>“Reviewer rating of whether The authors discuss the most recent relevant research on the topic”</i> |
| 13. Reproducibility | <i>“Whether the study can be replicated on current methods”</i> |

Steering committee meeting

The steering committee met on the 19/07/2019 to discuss the selection of items to include in the final version of the tool. Their decisions were based on the participants' quantitative and qualitative answers. The flow of the items is summarized in Figure 9.

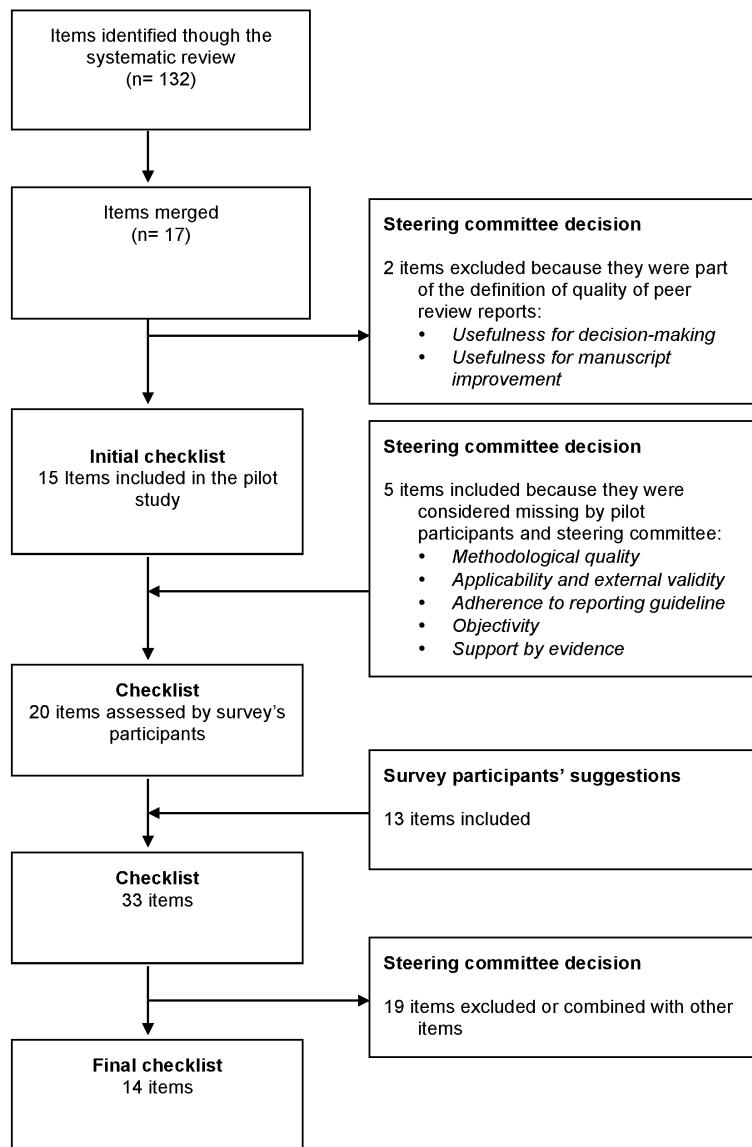


Fig. 8. Flowchart of items

The items *Relevance* and *Originality* were merged into a new item named *Contribution* (of the study). This decision was based on the high positive correlation found between the two items (0.55) and on the participants' opinions. Furthermore, participants suggested in their comments that the item *Relevance* was “*highly subjective*”, because “*each reviewer’s decision on relevance reflects what is relevant to them, which may not reflect relevance to the journal*”.

They also believed that the *Originality* of a study is not always an important aspect for comments in a peer review report, because some manuscripts “*are trying to duplicate findings from previous studies*”. They therefore suggested reformulating the two items by asking the reviewer what the study “*adds to our knowledge*”.

The steering committee decided to include the item *Interpretation of results* as a domain of the tool instead of a single item, changing the name into *Interpretation and discussion of the study results*. This decision resulted from the addition of two new items (*Conclusions* and *Limitations*), based on the suggestions of survey participants. The domain *Interpretation and discussion of the study results* now encompasses three items: 1) *Study conclusions*; 2) *Study limitations* and 3) *Applicability and generalizability*.

Overall, survey participants believed that the items *Strengths and weaknesses (general)* and *Strengths and weaknesses (methods)* were “*confusing to separate*”. Additionally, the steering committee agreed that *Strengths and weaknesses (methods)* and *Methodological quality* were also redundant; thus, it was ultimately decided to merge the three items into a new item named *Study methods*.

The items *Objectivity* and *Fairness* were merged because of both the moderate correlation between them (0.43) and the participants’ opinions. Participants suggested that the total objectivity of the reviewer’s comments is not possible because “*all decisions contain some personal biases and subjectivity*” and they also believed that the term fairness was “*very subjective*” and difficult to define. Additionally, the steering committee agreed to also combine these two items into *Supported by evidence*. The committee finally decided to merge all three items into *Objectivity*, and this was defined as “*comments provided in a peer review report should be as objective as possible and, if considered appropriate, include references to support the reviewer’s statements*”.

The steering committee agreed to merge *Structure of reviewer's comments* and *Clarity*, because participants considered both important for making the peer review report easy “*to read for both editors and authors*”. Moreover, participants suggested that the *Detail/Thoroughness* of a peer review report was mostly associated with the quality of a manuscript, because in certain occasions a study can be so poorly conducted that “*a reviewer can highlight one or two major methodological flaws*” without conducting a detailed review. They therefore believed that a detailed report is not “*always necessary*” and instead preferred a succinct report that “*cuts straight to the critical points*”. Taking into account the participants’ opinions, the steering committee finally decided to include a single item named *Clarity*, which is defined as “a peer review report should be clear, succinct and well organized in order to be understood correctly by editors and authors”.

The items *Tone* and *Constructiveness* were merged into *Constructiveness*, which is defined as “a peer review report should contain constructive and polite comments that allow the authors to improve the quality of their work”. This decision was based on the participants’ opinions that “*the comments should be polite and constructive*”.

The item *Adherence to reporting guideline* and the new item *Reproducibility* suggested by survey participants were merged into *Reporting* based on the steering committee decision. The item *Reporting* was defined as “the reviewer should comment if the reporting of the study is clear, complete and transparent enough for facilitating its reproducibility by verifying the adherence of the manuscript to the corresponding reporting guideline.”

The items *Timeliness* and *Knowledgeability* were not included in the final version of the tool. Survey participants suggested that *Timeliness* was not “*directly tied to review quality*” because “*some of the best reviews come in past the deadline*”. Furthermore, the steering committee agreed that the item *Knowledgeability* was generally difficult to assess, because it implied that anyone using the tool would have enough competence to evaluate the reviewer’s knowledge and

expertise. Five new items suggested by survey participants (*Data availability and software, Study protocol, Study conclusions, Study limitations and Relevant literature*) were finally included in the tool.

The ARCADIA tool

The ARCADIA (**A**ssessment of **R**eview reports with a **C**hecklist **A**vailable to **eD**itors and **A**uthors) tool was finally developed. The tool is a checklist that includes five domains and 14 items (Table 11). This tool aims to assess the quality of peer review reports in biomedical research, which was operationally defined 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*”. Since quality cannot be definite in absolute terms, this tool provides a proxy measure of the quality of peer review reports.

What is ARCADIA?

ARCADIA as a checklist provides a list of 14 items to “help editors make an informed and unbiased decision about the manuscripts’ outcome and, authors improve the quality of the submitted manuscript”. This checklist does not aim to assess the accuracy of the content of a peer review report (i.e., quality of the content), but it can be used to verify if a peer reviewer addresses the key elements of a manuscript in such a way that helps “editors make an informed and unbiased decision about the manuscripts’ outcome and, authors improve the quality of the submitted manuscript”.

It is hence the ARCADIA’s user which considers if a peer reviewer’s comment related to a specific item helps her/him to “make an informed and unbiased decision about the manuscripts’ outcome” or “improve the quality of the submitted manuscript” by ticking ‘Yes’ or ‘No’ in the checklist.

How to use ARCADIA?

Each item should be ticked as ‘Yes’ or ‘No’. However, 11 items also provide ‘Not applicable’ (NA) as a response option. An item could be checked ‘NA’ depending on whether the item is covered in the study, and/or the peer reviewer is qualified to comment on that specific item. For instance, theoretical essays do not have any methods and statistics, so the items corresponding to ‘study methods’ (ARCADIA item 2a) and ‘statistical methods’ (ARCADIA item 2b) should be ticked as ‘NA’. Brief explanations of the items included in the five domains are provided in Table 12.

Who can use ARCADIA?

ARCADIA can be used by editors to “make an informed and unbiased decision about the manuscripts' outcome” and, by authors, often acting as peer reviewers, to verify the quality of their own peer review report aimed to “improve the quality of the submitted manuscript”. However, a third party (e.g., researchers assessing an intervention aimed at improving the peer review process) can also use the tool taking into account both editor and author’s perspective. With the word ‘authors’ included in the name of the tool, we refer to all people conducting research and therefore publishing and/or reviewing biomedical articles.

ARCADIA is therefore not limited to authors and editors as potential users, but it can also be used by researchers interested in improving their peer review reports or in conducting research to evaluate interventions aimed at improving the peer review process.

Table 11. The ARCADIA tool

| In the peer review report, did the reviewer comment on... | | |
|---|--|--|
| Domain 1. Importance of the study | the contribution of the study to scientific knowledge? (item 1a) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | whether the relevant literature was accurately reviewed? (item 1b) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| Domain 2. Robustness of the study methods | the soundness of the study methods (e.g., study design, outcome measures, risk of bias)? (item 2a) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | the suitability of the statistical methods? (item 2b) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| Domain 3. Interpretation and discussion of the study results | whether the study conclusions answer the research question(s) and correctly summarize the study results? (item 3a) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | whether the study limitations are acknowledged? (item 3b) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | the applicability and generalizability (external validity) of the study results? (item 3c) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| Domain 4. Reporting and transparency of the manuscript | whether any major deviations from the study protocol are reported? (item 4a) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | whether the completeness of the reporting allows study reproducibility, by verifying the adherence of the manuscript to the corresponding reporting guideline? (item 4b) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | the presentation (e.g., quality of the written language, tables, figures, etc.) and organization of the manuscript? (item 4c) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | the availability of study data and materials? (item 4d) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| Were the peer reviewer's comments... | | |
| Domain 5. Characteristics of the peer reviewer's comments | clear? (item 5a) | <input type="checkbox"/> YES <input type="checkbox"/> NO |
| | constructive? (item 5b) | <input type="checkbox"/> YES <input type="checkbox"/> NO |
| | objective and, if opportune, supported by evidence? (item 5c) | <input type="checkbox"/> YES <input type="checkbox"/> NO |

NA=Not applicable

Table 12. ARCADIA items explanation

| | | |
|--|---|--|
| <p>Domain 1. Importance of the study</p> | <p>Item 1.a Contribution</p> | <p>A study can contribute to scientific knowledge in many ways: it can be a novel or confirmatory study with little or great impact on society and/or the research community. The contribution of a study is therefore not only associated to its novelty. Studies also need to be replicated in order to verify the validity of their results. The peer reviewer should discuss the importance of the study’s research question.</p> <p>Examples:</p> <ul style="list-style-type: none"> • This manuscript constitutes an excellent proof of concept that the repeat rate for chest radiography can be substantially improved by a behavioural skill model. • In both the Abstract and Introduction, the authors have failed to clearly articulate the underlying rationale for their study - or to highlight the added value of their study objectives and findings. |
| | <p>Item 1.b Relevant literature</p> | <p>The peer reviewer should check if the authors reviewed the relevant research related to the study’s topic in order to situate the study within the context of the existing literature.</p> <p>Examples:</p> <ul style="list-style-type: none"> • There are several studies that have been conducted in this area - and the authors need to clearly recognize the existing literature in this area and to illustrate the contribution of their work to this literature. • References are comprehensive. |
| <p>Domain 2. Robustness of the study</p> | <p>Item 2.a Study methods</p> | <p>The peer reviewer should evaluate the soundness of the study methods, such as the selection of the study design, assessment of the risk of bias, etc., to understand whether the methods were appropriate to the study’s aims, as well as if they were properly used and reported.</p> |

| | | |
|---|---------------------------------|--|
| methods | | <p>Examples:</p> <ul style="list-style-type: none"> • The program puts in ‘booklet and distribution to respondents’ as key factor to improve technical skills is not an appropriate approach. • Information about the pilot study is never discussed? How was this study administered? |
| | Item 2.b Statistical methods | <p>Data can be analysed in many ways, but the only appropriate statistical models are those that fit well with the study design and the characteristics of the variables. The peer reviewer with expertise in statistics should assess whether or not the study followed a suitable statistical procedure, as well as if they were correctly conducted and reported.</p> <p>Examples:</p> <ul style="list-style-type: none"> • Stats – you can’t do a ROC curve analysis for these results when the same results are actually used as the gold standard; this this doesn’t make statistical sense. • No description on how the missing data were handled |
| Domain 3. Interpretation and discussion of the study results | Item 3.a Study conclusions | <p>The reviewer should verify if the study conclusions answer the research question(s) and correctly summarize the study results.</p> <p>Examples:</p> <ul style="list-style-type: none"> • Some conclusions drawn were not supported by the analyses performed. • The discussion considers the findings of this research carefully and its potential implications. |
| | Item 3.b Study limitations | <p>The reviewer should check if the weaknesses of the study are correctly identified and discussed in order to interpret the validity of the research.</p> |

| | | |
|---|--|---|
| | | <p>Examples:</p> <ul style="list-style-type: none"> • Limitations of this study were not well addressed. In particular, it's been controversial to use tobit regression in the second stage analysis with DEA. • Study participants were recruited in community health centres. Could this be a source of selection bias? For example, are individuals with strong preferences for family care less likely to visit community health centres than individuals with preferences for nursing homes? Perhaps you could discuss this as part of the limitations of your study |
| | Item 3.c Applicability and generalizability | <p>The reviewer should comment on the applicability and generalizability of the study results. Applicability and generalizability are two underlying concepts of external validity [1]. The first concerns how “the results from a sample can be extended to the population from which the sample was drawn”, while the second how “the inferences drawn from study participants can be used in the care of patients drawn from any populations” (101).</p> <p>Examples:</p> <ul style="list-style-type: none"> • If possible, it might also be interesting to compare the descriptive statistics to the relevant population in Shanghai and discuss, whether your sample is somewhat representative of individuals in the relevant age group. • These findings are not only for China but also for many Asian countries, describing it in the discussion section will be more significant. |
| Domain 4. Reporting and transparency | Item 4.a Study protocol | Public access to study protocols is important to increase transparency and reduce waste of biomedical research. In the case of previous publication and/or inclusion as an additional file of a study protocol, the reviewer should verify that the major deviations from it are reported in the manuscript. |
| | Item 4.b | The reviewer should comment if the reporting of the study is clear, complete and transparent enough for facilitating |

| | | |
|---|---|---|
| of the manuscript | Reporting | <p>its reproducibility by verifying the adherence of the manuscript to the corresponding reporting guideline (if it has been developed). The Enhancing the Quality and Transparency of Health Research (EQUATOR) Network provides a toolkit to be used during the peer review process for selecting the appropriate reporting guideline (102).</p> <p>Examples:</p> <ul style="list-style-type: none"> • Important background information about the patients in the study is missing. What was the composition of patients in this study? How many patients were taking oestrogens? (These may bias all the cortisol results upwards, they should, perhaps, have been excluded.) How many patients were subsequently categorised on clinical grounds as adrenal insufficient and ACTH deficient? • In the method section, to avoid bias in interpreting the results is important to present the information on the sample size (formula and assumptions) or post hoc statistical power calculation. |
| | Item 4.c Presentation and organization | <p>The reviewer should discuss the quality of the written language used in the manuscript, shifts/reshuffling between sections, reduction of word count as well as of how the study results are presented (tables, figures, etc.).</p> <p>Examples:</p> <ul style="list-style-type: none"> • Table 1 could be re-arranged to avoid duplicating data and make it easier to read. • The standard of English is insufficient and needs a thorough revision and proofreading. |
| | Item 4.d Data availability | <p>When applicable, the reviewer should ensure that the data and materials (e.g., dataset, software codes), supported the results reported in the manuscript, are available.</p> |
| Domain 5. Characteristics of the | Item 5.a Clarity | <p>A peer review report should be clear (meaning that readers can easily understand its content), succinct and well organized (following the manuscript sections and, when it is necessary, providing line and page numbers) in order to be understood correctly by both editors and authors.</p> |

| | | |
|----------------------------|------------------------------|--|
| reviewer's comments | Item 5.b Constructiveness | A peer review report should contain constructive and polite comments that allow the authors to improve the quality of their work and editors to take a decision. |
| | Item 5.c Objectivity | <p>Comments provided in a peer review report should be as objective as possible and, if considered appropriate, include references to support the reviewer's statements.</p> <p>Examples:</p> <p><u>Objective PR report:</u></p> <ul style="list-style-type: none"> • Within the introduction many prognostic factors are mentioned but frailty is not - this is included within the discussion but I feel it is important to recognise it in the introduction as I believe there is a huge overlap between frailty and home help requirements. The clinical frailty score has been shown to be an independent predictor of inpatient mortality and length of stay. (See Wallis et al. Association of the clinical frailty scale with hospital outcome. QJIM. 2015;108(12):943-9 and Juma et al. Clinical Frailty Scale: A simple tool that predicts length of stay in an acute medicine unit. Journal of the American Geriatrics society. 2015;63:S121-5). • “First, one of the big issue of this study is that the objective is not well defined. Authors need to compare their approach with other approaches, if the objective is to prove that the four stage approach is superior to other approaches” <p><u>Non-objective PR report:</u></p> <ul style="list-style-type: none"> • Many modern studies of AI show that normal cosyntropin responses can be much lower than the cut-off used here. |

DISCUSSION

This study resulted in a tool, a checklist of items to assess the quality of peer review reports in biomedical research. To our best knowledge, it is the first such tool ever developed based on a review of the literature (54) and on empirical data from a large sample of both biomedical editors and authors. Further, it is the only tool that operationally defines the quality of peer review reports, as its definition was based on the perspectives of 446 authors and editors.

The present study also has some limitations. The response rate was very low, especially among authors (6%). This could be due to the mass mailings we used to contact participants. It has been shown that one of the factors that could contribute to a low response rate is the use of spamming filters, which block emails (103). Moreover, the survey questionnaire included some open-ended questions, which allowed participants to voluntarily express their opinions. However, we were not able to inquire further to clarify and verify some information provided by the study's participants. Therefore, the interpretation of some information could be affected by the perception of the three investigators who conducted the qualitative analysis. Additionally, since participants could comment voluntarily on the importance and wording of each item, the number of comments among items differed greatly. Finally, the majority of editors and authors who took part in the survey were from Europe and North America, which may limit the generalizability of the results. This result may be due to the recruitment strategy we used, especially to identify biomedical editors. Although we also utilized a snowballing strategy, we mainly contacted editors through European biomedical journals. Non-response and self-selection bias could have skewed the responses to be more positive toward our work.

CONCLUSIONS

ARCADIA is the first checklist that has been systematically developed to assess the quality of peer review reports. It is based on the perspectives of a large and heterogeneous sample of biomedical editors and authors.

CHAPTER 4. PSYCOMETRIC TESTING OF THE ARCADIA TOOL

This chapter will be submitted for publication to a peer-reviewed journal in the upcoming months.

Superchi C, Glonti K, Schroter S, Sánchez Espigares JA, Recchioni A, Hren D, Boutron I, Gonzalez José Antonio. Psychometric testing of ARCADIA, a tool for assessing peer review report quality in biomedical research. *Paper to be submitted for publication*

BACKGROUND

ARCADIA (Assessment of Review reports with a Checklist Available to eDItors and Authors) is the first tool that has been systematically developed to assess the quality of peer review reports in biomedical research (55). It is a checklist that includes five domains ('Importance of the study', 'Robustness of the study methods', 'Interpretation and discussion of the study results', 'Reporting and transparency of the manuscripts' and 'Characteristics of peer reviewers' comments') consisting of 14 items. This checklist is applicable for use with any study designs. The aim of the present study was to assess the psychometric properties of this newly developed tool.

METHODS

The study was approved by the Research Committee of the Governing Council of the Universitat Politècnica de Catalunya, Barcelona Tech, Spain (Reference: EC 01/2020, Date: 04/02/2020). A confidentiality agreement was signed with BMJ Publishing Group Ltd., before gaining access to the data.

Scoring of the ARCADIA tool

ARCADIA is a checklist and by definition, it does not include an overall score. However, for the purpose of the present study, a numerical summary score was needed to calculate measures of validity and reliability. Similarly to the total AMSTAR score (104), we included an *ARCADIA overall quality score* (from 0 to 100) which was calculated by dividing the total number of items ticked 'Yes' by the total number of items (n=14) excluding the items ticked as 'NA' and multiplying the final result by 100.

Assessment of the psychometric properties

We assessed the psychometric properties of ARCADIA in two phases, which were simultaneously conducted. The two phases mainly differed in the type of raters employed for assessing the peer review reports: in the first phase, raters were two researchers who are part of the research team and, consequently, more aware of the use of the tool (defined as ‘Fixed raters’); while in the second step, each peer review report was assessed by two raters which were randomly selected from a heterogeneous sample of biomedical editors and authors (defined as ‘Random raters’). We included both fixed and random raters to include a larger and diverse spectrum of assessors.

In phase I, we evaluated acceptability (i.e., distribution of the ARCADIA overall quality scores, ceiling/floor effect and endorsement frequencies), internal consistency, inter-rater and test-retest reliability. In phase II, we assessed practicability (i.e., time to complete the assessment using ARCADIA and feedback on the tool), inter-rater reliability and construct validity. As a gold standard does not exist for comparison, criterion validity was not assessed. Face and content validity were previously assessed in two ways: firstly, during the development of ARCADIA through the steering committee’s opinion; and secondly, responses to an open-ended question asking the participants “Are there any other items to assess the quality of peer review reports that you think should be included?” were examined to identify additional items not covered in the tool (55). Table 13 reports the definition of the psychometric tests used in the present study.

Table 13. Psychometric tests used in the present study

| Measure | Definition | Assessment | Criteria for acceptability |
|---|---|--|---|
| PHASE I | | | |
| Acceptability | Quality of data; assessed by skewness of score distribution, ceiling/floor effects and endorsement frequencies. | Distribution of scores (applied to overall scores); frequencies across response categories (applied to items) | Distribution should not be skewed; Even distribution of endorsement frequencies across response categories; Low ceiling/floor effects (percentage scoring lowest/highest score). |
| Reliability | | | |
| Internal consistency | The extent to which items in a tool measure the same concept; assessed through the degree of interrelatedness among the ARCADIA items. | Cronbach's α (applied to items) | The value of Cronbach's α ranges from 0 to 1, with higher values implying the items are consistently measuring the same dimension. A coefficient of ≥ 0.70 is considered acceptable (105). |
| | | Item-total correlations (applied to items) | An item should correlate with the total score above 0.30 (90). |
| Inter-rater reliability (assessment I) | The level of agreement between two or more independent raters of the same reports; assessed by administering the tool to two raters and examining the agreement between scores of the same report. | Cohen's kappa (applied to items) | The Cohen's kappa varies from ≤ 0 to 1. Values less than 0.40, between 0.40 and 0.75, and greater than 0.75 are respectively considered poor, fair and excellent reliability (106). |
| | | ICC coefficients (applied to overall scores) | The ICC (Intra-Class Correlation) ranges from 0 to 1. Values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.90 and greater than 0.90 are respectively considered poor, moderate, good and excellent reliability (107). |
| Test-retest reliability | The degree to which a tool reproduces stable scores over time by the same rater under the same conditions. A retest interval is usually of 2-14 days (90); assessed by administering the tool to two raters for a second time and examining the agreement between test and retest scores. | Cohen's kappa (applied to items) | See above |
| | | ICC coefficients (applied to overall scores) | See above |
| PHASE II | | | |
| Practicability | Ease of use of the tool; assessed by taking the time to complete the quality assessment and also feedback on the tool as indicator of survey participants' understanding. | Mean and SD of the time to complete the quality assessment General inductive approach for analysing comments from survey participants | No general criteria No general criteria |
| Reliability | | | |
| Inter-rater reliability (assessment II) | The level of agreement between two or more independent raters of the same report; assessed by administering the tool to two non-unique survey participants and examining the | Fleiss's kappa (applied to items) | The Fleiss's kappa range from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance and negative values indicate inverse |

| | | | |
|--------------------|--|---|---|
| | agreement between scores of the same report. | | agreement (108). |
| | | ICC coefficients (applied to overall scores) | See above |
| Validity | | | |
| Construct validity | The degree to which the scores of a tool are consistent with hypothesis based on the assumption that the tool validly measures the construct to be measured; assessed by examining the correlations between ARCADIA and external instruments' summary scores and results from factor analysis. | Pearson coefficients (External construct validity) Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) (Internal construct validity) | The Pearson correlation coefficient ranges from -1 to +1, with -1 indicating a total negative association, 0 no correlation and +1 a total positive association (109). Evidence from PCA and MCA that a single construct is measured (110) (97). Evidence from PCA supporting the structure of ARCADIA into five dimensions. |

Phase I. Acceptability, internal consistency, inter-rater and test-retest reliability

Sample of peer review reports

We used peer review reports associated with original biomedical research articles submitted to BMJ Open, a large general medical journal (111). We used only peer review reports made during the first round of peer review. Data was obtained from BMJ Open's manuscript tracking system, ScholarOne.

Data collection

SS randomly selected 60 manuscripts (30 accepted, 15 with a first decision of "major revisions" and consequently rejected and 15 with a first decision of "reject") that were submitted to BMJ Open between 1st of January to 31st March 2018, which had a final editorial decision by April 2019. A total number of 162 anonymized peer review reports associated with these 60 randomly selected manuscripts were used for assessing inter-rater reliability, test-retest reliability and internal consistency.

KG downloaded the 162 peer review reports associated with the 60 manuscripts from ScholarOne, anonymized the reports and gave each a unique identifier. To reduce researcher

bias, the main investigator (CS) was blinded to the editorial decision made for each manuscript for which the peer review reports were assessed. Two raters (CS, KG) independently assessed the quality of the peer review reports giving first an *overall score with a subjective scale* from 0 (extremely poor peer review report) to 100 (excellent peer review report) and then using the 14-item ARCADIA tool.

To assess test-retest reliability, a subsample of 30 randomly selected peer review reports was assessed for a second time by the same raters (CS, KG), after approximately 15 days from the first assessment. We used a simple randomization method stratified by the final decision on the manuscript (rejected and accepted) to ensure that 15 peer review reports were selected from each group.

Data analysis

Acceptability

We calculated the mean and median of the *ARCADIA overall quality score* distribution. Acceptability was assessed on the basis of skewness of the score distribution, endorsement frequencies and ceiling/floor effects.

Internal consistency

To assess the internal consistency, we calculated both 1) Cronbach's α and 2) item-total correlation.

1) The Cronbach's α expression is given by (112):

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2} \right)$$

where

k = Number of items

σ_i^2 = Variance of the i th item

σ_T^2 = Variance of the total score formed by summing all the items

2) The item-total correlation corresponds to the “*correlation of the individual item with the scale total omitting that items*” (90). The formula is given by:

$$r_{i(t-1)} = \frac{r_{it} \sigma_t - \sigma_i}{\sqrt{(\sigma_i^2 + \sigma_t^2 - 2\sigma_i \sigma_t r_{it})}}$$

where

$r_{i(t-1)}$ = Correlation of the item i with the total, removing the effect of item i

r_{it} = Correlation of the item i with the total score

σ_i = Standard deviation of item i

σ_t = Standard deviation of the total score

The calculations were performed using the R package Psych (113).

Inter-rater reliability (assessment I) and test-retest reliability

We calculated both 1) intra-class correlation (ICC) coefficients for continuous *ARCADIA overall quality scores* and 2) Cohen’s kappa coefficient for each categorical *ARCADIA* item.

- 1) We calculated the ICC coefficient using a linear mixed-effect model (LMM) with normal error distribution. ‘Rater’ was included as fixed effect predictor, while ‘Report’ as random effect factor. Additionally, the explanatory variable ‘Length of the peer review report’ was also added as fixed factor in the model. Effects were considered significant for p values < 0.05.

Mathematical representation of LMM

The general model, using a Laird and Ware formulation, was given by:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, i=1 \dots m,$$

$$\mathbf{b}_i \sim N(0, \Psi)$$

$$\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$$

where $\boldsymbol{\beta}$ is the p -dimensional vector of the fixed effect parameters, \mathbf{b}_i is the q -dimensional vector of random effects assumed to be normally distributed with mean 0 and a variance-covariance matrix Ψ ; $\boldsymbol{\varepsilon}_i$ is within subject error term and it conforms to a normal distribution with mean 0 and variance σ^2 . \mathbf{X}_i with dimensions $n \times p$ is the familiar design matrix of the general linear model; $\mathbf{X}_i\boldsymbol{\beta}$ is the fixed component of the model. \mathbf{Z}_i , with dimensions $n \times q$, is the design matrix for report i ; $\mathbf{Z}_i\mathbf{b}_i$ represents the random effects due to report i . The calculations were performed using the R package lme4 (114).

2) The Cohen's kappa is given by (115):

$$k = \frac{p_0 - p_e}{1 - p_e}$$

where

p_0 = Observed proportion of agreement

p_e = Proportion expected by chance

The calculations were performed using the R package Psych (113).

Phase II. Practicability, inter-rater reliability and construct validity

Survey

We conducted an online survey of editors and authors to test the tool and to evaluate its practicability, reliability and validity.

Survey questionnaire

The survey was designed by two researchers (CS, JAG) and was constructed as an interactive web application using the R package Shiny (116) (Appendix 9). To assess the response rate, JAG designed the system to be able to send the invitation emails to the participants, and tracked who opened the invitation emails, and who clicked to open the survey link. The system to track the invitation emails was not linked to the data entered by participants.

The survey questionnaire was structured into three main parts and included open and multiple-choice questions (Figure 10). Participants were first asked to assess the quality of a randomly selected peer review report giving an *overall score with a subjective scale* from 0 (extremely poor peer review report) to 100 (excellent peer review report) and then using the 14-item ARCADIA tool. They were also asked to report the time spent to complete the assessment using ARCADIA. To facilitate this task, a chronometer was provided to the participants before starting the assessment. Second, participants were asked to indicate if any ARCADIA item needed to be revised and also suggest how they would improve them. Finally, participants were asked some questions on their demographic characteristics (e.g., sex, age and education). The questionnaire was revised by all authors of the present study. Finally, it was piloted by 10 junior and senior researchers to ensure readability, comprehension, and acceptability.

PSYCHOMETRIC TESTING OF ARCADIA

Intro Instructions 1. Quality assessment 2. ARCADIA validation 3. Demographic questions

WELCOME

Thank you again for taking part in the first phase of our research. Based on your valuable response we were able to develop a new tool to measure the quality of peer review reports, the **ARCADIA** (Assessment of Review reports with a Checklist Available to eDitors and Authors). We would now like to invite you to take part in a subsequent survey which will enable us to evaluate its psychometric properties.

The survey will take approximately **20 minutes** to complete.

Your participation in this study is completely voluntary. If you decide to participate, all your answers will be de-identified and stored in a secured repository at Universitat Politècnica de Catalunya, Barcelona-Tech (Spain). The de-identified data from this study will be shared on Zenodo repository. If you opt out of sharing your data, you will still be able to participate in the study.

This survey has received ethics approval from the Research Ethics Committee of the Universitat Politècnica de Catalunya, Barcelona-Tech (Spain).

If you have any questions about this study or your rights as a participant, please email [Cecilia Superchi](mailto:Cecilia.Superchi@upc.edu).

Research team

This study is conducted by **Cecilia Superchi**, a PhD student at Universitat Politècnica de Catalunya, Barcelona-Tech and Université de Paris in collaboration with:

- Ketevan Glonti, PhD student at University of Split and Université de Paris
- Sara Schraier, Senior researcher at BMJ
- Alessandro Recchini, Senior editor at BMC Medicine
- Josep Anton Sánchez, Assistant prof. at Universitat Politècnica de Catalunya
- Danko Hren, Prof. at University of Split
- Isabelle Boutron, Prof. at Université de Paris
- José Antonio Gonzalez, Prof. at Universitat Politècnica de Catalunya

Related research:

1. Superchi C, Hren D, Blanco D, Riba R, Recchini A, Boutron I, González JA. The development of ARCADIA: a tool for assessing the quality of peer review reports in biomedical research. Submitted.
2. Superchi C, González JA, Solà I, Cobo E, Hren D, Boutron I. Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC Med Res Methodol*. 2019;19(43):1–14. DOI: <https://doi.org/10.1186/s12874-019-0555-x>

Do you agree to take part in the study?

Yes, I agree

No, I do not agree


Do you agree to share your de-identified data?

Yes, I agree

No, I do not agree

Start the survey

The MiRoR Project



This study is part of the **Methods in Research on Research (MiRoR)** project, a joint doctoral training programme in the field of clinical research funded by Marie Skłodowska-Curie Action.

Fig. 9. Survey welcome page

Participants

A total number of 151 biomedical editors and authors, identified from a previous study (55) and agreed to be contacted, were invited by email to participate in an online survey (Appendix 10).

Sample of peer review reports

We assigned a single peer review report to two survey participants entering the survey successively. A total number of 76 peer review reports were therefore assigned to 151 survey participants.

We used two different sources to identify the 76 peer review reports to assess in phase II to potentially include all study designs. A total of 31 reviews was selected from the sample of peer review reports (n=162) used in phase I. Since review length varied greatly across BMJ Open reports, we picked only peer review reports with more than 150 words (n=127). We obtained the number of words for each peer review report by means of the 'pdftotext' linux command in order to convert the pdf file to a single text file, and successively the 'wc' command to count the words. The peer reviewers corresponding to the selected peer review reports were contacted to get their approval to use their reports in the present study. From those peer review reports that received permission (n=37), we randomly selected 31 peer review reports. We used a simple randomization method stratified by the final decision on the manuscript to ensure that the peer review reports were balanced between rejected and accepted manuscripts.

The remaining 45 peer review reports were selected from a sample of 522 reviews related to a previously published RCT for evaluating the effect of training on the quality of peer review (51). These reports were associated with three RCTs and were independently assessed by two editors using the RQI (79). The RQI is a tool developed by Van Rooyen et al. in 1999 consisting of seven items, each scored on a 5-points Likert scale. The mean score of the seven items averaged over the two ratings was calculated. We used a simple randomization method stratified by low, medium and high values of *mean averaged total RQI score* to ensure that 15 peer review reports were selected from each group. Low, medium and high categories of *mean averaged total RQI score* variable were calculated by tertiles.

As shown in Table 14, two participants accessing the survey immediately one after the other were assigned the same report. We assumed that pairs of participants matched using this method were independent of each other. In addition, we organized the selected reports by randomly choosing: 1) a report from The BMJ with a high RQI score, 2) one from an accepted manuscript in BMJ Open, 3) one from The BMJ with medium RQI score, 4) one from a rejected manuscript in BMJ Open, 5) one from The BMJ with low RQI score; and so on.

Table 14. Method used to assign peer review reports to survey participants

| Report | Participant 1 | Participant 2 |
|---|----------------------|----------------------|
| R1 The BMJ with high RQI score | A | B |
| R2 BMJ Open (accepted manuscript) | C | D |
| R3 The BMJ with medium RQI score | E | F |
| R4 BMJ Open (rejected manuscript) | G | H |
| R5 The BMJ with low RQI score | I | L |
| R6 The BMJ with high RQI score | M | N |
| ... | ... | ... |

Data collection

The participants received a personalized email explaining the purpose of the study with a link to the survey. Two reminder emails were sent to non-respondents in order to incentivize their participation.

Data analysis

We described the demographic data in terms of frequencies and percentages.

Practicability

Time to complete the quality assessment

We calculated the mean and standard deviation for the time spent by survey participants in completing the assessment using ARCADIA.

Participant's feedback

We used a general inductive approach for analysing the survey participants' responses to the open-ended question regarding if any ARCADIA items needed to be revised and how they may be improved. In particular, we followed the five steps of inductive analysis proposed by David R. Thomas: 1) Preparation of raw data files; 2) Close reading of text; 3) Creation of codes; 4) Overlapping coding and uncoded text; 5) Continuing revision and refinement of themes system (99). In the third phase, an investigator (CS) created the initial codes from all responses. In order to ensure consistency and credibility, the initial codes were discussed with a second investigator (DH) and a codebook was developed. Two investigators (CS and DH) reviewed and refined the codebook. We used the software NVivo V.12 for data management and analysis (100).

Inter-rater reliability (assessment II)

We calculated both 1) intra-class correlation (ICC) coefficients for continuous *ARCADIA overall quality scores* and 2) Fleiss's kappa coefficient for each categorical ARCADIA item.

- 1) We calculated the ICC coefficient using a linear mixed-effect model (LMM) with normal error distribution. We used the same model previously described in Phase I. However, for this model, 'Report' and 'Rater' were both included as random effect factors. The within-rater variability was adjusted by including the interaction between 'Rater' and 'Length of the peer review report' in words. Effects were considered significant for p values < 0.05 .

- 2) We used Fleiss's kappa instead of Cohen's kappa because "the raters responsible for rating one subject are not assumed to be the same as those responsible for rating another" (106). The Fleiss's kappa is given by (16):

$$k = \frac{\bar{P}_0 - \bar{P}_e}{1 - \bar{P}_e}$$

where

$1 - \bar{P}_e$ = degree of agreement attainable over and above what would be predicted by chance

$\bar{P}_0 - \bar{P}_e$ = degree of agreement actually attained in excess of chance

The calculations were performed using the R package irr (117).

Internal construct validity

We conducted both principal component analysis (PCA) and multiple correspondence analysis (MCA) to investigate the internal construct validity of ARCADIA, meaning if the tool measures a single construct (i.e., the quality of peer review reports). In addition, PCA was also performed to confirm the structure of ARCADIA into five dimensions (i.e., the five *ARCADIA domains*).

PCA and MCA are multivariate statistical techniques designed for use with continuous and categorical variables, respectively. MCA is an extension of correspondence analysis (CA) when multiple categorical variables are considered.

Particularly, PCA is utilized to reduce the number of variables in a dataset to a smaller number of dimensions (97); while MCA aims to illustrate the most important relationships among variables' response categories (110).

These exploratory analyses help reveal simple underlying structures in complex datasets. Both techniques decompose the variance of the individual profiles around the average profile by

identifying mutually independent dimensions. In particular, the new dimensions are determined by choosing the directions that explain the most variation in the data. The first dimension accounts for the largest possible variance in the data, and each succeeding dimension accounts for decreasing amounts of the remaining. These mutually independent dimensions are commonly named *principal components* (PC) in PCA. We performed MCA and PCA using the R package FactoMineR (98).

The variables used in PCA were the five *ARCADIA domain scores* and the *overall score with a subjective scale* (for definitions see Table 14); while the three response categories - 'Yes', 'No' and 'NA' - for each ARCADIA items were used as variables in MCA.

External construct validity

We calculated Pearson correlations to examine the relationships between *ARCADIA overall quality score*, *mean averaged total RQI score* and *overall quality score with a subjective scale* (for definitions see Table 15).

Table 15. Definition of quality measures used in the present study

| Quality measures | Definition |
|---|---|
| ARCADIA overall quality score | $\left(\frac{\text{N items ticked as "Yes"}}{14 - \text{N items ticked as "NA"}} \right) * 100$ |
| ARCADIA domain score (for each domain) | $\left(\frac{\text{N items ticked as "Yes" w/i domain}}{\text{N items w/i domain} - \text{N items ticked as "NA" w/i domain}} \right) * 100$ |
| Mean averaged total RQI score | It corresponds to the mean score of the 7 items averaged over the two ratings. The RQI is a tool developed by Van Rooyen et al. in 1999 consisting of eight items, each scored on a 5-points Likert scale (79). Each of the first seven items reflects a different aspect of a review, while item 8 is a global question on the overall quality of the review. This tool is used at The BMJ journal for research studies. |
| Overall quality score with a subjective scale | It is a score from 0 (extremely poor peer review report) to 100 (excellent peer review report). A subjective scale represents the most common practice used among biomedical editors to assess the quality of peer review reports. |

RESULTS

Phase I. Acceptability, internal consistency, inter-rater and test-retest reliability

Assessment peer review reports

Between January and March 2020, two raters (CS and KG) independently assessed the sample of 162 peer review reports. Figure 11 shows the assessment of each ARCADIA item performed by the two raters.

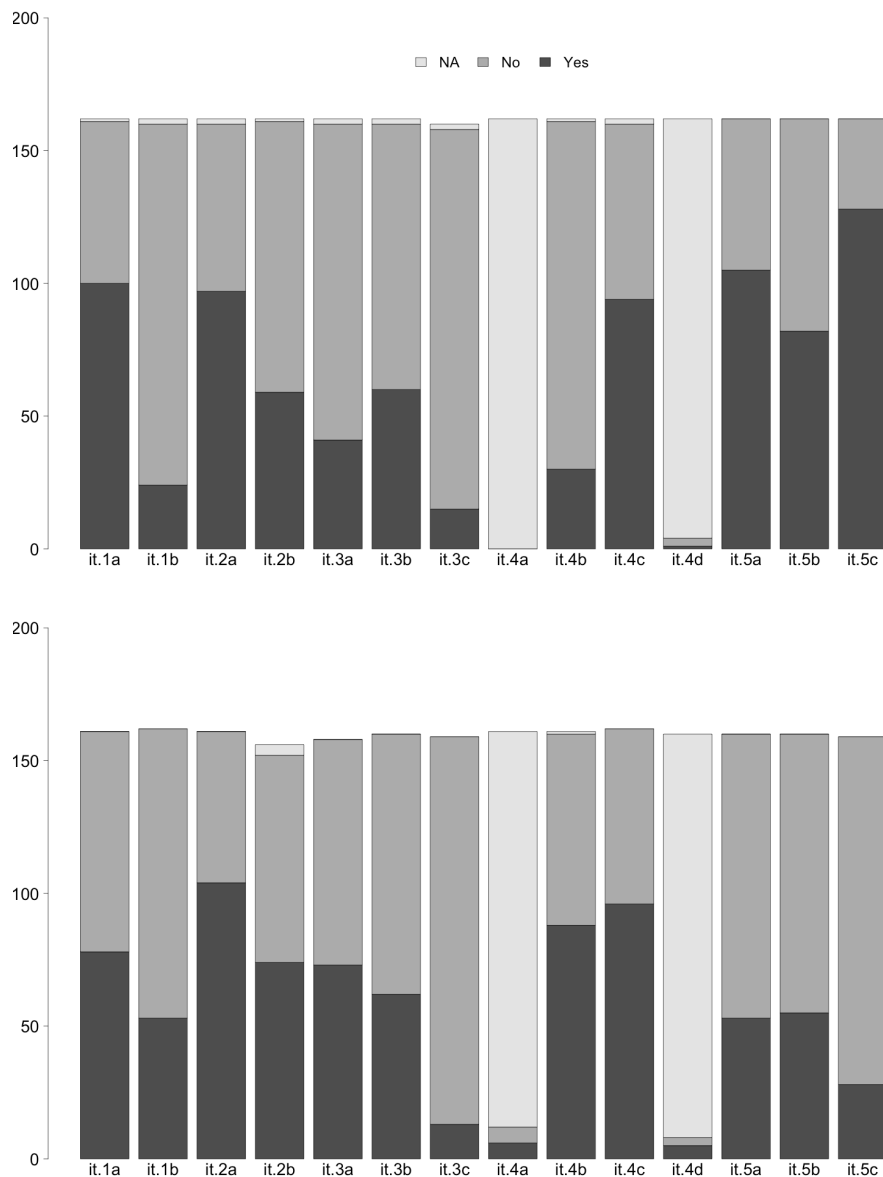


Fig. 10. Comparison of the assessment of each ARCADIA item between the two raters on the sample of 162 peer review reports

The y-axis indicates the number of peer review reports assessed by the two raters. The upper graph corresponds to the assessment carried out by the rater 1, while the lower one shows the assessment by rater 2.

The global mean (SD) for the *ARCADIA overall quality scores* was 42.3 (21.9), with values of 43.6 (22.3) and 40.9 (21.4) corresponding to each of the two raters; while the global mean (SD) for the *overall scores with a subjective scale* was 43.3 (27.9), with values of 60.2 (19.4) and 26.4 (24.8) for each assessor.

Acceptability

The *ARCADIA overall quality scores* followed a bell-shaped distribution (mean=42.3, median=41.6) and encompassed the full range of the measure from 0 to 100. Although the distribution was quite symmetric, we found a slight floor effect since a number of scores were distributed at the bottom of the scale (Figure 12).

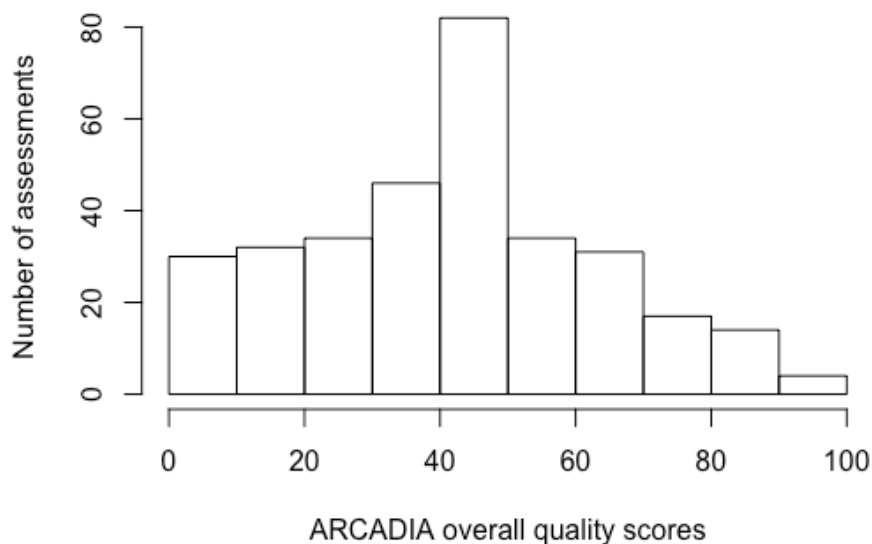


Fig. 11. Distribution of the ARCADIA overall quality scores

As expected, analysis of the endorsement frequencies showed that the items were mainly assessed as ‘Yes’ and ‘No’, except for the items 4.a Study protocol and 4.d Data availability, which were frequently assessed as ‘NA’ (Table 16).

Table 16. Endorsement frequencies for each ARCADIA item

| ARCADIA items | Endorsement frequencies | | | | | |
|---|-------------------------|----------------|----------------|----------------|----------------|----------------|
| | YES | | NO | | NA | |
| | Rater 1 | Rater 2 | Rater 1 | Rater 2 | Rater 1 | Rater 2 |
| Item 1.a Contribution | 100 (61.7%) | 78 (48.1%) | 61 (37.7%) | 83 (51.2%) | 1 (0.6%) | 0 (0.0%) |
| Item 1.b Relevant literature | 24 (14.8%) | 53 (32.7%) | 136 (84.0%) | 109 (67.3%) | 2 (1.2%) | 0 (0.0%) |
| Item 2.a Study methods | 97 (59.9%) | 104 (64.2%) | 63 (38.9%) | 57 (35.2%) | 2 (1.2%) | 0 (0.0%) |
| Item 2.b Statistical methods | 59 (36.4%) | 74 (45.7%) | 102 (63.0%) | 78 (48.1%) | 1 (0.6%) | 4 (2.5%) |
| Item 3.a Study conclusions | 41 (25.3%) | 73 (45.1%) | 119 (73.5%) | 85 (52.5%) | 2 (1.2%) | 0 (0.0%) |
| Item 3.b Study limitations | 60 (37.0%) | 62 (38.3%) | 100 (61.7%) | 98 (60.5%) | 2 (1.2%) | 0 (0.0%) |
| Item 3.c Applicability and generalizability | 15 (9.3%) | 13 (8.0%) | 143 (88.3%) | 146 (90.1%) | 2 (1.2%) | 0 (0.0%) |
| Item 4.a Study protocol | 0 (0.0%) | 6 (3.7%) | 0 (0.0%) | 6 (3.7%) | 162 (100%) | 149 (92.0%) |
| Item 4.b Reporting | 30 (18.5%) | 88 (54.3%) | 131 (80.9%) | 72 (44.4%) | 1 (0.6%) | 1 (0.6%) |
| Item 4.c Presentation and organization | 94 (58.0%) | 96 (59.3%) | 66 (40.7%) | 66 (40.7%) | 2 (1.2%) | 0 (0.0%) |
| Item 4.d Data availability | 1 (0.6%) | 5 (3.1%) | 3 (1.9%) | 3 (1.9%) | 158 (97.5%) | 152 (93.8%) |
| Item 5.a Clarity | 105 (64.8%) | 53 (32.7%) | 57 (35.2%) | 107 (66.0%) | 0 (0.0%) | 0 (0.0%) |
| Item 5.b Constructiveness | 82 (50.6%) | 55 (34.0%) | 80 (49.4%) | 105 (64.8%) | 0 (0.0%) | 0 (0.0%) |
| Item 5.c Objectivity | 128 (79.0%) | 28 (17.3%) | 34 (21.0%) | 131 (80.9%) | 0 (0.0%) | 0 (0.0%) |

Internal consistency

Cronbach's alpha coefficients did not exceed the criterion of 0.70. Particularly, the global Cronbach's alpha was 0.58, with values of 0.54 and 0.68 corresponding to each of the two raters. We found just minimal improvements on Cronbach's alpha by removing the items on study protocol (0.59), reporting (0.61) and presentation and organization (0.59) (Table 17). For nine out of 14 items, we found that the item-total correlations exceeded the criterion of 0.30 and were quite similar, indicating that those items contributed in an analogous way to the underlying construct of the tool (Table 17). Five items (2.b Statistical methods, 3.c Applicability and generalizability, 4.a Study protocol, 4.b Reporting, and 4.c Presentation and organization) showed item-total correlations below 0.30, meaning that they were less correlated to the *ARCADIA overall quality score*, compared to the other items.

Table 17. Internal consistency of ARCADIA tested on 162 peer review reports assessed by 2 raters

| ARCADIA items | Cronbach's alpha if an item is dropped | Item-total correlations |
|---|--|-------------------------|
| Item 1.a Contribution | 0.57 | 0.32 ² |
| Item 1.b Relevant literature | 0.56 | 0.31 ² |
| Item 2.a Study methods | 0.55 | 0.54 ² |
| Item 2.b Statistical methods | 0.60 | 0.15 |
| Item 3.a Study conclusions | 0.57 | 0.37 ² |
| Item 3.b Study limitations | 0.56 | 0.30 ² |
| Item 3.c Applicability and generalizability | 0.57 | 0.19 |
| Item 4.a Study protocol | 0.59 ¹ | 0.12 |
| Item 4.b Reporting | 0.61 ¹ | 0.18 |
| Item 4.c Presentation and organization | 0.59 ¹ | 0.00 |
| Item 4.d Data availability | 0.58 | 0.40 ² |
| Item 5.a Clarity | 0.50 | 0.59 ² |
| Item 5.b Constructiveness | 0.49 | 0.38 ² |
| Item 5.c Objectivity | 0.54 | 0.32 ² |

¹Higher Cronbach's alpha if the item is dropped (compared to the global Cronbach's alpha)

²Items with item-total correlation above 0.30

Inter-rater reliability

The inter-rater reliability of the *ARCADIA overall quality score* was moderate (ICC=0.57). By adjusting the LMM for the length of the peer review reports (i.e., the number of words used in a peer review report), we found that the intra-class correlation considerably decreased (ICC=0.34). No statistically significant difference was found between the two raters; while a statistically significant difference was found in length among peer review reports (Table 18). Cohen's Kappa (K) values varied between -0.02 (poor reliability) and 0.54 (fair reliability) across items. For five items K was around 0.5 (1.a Contribution, 2.a Study methods, 2.b Statistical methods, 3.b Study limitation and 4.c Presentation and organization), while for nine items it was below 0.4 (1.b Relevant literature, 3.a Study conclusions, 3.c Applicability and generalizability, 4.a Study protocol, 4.b Reporting, 4.d Data availability, 5.a Clarity, 5.b Constructiveness and 5.c Objectivity) (Table 19).

Table 18. Effect of ‘Rater’, ‘Report’ and ‘Words’ on the assessment of peer review report quality in Phase I

| | Inter-rater reliability (Unadjusted model) | | Inter-rater reliability (Adjusted model) | | Test-retest reliability (Unadjusted model) | |
|----------------------|---|---------------|---|---|---|----------------|
| | Estimate (SE) | 95%CI | Estimate (SE) | 95%CI | Estimate (SE) | 95%CI |
| Fixed effect | | | | | | |
| Intercept | 0.463 (0.028) | 0.407, 0.519 | 0.293 (0.031) | 0.233, 0.353 | 0.584 (0.053) | 0.479, 0.688 |
| Rater | -0.027 (0.016) | -0.058, 0.005 | -0.027 (0.016) | -0.058, 0.005 | -0.114 (0.025) | -0.165, -0.063 |
| Words | - | - | $4.2 \cdot 10^{-4}$ ($4.0 \cdot 10^{-5}$) | $3.5 \cdot 10^{-4}$, $4.9 \cdot 10^{-4}$ | - | - |
| Random effect | | | | | | |
| | Variance | ICC | Variance | ICC | Variance | ICC |
| Report | 0.0273 | 0.57 | 0.0107 | 0.34 | 0.0341 | 0.63 |
| Residual | 0.0207 | - | 0.0207 | - | 0.0197 | - |

Table 19. Inter-rater reliability tested on 162 peer review reports assessed by 2 raters

| ARCADIA items | Cohen’s Kappa (K) | Weighted Cohen’s kappa (Kw) |
|---|----------------------|--------------------------------|
| Item 1.a Contribution | 0.54 | 0.55 |
| Item 1.b Relevant literature | 0.38 | 0.40 |
| Item 2.a Study methods | 0.43 | 0.37 |
| Item 2.b Statistical methods | 0.49 | 0.35 |
| Item 3.a Study conclusions | 0.31 | 0.32 |
| Item 3.b Study limitations | 0.50 | 0.45 |
| Item 3.c Applicability and generalizability | 0.31 | 0.15 |
| Item 4.a Study protocol | 0.00 | 0.00 |
| Item 4.b Reporting | 0.06 | 0.06 |
| Item 4.c Presentation and organization | 0.48 | 0.47 |
| Item 4.d Data availability | -0.02 | -0.02 |
| Item 5.a Clarity | 0.03 | 0.04 |
| Item 5.b Constructiveness | 0.21 | 0.17 |
| Item 5.c Objectivity | 0.06 | 0.04 |

Test-retest reliability

The two raters tested the tool for a second time using a random sample of 30 peer review reports. The re-test interval mean was 15.5 days. Figure 13 shows the comparison of before-after *ARCADIA overall quality scores* between raters.

The test-retest reliability was moderate (ICC=0.63). Contrary to inter-rater reliability, we found a significant difference between raters (Table 18). Cohen’s Kappa (K) values varied between -

0.03 (poor reliability) and 1.00 (excellent reliability) across items. In particular, for five items K was around 0.55 (1.a Contribution, 2.b Statistical methods, 3.a. Study conclusions, 3.b Study limitation and 4.c Presentation and organization); while for eight items it was below 0.4 (1.b Relevant literature, 2.a Study methods, 3.c Applicability and generalizability, 4.a Study protocol, 4.b Reporting, 5.a Clarity, 5.b Constructiveness and 5.c Objectivity). Finally, for an item (4.d Data availability) K was equal to 1 (Table 20).

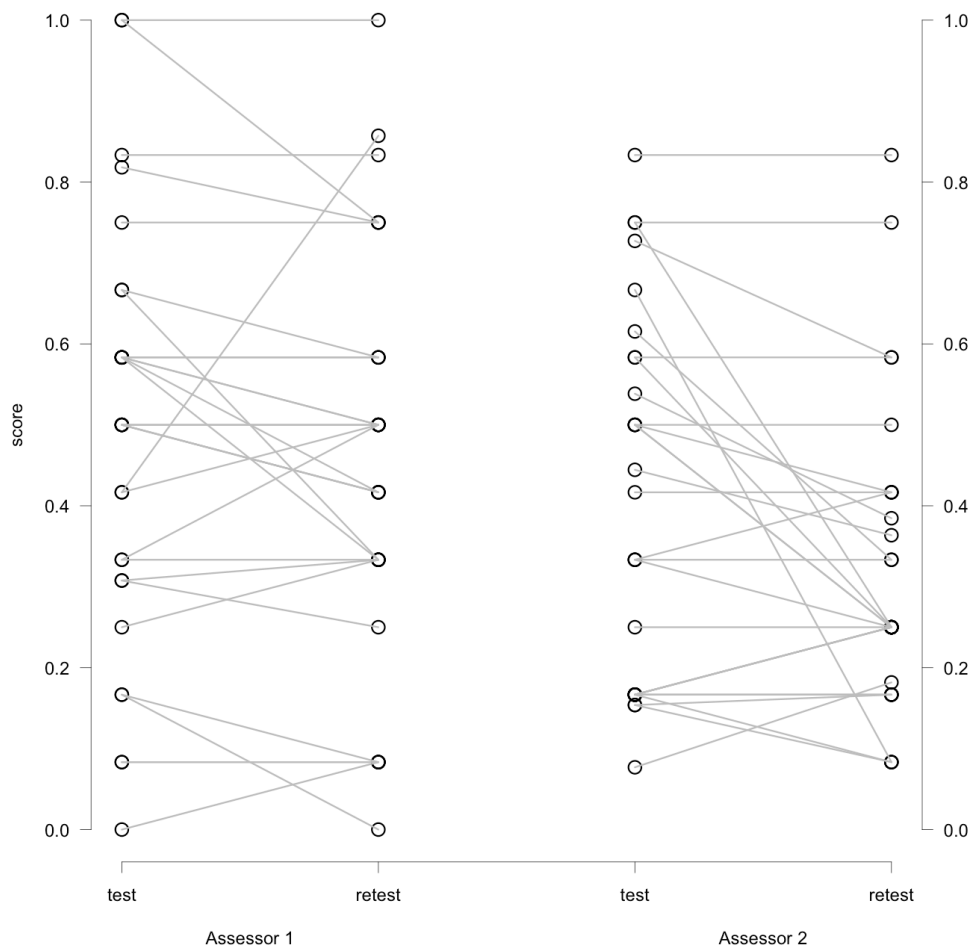


Fig. 12. Comparison of before-after ARCADIA overall scores between raters on a subsample of 30 peer review reports

In the figure, a dot corresponds to the *ARCADIA overall score* of a peer review report. Each line connects the before-after *ARCADIA overall scores* of the same peer review report. In some cases, the same *ARCADIA overall score* corresponds to various peer review reports and therefore multiple lines start from a single dot.

Table 20. Test-retest reliability tested on a subsample of 30 peer review reports assessed by 2 raters

| ARCADIA items | Cohen's Kappa (K) | Weighted Cohen's Kappa (Kw) |
|---|-------------------|-----------------------------|
| Item 1.a Contribution | 0.41 | 0.41 |
| Item 1.b Relevant literature | 0.08 | 0.17 |
| Item 2.a Study methods | 0.29 | 0.42 |
| Item 2.b Statistical methods | 0.68 | 0.64 |
| Item 3.a Study conclusions | 0.55 | 0.59 |
| Item 3.b Study limitations | 0.53 | 0.59 |
| Item 3.c Applicability and generalizability | -0.03 | 0.02 |
| Item 4.a Study protocol | 0.00 | 0.00 |
| Item 4.b Reporting | 0.14 | 0.14 |
| Item 4.c Presentation and organization | 0.62 | 0.62 |
| Item 4.d Data availability | 1.00 | 1.00 |
| Item 5.a Clarity | 0.08 | 0.08 |
| Item 5.b Constructiveness | 0.08 | 0.08 |
| Item 5.c Objectivity | 0.04 | 0.04 |

Phase II. Practicability, inter-rater reliability and construct validity

Survey participants

Between February 4, and March 9, 2020, 151 participants were contacted to participate in the study. Of the 57 (38%) participants who gave their consent, 48 (32%) completed the survey by assessing the peer review report using both the subjective scale and the ARCADIA tool. A total number of 35 (23%) participants responded to the eight questions related to the demographic characteristics.

Participants were mostly male (24/35, 68.9%) with a PhD degree (22/35, 63%) and their ages were equally distributed across ranges. Most of them were located in Europe (21/35, 60%) with more than 20 years of work experience as authors and/or editors in the biomedical field (15/35, 43%). The majority of them (28/35, 80%) used to review six or more manuscripts per year. Moreover, they mainly used to publish 10 or more than 10 (13/35, 37.1%) or two or less than two (11/35, 31.4%) articles per year. 65.7% of the participants had never received training on how to peer review a manuscript (Table 21).

Assessment of the peer review reports

The same randomly selected peer review report was assigned to two participants, which entered the survey successively. A total number of 48 survey participants assessed 30 peer review reports: 18 were assessed by two participants, while 12 by only one person because of the drop out of some participants during the study. Figure 14 shows the total number of ‘Yes’, ‘No’ and ‘NA’ responses for each item. As we found in Phase I, the items were mainly assessed as ‘Yes’ and ‘No’, except for the item 4.a Study protocol and item 4.d Data availability, which were frequently assessed as ‘NA’.

The global mean (SD) for the *ARCADIA overall quality scores* and *overall scores with a subjective scale* was 51.3 (22.1) and 65.1 (22.3), respectively.

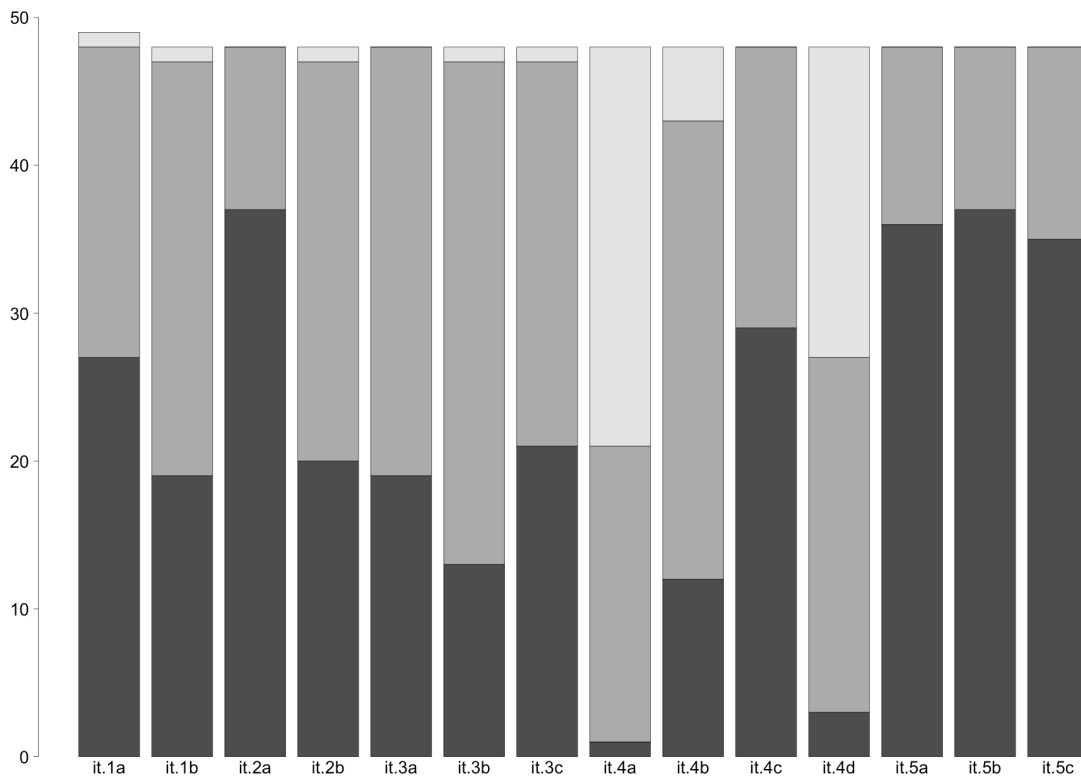


Fig. 13. Assessment of each ARCADIA item by survey participants

The y-axis indicates the number of survey participants.

Table 21. Participants' characteristics

| Characteristics | Total (N=48) |
|--|---------------------|
| Sex | N=35 |
| Female | 10 (28.6%) |
| Male | 24 (68.6%) |
| Other | 1 (2.8%) |
| Age | N=35 |
| <40 | 10 (28.6%) |
| 40-49 | 5 (14.2%) |
| 50-59 | 10 (28.6%) |
| >60 | 10 (28.6%) |
| Education | N=35 |
| Bachelor Degree | 0 (0.0%) |
| Master Degree | 1 (2.8%) |
| PhD | 22 (63%) |
| M.D. or equivalent | 10 (28.6%) |
| Prefer not to answer | 1 (2.8%) |
| Other | 1 (2.8%) |
| Location | N=35 |
| Europe | 21 (60.0%) |
| North America | 9 (25.7%) |
| South America | 3 (8.6%) |
| Africa | 0 (0.0%) |
| Asia | 0 (0.0%) |
| Australia | 2 (5.7%) |
| Number of years of experience | N=35 |
| <=5 years | 4 (11.4%) |
| 6-10 years | 6 (17.1%) |
| 11-15 years | 3 (8.6%) |
| 16-20 years | 7 (20.0%) |
| >20 years | 15 (42.9%) |
| Number of peer reviews per year | N=35 |
| <=2 | 4 (11.4%) |
| 3-5 | 3 (8.6%) |
| >=6 | 28 (80%) |
| Number of publications per year | N=35 |
| <=2 | 11 (31.4%) |
| 3-5 | 8 (22.9%) |
| 6-9 | 3 (8.6%) |
| >=10 | 13 (37.1%) |
| Training in peer review | N=35 |
| Yes | 12 (34.3%) |
| No | 23 (65.7%) |

We illustrated the five *ARCADIA* domains mean scores and overall score with a subjective scale (for definitions see Table 14) for each of the 30 reports to compare the quality assessments realized by the two raters (Figure 15). Overall, we found that the assessments were quite diverse: some peer review reports were assessed in a similar way (e.g., report ‘43B’ and report ‘24A’), while for others, the assessments carried out by the raters were very distinct (e.g., report ‘1068(low)’ and report ‘908(low)’).

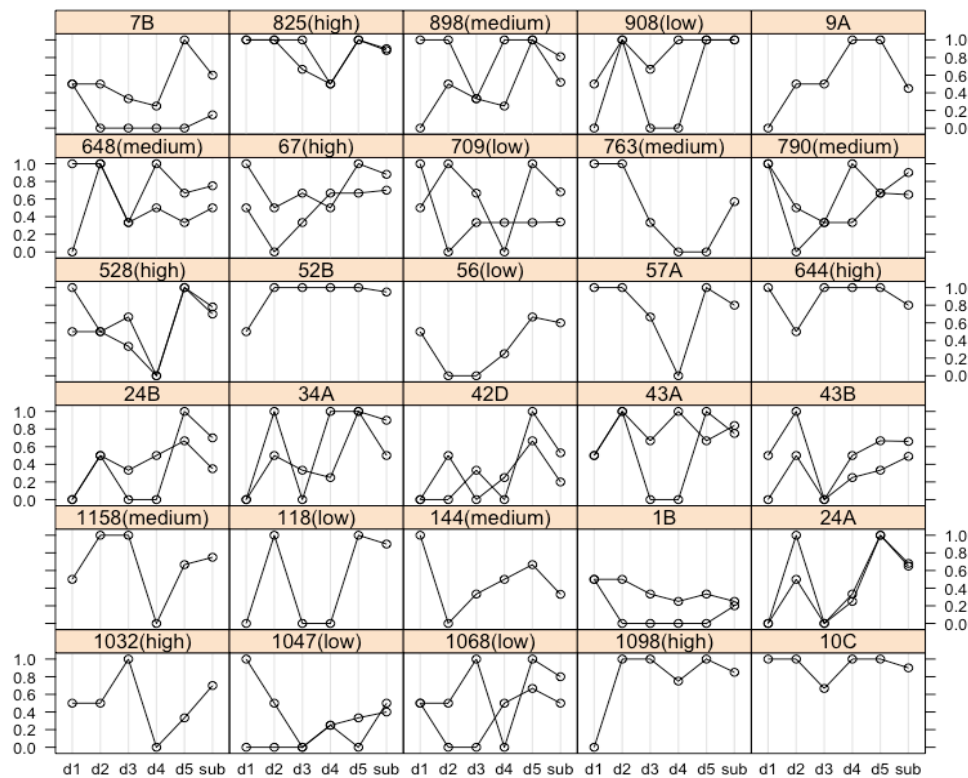


Fig. 14. Comparison of ARCADIA domain mean scores and overall score with a subjective scale between raters

d1= ARCADIA domain 1; d2= ARCADIA domain 2; d3= ARCADIA domain 3; d4= ARCADIA domain 4; d5=ARCADIA domain 5; sub=overall score with a subjective scale

Practicability

Time to complete the quality assessment

Out of 48, 47 (98%) participants reported the time spent to complete the assessment using ARCADIA. The average time was 5.44 minutes (SD=3.15) per report with values ranging from 2.00 (min) to 15.00 (max) minutes.

Participant's feedback

A total number of 32 participants (67%) made some comments on whether they believed any ARCADIA items needed revising and if so, how they may be improved. Based on the coding of the comments, we identified three main themes they addressed: 1) ARCADIA items, 2) ARCADIA tool and, 3) Missing items (Fig. 16). The code including more comments from the survey participants (8/32, 25%) was 'Response Type' (in bold in Fig.16). All 32 comments are reported in Appendix 11 and the entire codebook is found in Appendix 12.

ARCADIA items

Out of 32, 24 participants (75%) made some comments on how to improve the items encompassed in the ARCADIA tool. Particularly, they suggested improving the wording of ten items (1a.Contribution, 1b.Relevant literature, 2b.Statistical methods, 3b.Study limitations, 4a.Study protocol, 4b.Reporting, 4d.Data availability, 5a.Clarity, 5b.Constructiveness and 5c.Objectivity). In addition, they indicated three items as double-barrelled questions (3a. Study conclusions, 4b.Reporting and 4c.Presentation and organization) and, therefore they recommended splitting them into two separate questions.

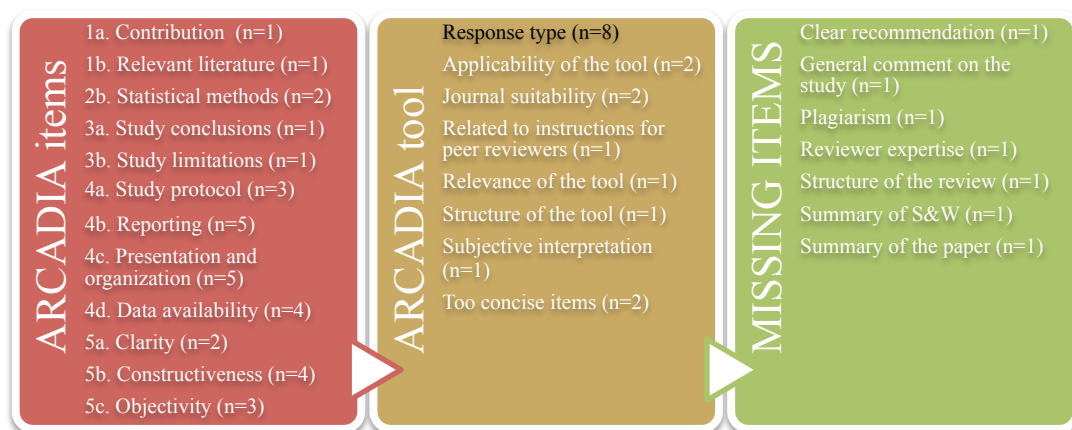


Fig. 15. Survey participants' feedback on ARCADIA

The number of comments included in each code is reported in brackets.

ARCADIA tool

Out of 32, 15 participants (47%) made some comments especially related to the tool. Particularly, the majority of participants (8/15, 53%) found the dichotomization of the response ('Yes' or 'No') not adequate for the tool since "*many items are PARTLY answered by the reviewer*". In addition, they suggested that "*a reviewer could mention an item but do a terrible job at reviewing/giving feedback on the item*".

Missing items

Some participants (5/32, 16%) suggested seven items, which should be included in the tool:

1. Clear recommendations for manuscript improvement;
2. General comment on the study;
3. Plagiarism;
4. Reviewer expertise;
5. Structure of the peer review report;
6. Summary of strengths and weaknesses of the paper;
7. General summary of the paper.

Inter-rater reliability (assessment II)

We found that the agreement between *ARCADIA overall quality scores* of the same rater using the same report was good (ICC=0.75); while the correlation of scores from any rater (regardless of which report they assessed) and also the correlation of scores for any report (regardless of the rater assessing the report) were moderate (ICC=0.59) and poor (0.15), respectively. By adjusting the LMM for the length of the peer review reports (i.e., the number of words used in a peer review report), the intra-class correlation slightly decreased. Finally, a statistically significant difference was found among *ARCADIA overall quality scores* (Table 22).

Table 22. Effect of ‘Rater’, ‘Report’ and ‘Words’ on the assessment of peer review report quality in Phase II

| | Inter-rater reliability Unadjusted model | | | Inter-rater reliability Adjusted model | | |
|----------------------|---|------------------|------------------------|---|--|------------------------|
| | Estimate (SE) | 95%CI | | Estimate (SE) | 95%CI | |
| Fixed effect | | | | | | |
| Intercept | 0.626 (0.036) | [0.552, 0.699] | | 0.540 (0.063) | [0.411, 0.669] | |
| ARCADIA_score | -0.130 (0.025) | [-0.181, -0.079] | | -0.129 (0.025) | [-0.180, -0.079] | |
| Words | - | - | | $1.4 \cdot 10^{-4}$ ($8.0 \cdot 10^{-5}$) | [- $4 \cdot 10^{-5}$, $3.1 \cdot 10^{-4}$] | |
| Random effect | | | | | | |
| | Variance | ICC | ICC^a | Variance | ICC | ICC^a |
| Rater | 0.0362 | 0.59 | 0.75 | 0.0357 | 0.61 | 0.74 |
| Report | 0.0094 | 0.15 | | 0.0074 | 0.13 | |
| Residual | 0.0154 | - | - | 0.0155 | - | - |

^a Global ICC

Fleiss’s Kappa values varied between -0.33 and 0.55 across items (Table 23). In particular, for nine items (1.a Contribution, 1.b Relevant literature, 2.b Statistical methods, 3.b Study limitation, 4.a Study protocol, 4.b Reporting, 5.a Clarity, 5.b Constructiveness and 5.c Objectivity) K values were positive; while for four items (2.a Study methods, 3.c Applicability and generalizability, 4.c Presentation and organization and 4.d Data availability) the inter-rater reliability was negative. Finally, only one item (3.a. Study conclusion) showed a value equal to 0, meaning that the agreement between raters was no better than what would be obtained by chance.

Table 23. Inter-rater reliability tested on 18 peer review reports assessed by 2 raters

| ARCADIA items | Fleiss's Kappa |
|---|-----------------------|
| Item 1.a Contribution | 0.39 |
| Item 1.b Relevant literature | 0.33 |
| Item 2.a Study methods | -0.33 |
| Item 2.b Statistical methods | 0.55 |
| Item 3.a Study conclusions | 0.00 |
| Item 3.b Study limitations | 0.20 |
| Item 3.c Applicability and generalizability | -0.17 |
| Item 4.a Study protocol | 0.26 |
| Item 4.b Reporting | 0.09 |
| Item 4.c Presentation and organization | -0.32 |
| Item 4.d Data availability | -0.09 |
| Item 5.a Clarity | 0.36 |
| Item 5.b Constructiveness | 0.26 |
| Item 5.c Objectivity | 0.08 |

Internal construct validity

Principal Component Analysis (PCA)

We described the first two principal components because among all, these dimensions better explained the data variability. Particularly, these components (PC1 and PC2) accounted for 63.3% of cumulative variability (Figure 17). PC1 was positively correlated to all *ARCADIA domain scores* and also to the *overall quality score with a subjective scale* (or variables), and it showed correlations higher than 0.4 —which is the figure commonly used as a threshold reference for factor loadings — for 5 out of 6 variables (Table 24). Particularly, PC1 distinguishes the peer review reports according to their quality: reports, which were generally rated higher, were situated on the right of the factorial plan, while those which were rated lower were on the left side (Figure 18). The second principal component (PC2) explained the 20.8% of the data variation. PC2 differentiated the peer review reports according to their journal's origin: reports, which were from The BMJ, were mostly situated on the upper part of the factorial plan, while those from BMJ Open were on the lower side. No difference was found between peer review reports associated with accepted or rejected manuscripts.

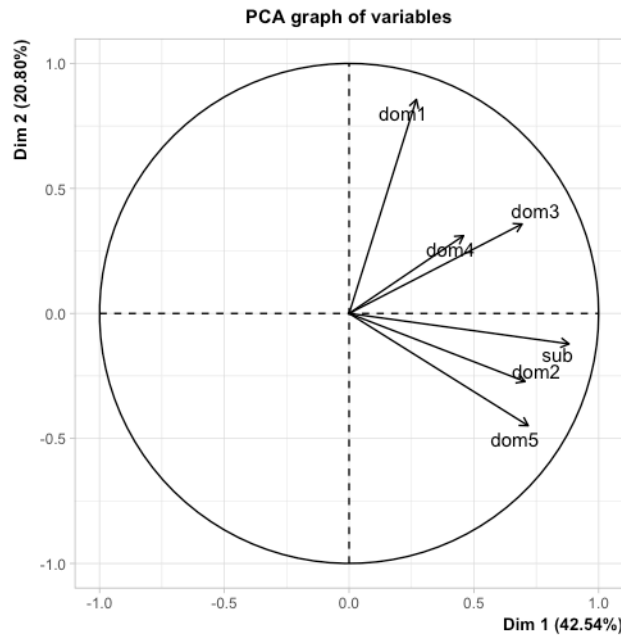


Fig. 16. Principal component analysis (PCA) of variables (i.e., ARCADIA domain scores and overall quality score with a subjective scale)

dom1= ARCADIA domain 1; dom2= ARCADIA domain 2; dom3= ARCADIA domain 3; dom4= ARCADIA domain 4; dom5=ARCADIA domain 5; sub=overall score with a subjective scale

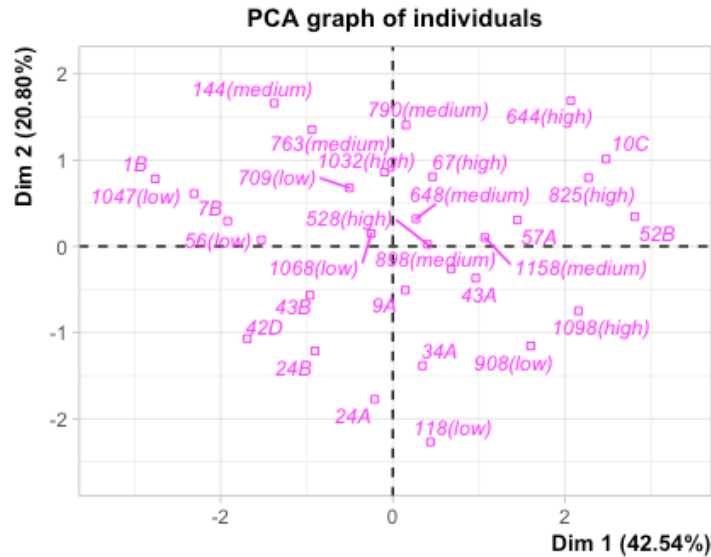


Fig. 17. Principal component analysis (PCA) of individuals (i.e., peer review reports)

In addition, we found that the five ARCADIA domains were not supported, since the variables, corresponding to each domain, were not clearly distinct in the factorial plan (Fig. 17). In particular, ARCADIA domain 3 (i.e., ‘Interpretation and discussion of the study results’) and

ARCADIA domain 4 (i.e., ‘Reporting and transparency of the manuscript’) were linked together as well as ARCADIA domain 2 (i.e., ‘Robustness of the study methods’) and ARCADIA domain 5 (i.e., ‘Characteristics of the peer reviewer’s comments’). Contrary to the other domains, ARCADIA domain 1 (i.e., ‘Importance of the study’) was clearly separated from the rest of the variables.

Table 24. Factor loadings

| Score | PC1 | PC2 |
|--------------|------------|------------|
| Dom1 | 0.270 | 0.856 |
| Dom2 | 0.705 | -0.272 |
| Dom3 | 0.693 | 0.358 |
| Dom4 | 0.458 | 0.311 |
| Dom5 | 0.718 | -0.449 |
| Sub | 0.881 | -0.122 |

dom1= ARCADIA domain 1; dom2= ARCADIA domain 2; dom3= ARCADIA domain 3; dom4= ARCADIA domain 4; dom5=ARCADIA domain 5; sub=overall score with a subjective scale

Multiple Correspondence Analysis (MCA)

The first two dimensions better explained the data variability, even though accounted for relatively small percentages of variances. These dimensions accounted for 27.55% of the cumulative variability. The other dimensions contributed gradually with small increments. Similarly to the results obtained with PCA, the dimension 1 distinguishes the ARCADIA items according to their quality: items which were rated as ‘Yes’ were situated on the right of the factorial plan, while those which were rated as ‘No’ were on the left side. The items rated as ‘NA’ were furthest away from the origin of the figure corresponding to the centroid of each variable (Figure 19).

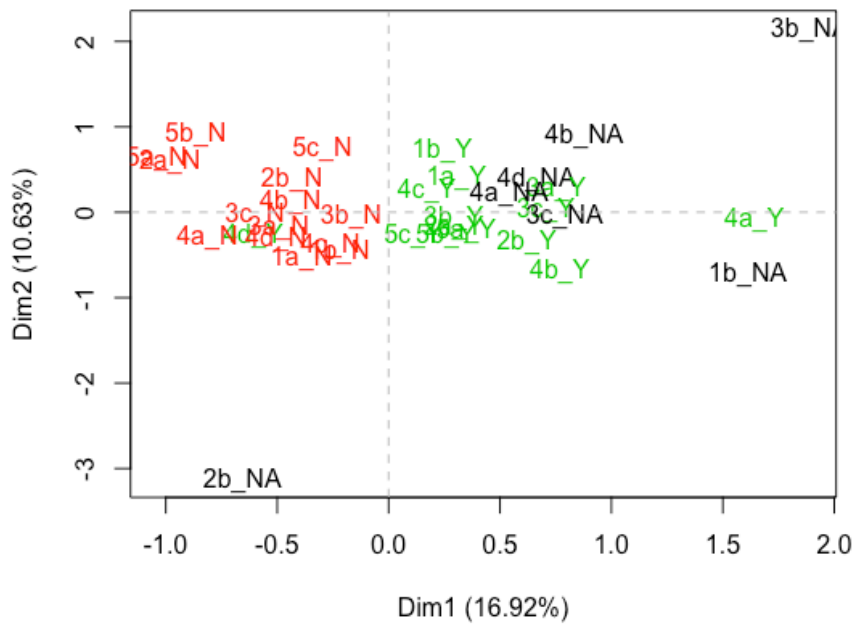


Fig. 18. Correspondence Analysis of variables (i.e. three response categories -‘Yes’, ‘No’ and ‘NA’- for each ARCADIA items)

Items assessed as ‘Yes’ are in green, items assessed as ‘No’ in red and items assessed as ‘NA’ in black.

External construct validity

Overall, we found strong positive correlations between the *ARCADIA overall quality score* and the *mean averaged total RQI score* ($r=0.57$) and *overall quality score with a subjective scale* ($r=0.72$).

DISCUSSION

ARCADIA was validated by a heterogeneous sample of both biomedical editors and authors using a sample of peer review reports from two different biomedical journals (i.e., The BMJ and BMJ Open). Field-testing demonstrated that the psychometric properties of ARCADIA are not entirely satisfactory.

The present validation study also has some limitations. We used only peer review reports made during the first round of peer review. Only 32% of individuals, who previously agreed to take in the validation of ARCADIA, actually participated in the study. Furthermore, we conducted the survey electronically and it may appear daunting and demanding to the participants and for this reason, some of the participants dropped out during the study. In addition, although the tool was used by a heterogeneous sample of biomedical editors and authors, the testing was confined to only two general biomedical journals.

CONCLUSIONS

Results from this study show that the psychometric properties of ARCADIA are not entirely satisfactory. These results should be used to inform a new version of the ARCADIA tool which it should be validated in a real-editorial setting. To ensure generalizability of validity and reliability, it is also necessary to apply ARCADIA to different types of journals, such as specialty journals, including peer review reports associated with manuscripts with diverse study designs.

CHAPTER 5. GENERAL DISCUSSION

This PhD thesis reports the development and validation of a new tool for assessing the quality of peer review reports applicable to all biomedical research.

As a starting point for the development of a new tool, we performed a comprehensive review to identify the tools used in biomedical research for assessing the quality of peer review reports (Chapter 2). We identified 24 tools from both the medical literature and an internet search: 23 scales and 1 checklist. One out of four tools consisted of a single item that simply asked the evaluator for a direct assessment of the peer review report's 'overall quality'. The remaining tools had between 4 to 26 items in which the overall quality was assessed as the sum of all items, their mean, or as a summary score.

Since the identified tools did not provide a definition of overall quality, these instruments consisted exclusively of a subjective quality assessment by the evaluators. Moreover, we found that only one study reported a rigorous development process of the tool, although it included a very limited number of people. This is of concern because it means that the identified tools were, in fact, not suitable to assess the quality of a peer review report, particularly because they lack a focused theoretical basis. We found 10 tools that were evaluated for validity and reliability; in particular, criterion validity was not assessed for any tool. Most of the scales with more than one item resulted in a summary score. These scales did not consider how items could be weighted differently.

Although commonly used, scales are controversial tools in assessing quality primarily because using a score "*in summarization weights*" would cause a biased estimation of the measured object (118). It is not clear how weights should be assigned to each item of the scale (58). Thus different weightings would produce different scales, which could provide varying quality assessments of an individual study (119).

In our methodological systematic review, we found only one checklist. However, it was neither rigorously developed nor validated and therefore we could not consider it adequate for assessing peer review quality. We believe that checklists may be a more appropriate means for assessing quality because they do not present an overall score, meaning they do not require a weight for the items.

It is necessary to clearly define what the tool measures. For example, the Risk of Bias (RoB) tool (120) has a clear aim (i.e., to assess trial conduct and not reporting), and it provides a detailed definition of each domain in the tool, including support for judgment. Furthermore, it was developed with transparent procedures, including wide consultation and review of the empirical evidence. Bias and uncertainty can arise when using tools that are not evidence-based, rigorously developed, validated and reliable; and this is particularly true for tools that are used for evaluating interventions aimed at improving the peer review process in RCTs, thus affecting how trial results are interpreted.

We found that most of the items included in the different tools did not cover the scientific aspects of a peer review report nor were constrained to biomedical research. Surprisingly, few tools included an item related to the methods used in the study, and only one inquired about the statistical methods.

In line with a previous study published in 1990 (69), we believe that the quality components found across all tools could be evaluated according to the perspective of either an editor or author, specifically by taking into account the different yet complementary uses of a peer review report (i.e., “to help editors make an informed and unbiased decision about the manuscripts’ outcome and, authors improve the quality of the submitted manuscript”).

As a second step to develop the tool, we investigated the perspectives of biomedical editors and authors towards the quality of peer review reports by conducting an international online survey

(Chapter 3). We also included patient editors as survey participants as their involvement in the peer review process can further ensure that research manuscripts are relevant and appropriate to end-users (121).

We recruited a large sample of biomedical editors and authors with varying experience and backgrounds. We found the percentage of female participants who took part in the survey to be low (129/399, 32.3%). This is in line with evidence showing that gender equity in academic medicine careers remains far behind (122). Moreover, we recruited corresponding authors (who are usually first or last authors) from the top 30-biomedical journals. Evidence also shows that women are underrepresented as first or last authors among biomedical journals with high Journal Impact Factors (123,124).

Based on a comprehensive systematic review and the empirical data of a large and heterogeneous sample of biomedical editors and authors, we developed the ARCADIA (Assessment of **R**eview reports with a **C**hecklist **A**vailable to **eD**itors and **A**uthors) tool. It is a checklist, simple to use and applicable to any biomedical field and it consists of five domains covering 14 items, each of which is phrased as a question.

‘Quality’ is a multidimensional and amorphous notion, which is difficult (and probably impossible, as demonstrated by Robert M. Pirsig’s story) to define. Robert M. Pirsig wrote in his book ‘Zen and the Art of Motorcycle Maintenance’: *“But even though Quality cannot be defined, you know what Quality is!”*

We believe that journal editors and authors could generally recognize a high-quality peer view report, but then they struggle to define its ‘quality’. As we found in our methodological systematic review, the concepts evaluated by the identified tools vary widely (54). Despite the difficulties to define this amorphous concept, our research contributes defining it for the first time taking into account the perspectives of 446 biomedical editors and authors.

We operationally defined peer review report quality assessed by ARCADIA 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”*.

Given our awareness that the quality as a whole cannot be objectively grasped because it would necessarily need to include specific expertise in each particular manuscript content area, we aimed to develop a proxy measure.

We regarded content areas covered by the peer review report as an aspect of quality that may be considered common across the spectrum of biomedical peer review reports. We are aware that this is admittedly a superficial level of defining quality, but we do believe that generally, peer review reports that cover more content areas can be considered of higher quality. That is not to say that there may not be opposite cases such as extensive review where the peer reviewer got something wrong or a review addressing a single but vital issue. There may well be such cases, but in large, we would still expect more complete reviews to reflect higher quality.

ARCADIA as a checklist, provides a list of 14 items, to “help editors make an informed and unbiased decision about the manuscripts’ outcome and, authors improve the quality of the submitted manuscript”. Therefore, the tool does not ask the user to rate the quality of a peer review report, but to assess if the reviewer appropriately comments on the key aspects of a manuscript.

This tool could be used by editors from any biomedical journal to evaluate the reviewer’s work. In addition, authors, commonly acting also as researchers and/or peer reviewers, could use it to improve their peer review reports (e.g, in training programme) and also as an outcome to evaluate interventions to improve the peer review process (e.g, in RCTs).

Findings from the field-testing of ARCADIA (Chapter 4) show that the internal consistency (global Cronbach alpha 0.58) was below the widely used criterion of 0.70. Moreover, the item-total correlations ranged from 0.00 to 0.59, indicating that the items contribute differently to the underlying construct of the tool. These results also show that equal weights cannot be applied to all items within the tool.

The inter-rater and test-retest reliability of the *ARCADIA overall quality score* was moderate with ICC values ranging from 0.57 to 0.75. In addition, we found that *ARCADIA overall quality scores* from any rater were mostly correlated regardless of which report they assessed (ICC=0.59), while the *ARCADIA overall quality scores* for any peer review report were quite uncorrelated regardless of who assessed it (ICC=0.15).

We found no improvements in ICC when adjusting the model with the word length of peer review reports (i.e., the number of words used in a peer review report). Particularly, in phase I, we found that the ICC decreased (ICC=0.34) compared to the ICC obtained using an unadjusted model (ICC=0.57). These results showed that the number of words in a report affected the assessment of its quality, meaning that the raters evaluated differently the quality of peer review reports characterized by diverse words length.

Moreover, we found that the difference in length among peer review reports used in phase I was statistically significant (95%CI 0.35 to 0.49) compared to those used in phase II (95%CI -0.04 to 0.31). This result may be due to the different sample of peer review reports used in the two phases. Particularly, in phase I, we included only peer review reports from BMJ Open, while in phase II, from both BMJ Open and The BMJ. Peer review reports from The BMJ were part of a trial on the effect of the training in quality of peer review published in 2004 (51). The peer reviewers of these reports were aware of being part of a trial and consequently, they may have overperformed by writing longer reviews.

In addition, in phase II, we selected only those BMJ Open reports with more than 150 words. This decision was made because the length was greatly different across BMJ Open reports. By selecting peer review reports with more than 150 words, all survey participants were given a similar workload when assessing the report.

We found that the values of Cohen's kappa varied greatly across items. The use of Cohen's kappa is a matter of debate. In particular, it has been shown that 'prevalence' is one of the two paradoxes, which jeopardised the use of Cohen's kappa as a measure to assess the reliability of a tool. 'Prevalence' refers to the relative numbers in 'Yes' and 'No' categories (90). It has been shown by Cicchetti and Feinstein that although the agreement of the two raters is high, kappa can result in a low value (125). This is due to the fact that the "*maximum value occurs when the prevalence is 50%, and decreases rapidly as the ratio deviates from a 50:50 split*" (90). The 'prevalence' paradox occurred with two ARCADIA items, i.e., 4.a Study protocol and 4.d Data availability. These two items were mainly assessed as 'NA' due to the fact that information about the availability of study protocols and/or data and materials was not accessible to the raters. Although the raters mostly agreed in assessing these items as 'NA', Cohen's kappa resulted in a low value. Moreover, we found that the Fleiss Kappa varied greatly across items with values ranging from -0.33 and 0.55. This result may be due to the small sample size of reports assessed in phase II (n=18).

The distribution of the *ARCADIA overall quality scores* showed evidence of a slight floor effect since most of the scores were situated on the bottom part of the scale. In addition, the majority of participants who gave feedback on the tool (8/15, 53%) found that the dichotomization of the responses was not adequate since "*many of these questions could have partial answers*". Similarly to the RQI (79), ARCADIA does not assess "*the degree to which the content of reviewer's comments is accurate*" or complete, but it presents a list of quality items that should be commented on in a peer review report to help editors make an informed and unbiased decision about the manuscript's outcome and, authors improve the quality of the submitted

manuscript. Therefore, the ARCADIA's user considers whether the peer reviewer's comment related to a specific item assists her/him by simply ticking 'Yes' or 'No' in the checklist.

A few participants suggested incorporating new items. We believe that the seven items suggested by five participants (5/48, 11%) are not key elements that should be commented on in a peer review report. In addition, some of them could be incorporated into already existing items by expanding their explanations.

The internal and external construct validity was satisfactory showing that the tool measured a single construct (i.e., the quality of peer review reports) and the *ARCADIA overall quality score* was positively correlated to other external instruments' summary scores (i.e., *mean averaged total RQI score* and *average quality score with a subjective scale*). However, findings from PCA also show that the five *ARCADIA domains* were not entirely supported.

Implications

This thesis makes a significant contribution to improving the efficiency and transparency of the peer review process by developing and evaluating a new tool for assessing peer review report quality in biomedical research. A more efficient and transparent peer review process is strongly needed in biomedical research as it will help prevent research misconduct and improve research integrity and study reproducibility.

Future research

The first version of ARCADIA presents unsatisfactory psychometric properties. Results from this PhD thesis should inform a new version of the instrument. Moreover, the new version should be developed online to be more accessible, user-friendly, and publicly available to all users.

As suggested by survey participants (24/32, 75%), the wording of some items, as well as their definitions reported in the 'ARCADIA items explanation' document, should be improved. Moreover, those items, identified as missing by the survey participants, should be incorporated into already existing items by expanding their explanations. Clearer instructions on how to use ARCADIA and on the meaning of its items may improve the acceptability and practicability of the tool.

Three items were identified as double-barrelled questions. A double-barrelled item is defined as “*one that asks two or more questions at the same time, each of which can be answered differently*” (88). We should therefore consider splitting these items into separate questions.

In addition, the majority of comments by the survey participants were about the dichotomization of the responses, which was considered not adequate for the tool. We should therefore consider investigating whether alternative response options are more appropriate for use of ARCADIA.

Finally, ARCADIA should be validated in a real-editorial setting using peer review reports associated with manuscripts with diverse study designs and from different types of biomedical journals.

General conclusions

The peer review process is the cornerstone of biomedical research. Low-quality biomedical research implies a tremendous waste of resources and explicitly affects patients' lives. This PhD project contributes to develop and validate ARCADIA, a new tool for assessing peer review report in biomedical research. We found that multiple tools have been used to assess peer review report quality, but none of them reported a definition of quality. In addition, the development and the validation process of those tools was unclear and the concepts evaluated by these tools vary widely. We developed ARCADIA, which constitutes the first tool that has been systematically developed to assess the quality of peer review reports and its development

is based on a large and diverse sample of biomedical editors and authors. In addition, it is the first tool, which provides an operational definition of the quality of peer review reports in biomedical research. The psychometric testing shows that improvements to ARCADIA should be pursued to create a new version that it can be recommended for routine use and in the study of peer review.

ARTICLES

Superchi C, González JA, Solà I, Cobo E, Hren D, Boutron I. Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC Med Res Methodol.* 2019;19(1):48. doi:10.1186/s12874-019-0688-x.

Superchi C, Hren D, Blanco D, Rius R, Recchioni A, Boutron I, González JA. Development of ARCADIA: a tool for assessing the quality of peer-review reports in biomedical research. *BMJ Open* 2020;0:e035604. doi:10.1136/bmjopen-2019-035604

OTHER PUBLICATIONS

Other papers published during the PhD programme:

Vo T-T, **Superchi C**, Boutron I, Vansteelandt S. The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. *Journal of Clinical Epidemiology*. 2019 Oct;117:78–88. doi: 10.1016/j.jclinepi.2019.10.001

Nguyen VT, **Superchi C**, Boutron I. 2-Year outcome from two parallel randomized controlled trials. Reporting considerations. *Osteoarthritis and Cartilage*. 2019 Mar;27(3):e3–4. doi: 10.1016/j.joca.2018.10.015

Kujinga P, Borgonjen-van den Berg KJ, **Superchi C**, ten Hove HJ, Onyango EO, Andang'o P, et al. Combining food-based dietary recommendations using Optifood with zinc fortified water potentially improves nutrient adequacy among 4- to 6-year-old children in Kisumu West district, Kenya. *Maternal & Child Nutrition*. 2017 Sep;14(2):e12515. doi: 10.1111/mcn.12515

Martínez García L, Pardo-Hernandez H, Niño de Guzman E, **Superchi C**, Ballesteros M, McFarlane E, et al. Development of a prioritisation tool for the updating of clinical guideline questions: the UpPriority Tool protocol. *BMJ Open*. 2017 Aug;7(8):e017226. doi: 10.1136/bmjopen-2017-017226

Martínez García L, Pardo-Hernandez H, **Superchi C**, Niño de Guzman E, Ballesteros M, Ibarгойen Roteta N, et al. Methodological systematic review identifies major limitations in prioritization processes for updating. *Journal of Clinical Epidemiology*. 2017 Jun;86:11–24. doi: 10.1016/j.jclinepi.2017.05.008

SCIENTIFIC PORTFOLIO

Conferences

Oral presentations

- 2nd PEERE International Conference on Peer Review, 29 September-1 October 2020, Valencia, Spain
- 16th Mediterranean Editors and Translators Meeting (METM), 26-28 September 2019, Split, Croatia
- 4th BIOSTATNET General Meeting, 25-26 January 2019, Santiago, Spain
- 3rd Conference for Young Researchers of the Spanish Society of Biostatistics, 18-19 January 2018, Bilbao, Spain

Posters

- MiRoR Conference on Meta-Research For Transforming Clinical Research, 25 November 2019, Paris, France
- 1st PEERE International Conference on Peer Review, 7-9 March 2018, Roma, Italy

Others

- PUBMET 5th Conference on Scholarly Publishing in the Context of Open Science, 20-21 September 2018, Zadar, Croatia
- 8th International Congress on Peer Review and Scientific Publication, 10-12 September 2017, Chicago, USA
- 38th Annual Conference of the International Society for Clinical Biostatistics, 9-13 July 2017, Vigo, Spain

Training Courses

- XII Summer School in Statistics and Operations Research, How to create an application with Shiny, 25-29 June 2018, Barcelona, Spain

- PEERE Training School on Peer Review, 15-17 May 2018, Split, Croatia
- XI Summer School in Statistics and Operations Research, Design aspects of individually randomised trials, cluster randomised trials and stepped wedge designs, 4-7 July 2017, Barcelona, Spain
- XI Summer School in Statistics and Operations Research, Data Science con R: Tidyverse, 3-7 July 2017, Barcelona, Spain

Peer reviewer

- Medicina Clínica
- BMJ Open

MiRoR Newsletter and journal clubs

- Interview to John Ioannidis (MiRoR newsletter [November 2017]) available at <https://www.youtube.com/watch?v=5wD4aSA2JLI>
- Letter to the editor published in the Osteoarthritis and Cartilage journal (MiRoR journal club [June 2018])
 - Nguyen VT, **Superchi C**, Boutron I. 2-Year outcome from two parallel randomized controlled trials. Reporting considerations. Osteoarthritis and Cartilage. 2019 Mar;27(3):e3–4
- Publication in the Journal of Clinical Epidemiology (MiRoR journal club [January 2019])
 - Vo T-T, **Superchi C**, Boutron I, Vansteelandt S. The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. Journal of Clinical Epidemiology. 2019 Oct;117:78–88.

Secondments

- Visiting Researcher at French Cochrane Centre, Paris, France, April-June 2019
- Visiting Researcher at University of Split, Split, Croatia, September-December 2018
- Visiting Researcher at Centre for Statistics in Medicine (CSM) and EQUATOR Network Centre, Oxford, United Kingdom, February-April 2018

REFERENCES

1. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biol.* 2015 Oct 2;13(10):e1002264.
2. Ioannidis JPA. Meta-research: Why research on research matters. *PLOS Biol.* 2018 Mar 13;16(3):e2005468.
3. Moher D, Altman DG. Four Proposals to Help Improve the Medical Research Literature. *PLOS Med.* 2015 Sep 22;12(9):e1001864.
4. Baker M, Dan P. Is there a reproducibility crisis? *Nature.* 2016 May 26;533:452–4.
5. Chalmers I, Glasziou P. Avoidable Waste in the Production and Reporting of Research Evidence. *The Lancet.* 2009;374:86-89.
6. Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ.* 2018 Nov 12;k4645.
7. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. *The Lancet.* 2014 Jan;383(9912):156–65.
8. Meerpohl JJ, Schell LK, Bassler D, Gallus S, Kleijnen J, Kulig M, et al. Evidence-informed recommendations to reduce dissemination bias in clinical research: conclusions from the OPEN (Overcome failure to Publish nEgative fiNDings) project based on an international consensus meeting. *BMJ Open.* 2015 May 5;5(5):e006666.
9. The MiRoR Project [Internet]. [cited 2019 Dec 3]. Available from: <http://mirror-ejd.eu/>
10. Oxford Dictionaries. Peer review definition [Internet]. Available from: https://en.oxforddictionaries.com/definition/peer_review
11. Smith R. Peer review: a flawed process at the heart of science and journals. *J R Soc Med.* 2006;99:178–82.
12. Jefferson T, Alderson P, Wager E, Davidoff F. Effects of Editorial Peer Review. *JAMA.* 2002;287(21):2784–6.
13. Publons. Global State of Peer Review [Internet]. 2018. Available from: <https://www.sciping.com/wp-content/uploads/2018/09/Publons-Global-State-Of-Peer-Review->

2018.pdf

14. Ross-Hellauer T. What is open peer review? A systematic review. *F1000*. 2017;(6).
15. Smith R. Opening up BMJ peer review. *BMJ*. 1999;318:4–5.
16. Das Sinha S, Sahni P, Nundy S. Does exchanging comments of Indian and non-Indian reviewers improve the quality of manuscript reviews? *Natl Med J India*. 1999;12:4.
17. van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of Blinding and Unmasking on the Quality of Peer Review. *JAMA*. 1998;280(3).
18. van Rooyen S, Godlee F, Evans S, Black N, Smith R. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ*. 1999 Jan 2;318(7175):23–7.
19. Vinther S, Nielsen OH, Rosenberg J, Keiding N, Schroeder TV. Same review quality in open versus blinded peer review in “Ugeskrift for Læger.”. *Dan Med J*. 2012; 59(8):A4479
20. Walsh E, Rooney M, Appleby L, Wilkinson G. Open peer review: A randomised controlled trial. *Br J Psychiatry*. 2000 Jan;176(1):47–51.
21. 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. *Nat Commun*. 2019 Dec;10(1):322.
22. Ross-Hellauer T, Deppe A, Schmidt B. Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. Wicherts JM, editor. *PLOS ONE*. 2017 Dec 13;12(12):e0189311.
23. Kronick DA. Peer review in 18th-century scientific journalism. *JAMA*. 1990;263(10):1321–2.
24. Moxham N, Fyfe A. The Royal Society And The Prehistory Of Peer Review, 1665–1965. *Hist J*. 2018 Dec;61(4):863–89.
25. Burnham JC. The evolution of editorial peer review. *JAMA*. 1990;263(10):1323–9.
26. Lock S. *A Difficult Balance: Editorial peer review in medicine*. Nuffield Prov Hosp Trust. 1985
27. Rennie D. Guarding the guardians: a conference on editorial peer review. *JAMA*.

1986;256(17):2391–2.

28. Rennie D. Make peer review scientific. *Nature*. 2016;535:31–3.

29. Ioannidis JPA, Berkwits M, Flanagin A, Godlee F, Bloom T. The Ninth International Congress on Peer Review and Scientific Publication: A Call for Research. *JAMA*. 2019 Nov 5;322(17):1658.

30. Moher, David. Custodians of High-Quality Science: Are Editors and Peer Reviewers Good Enough? [Internet]. Available from: <https://www.youtube.com/watch?v=RV2tknDtyDs>

31. World Medical Association. World Medical Association Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects. *JAMA*. 2013;310(20):2191-2194

32. Galipeau J, Barbour V, Baskin P, Bell-Syer S, Cobey K, Cumpston M, et al. A scoping review of competencies for scientific editors of biomedical journals. *BMC Med*. 2016 Dec;14(1):16.

33. Moher D, Galipeau J, Alam S, Barbour V, Bartolomeos K, Baskin P, et al. Core competencies for scientific editors of biomedical journals: consensus statement. *BMC Med*. 2017 Dec;15(1):167.

34. Bruce R, Chauvin A, Trinquart L, Ravaud P, Boutron I. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. *BMC Med*. 2016;14(1):85.

35. Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. Cochrane Methodology Review Group, editor. *Cochrane Database Syst Rev*. 2007;2:MR000016

36. Ghimire S, Kyung E, Kang W, Kim E. Assessment of adherence to the CONSORT statement for quality of reports on randomized controlled trial abstracts from four high-impact general medical journals. *Trials*. 2012 Dec;13(1):77.

37. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and Interpretation of Randomized Controlled Trials With Statistically Nonsignificant Results for Primary Outcomes. *JAMA*. 2010 May 26;303(20):2058.

38. Hopewell S, Collins GS, Boutron I, Yu L-M, Cook J, Shanyinde M, et al. Impact of

peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. *BMJ*. 2014 Jul 1;349:g4145–g4145.

39. Lazarus C, Haneef R, Ravaud P, Hopewell S, Altman DG, Boutron I. Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. *J Clin Epidemiol*. 2016 Sep;77:44–51.

40. Hopewell S, Witt CM, Linde K, Icke K, Adedire O, Kirtley S, et al. Influence of peer review on the reporting of primary outcome(s) and statistical analyses of randomised trials. *Trials*. 2018 Dec;19(1):30.

41. Chauvin A, Ravaud P, Baron G, Barnes C, Boutron I. The most important tasks for peer reviewers evaluating a randomized controlled trial are not congruent with the tasks most often requested by journal editors. *BMC Med*. 2015 Dec;13(1):158.

42. Glonti K, Cauchi D, Cobo E, Boutron I, Moher D, Hren D. A scoping review on the roles and tasks of peer reviewers in the manuscript review process in biomedical journals. *BMC Med*. 2019 Dec;17(1):118.

43. Glonti K, Boutron I, Moher D, Hren D. Journal editors' perspectives on the roles and tasks of peer reviewers in biomedical journals: a qualitative study. *BMJ Open*. 2019 Nov;9(11):e033421.

44. Baxt WG, Waeckerle JF, Berlin JA, Callahan ML. Who Reviews the Reviewers? Feasibility of Using a Fictitious Manuscript to Evaluate Peer Reviewer Performance. *Ann Emerg Med*. 1998;32(3):310–7.

45. Kravitz RL, Franks P, Feldman MD, Gerrity M, Byrne C, William M. Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care? 2010;5(4):2–6.

46. Yaffe MB; Re-reviewing Peer Review. 2009;2(85):1–3.

47. Stahel PF, Moore EE. Peer review for biomedical publications: we can improve the system. 2014;1–4.

48. Grimaldo F, Marušić A, Squazzoni F. Fragments of peer review: A quantitative analysis of the literature (1969-2015). Bornmann L, editor. *PLOS ONE*. 2018 Feb 21;13(2):e0193148.

49. Tennant JP, Ross-Hellauer T. The limitations to our understanding of peer review. *Res Integr Peer Rev.* 2020 Dec;5(1):6.
50. Houry D, Green S, Callaham M. Does mentoring new peer reviewers improve review quality? A randomized trial. *BMC Med Educ.* 2012 Dec;12(1):83.
51. Schroter S, Black N, Evans S, Carpenter J, Godlee F, Smith R. Effects of training on quality of peer review: randomised controlled trial. *BMJ.* 2004 Mar 20;328(7441):673.
52. Cobo E, Selva-O'Callaghan A, Ribera J-M, Cardellach F, Dominguez R, Vilardell M. Statistical Reviewers Improve Reporting in Biomedical Articles: A Randomized Trial. *PLoS ONE.* 2007 Mar 28;2(3):e332.
53. Heim A, Ravaud P, Baron G, Boutron I. Designs of trials assessing interventions to improve the peer review process: a vignette-based survey. *BMC Med.* 2018 Dec;16(1):191.
54. Superchi C, González JA, Solà I, Cobo E, Hren D, Boutron I. Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC Med Res Methodol.* 2019;19(1):48.
55. Superchi C, Hren D, Blanco D, Rius R, Recchioni A, Boutron I, et al. Development of ARCADIA: a tool for assessing the quality of peer-review reports in biomedical research. *BMJ Open.* 2020 Jun;10(6):e035604.
56. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* 2009;6(7):6.
57. NHS. PROSPERO International prospective register of systematic reviews [Internet]. Available from: <https://www.crd.york.ac.uk/prospero/>
58. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol.* 2007 Jun 1;36(3):666–76.
59. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; Available from: <https://www.R-project.org/>
60. Gentles SJ, Charles C, Nicholas DB, Ploeg J, McKibbin KA. Reviewing the research

methods literature: principles and strategies illustrated by a systematic overview of sampling in qualitative research. *Syst Rev*. 2016 Dec;5(1):172.

61. Glaser B, Strauss A. *The discovery of grounded theory*. Chicago: Aldine; 1967.
62. Landkroon AP, Euser AM, Veeken H, Hart W, Overbeke AJPM. Quality Assessment of Reviewers' Reports Using a Simple Instrument: *Obstet Gynecol*. 2006 Oct;108(4):979–85.
63. Rajesh A, Cloud G, Harisinghani MG. Improving the Quality of Manuscript Reviews: Impact of Introducing a Structured Electronic Template to Submit Reviews. *Am J Roentgenol*. 2013 Jan;200(1):20–3.
64. Friedman DP. Manuscript peer review at the AJR: facts, figures, and quality assessment. *Am J Roentgenol*. 1995 Apr;164(4):1007–9.
65. Black N. What Makes a Good Reviewer and a Good Review for a General Medical Journal? *JAMA*. 1998 Jul 15;280(3):231.
66. Henly SJ, Dougherty MC. Quality of manuscript reviews in nursing research. *Nurs Outlook*. 2009;57(1):18–26.
67. Callaham ML. Reliability of Editors' Subjective Quality Ratings of Peer Reviews of Manuscripts. *JAMA*. 1998 Jul 15;280(3):229.
68. Callaham ML, Knopp RK, John Gallagher E. Effect of written feedback by editors on quality of reviews: Two randomized trials. *JAMA*. 2002;287(21):2781–3.
69. McNutt RA. The Effects of Blinding on the Quality of Peer Review: A Randomized Trial. *JAMA*. 1990 Mar 9;263(10):1371.
70. Moore A, Jones R. Supporting and enhancing peer review in the *BJGP*. *Br J Gen Pract*. 2014 Jul;64(624):e459–61.
71. Thompson SR, Agel J, Losina E. The JBJS Peer-Review Scoring Scale: A valid, reliable instrument for measuring the quality of peer review reports: The JBJS Peer-Review Scoring Scale. *Learn Publ*. 2016 Jan;29(1):23–5.
72. Stossel T. Reviewer status and review quality. *N Engl J Med*. 1985;312(10):658–9.
73. Hettyey A, Griggio M, Mann M, Raveh S, Schaedelin FC, Thonhauser KE, et al. Peerage of Science: will it work? *Trends Ecol Evol*. 2012 Apr;27(4):189–90.

74. Publons. Publons for Editors : Overview [Internet]. Available from: https://static1.squarespace.com/static/576fcda2e4fcb5ab5152b4d8/t/58e21609d482e9ebf98163be/1491211787054/Publons_for_Editors_Overview.pdf
75. Shattell MM, Chinn P, Thomas SP, Cowling WR. Authors' and Editors' Perspectives on Peer Review Quality in Three Scholarly Nursing Journals. *J Nurs Scholarsh*. 2010 Mar;42(1):58–65.
76. Jawaid SA, Jawaid M, Jafary MH. Characteristics of reviewers and quality of reviews: A retrospective study of reviewers at Pakistan Journal of Medical Sciences. *Pak J Med Sci*. 2006;22(2):101–6.
77. Justice AC, Cho MK, Winker MA, Berlin JA, Rennie D, and the PEER Investigators. Does Masking Author Identity Improve Peer Review Quality?: A Randomized Controlled Trial. *JAMA*. 1998 Jul 15;280(3):240.
78. Henly SJ, Bennett JA, Dougherty MC. Scientific and statistical reviews of manuscripts submitted to *Nursing Research*: Comparison of completeness, quality, and usefulness. *Nurs Outlook*. 2010;58(4):188–99.
79. van Rooyen S, Black N, Godlee F. Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts. *J Clin Epidemiol*. 1999 Jul;52(7):625–9.
80. Evans AT, McNutt RA, Fletcher SW, Fletcher RH. The characteristics of peer reviewers who produce good-quality reviews. *J Gen Intern Med*. 1993 Aug;8(8):422–8.
81. Feurer ID, Becker GJ, Picus D, Ramirez E, Darcy MD, Hicks ME. Evaluating peer reviews: Pilot testing of a grading instrument. *JAMA*. 1994;272(2):98–100.
82. Henly SJ, Dougherty MC. Quality of manuscript reviews in nursing research. *Nurs Outlook*. 2009 Jan;57(1):18–26.
83. Kliever MA, Freed KS, DeLong DM, Pickhardt PJ, Provenzale JM. Reviewing the reviewers: Comparison of review quality and reviewer characteristics at the *American Journal of Roentgenology*. *Am J Roentgenol*. 2005;184(6):1731–5.
84. Berquist TH. Improving your reviewer score: It's not that difficult. *Am J Roentgenol*. 2017;209(4):711–2.

85. Callaham M, McCulloch C. Longitudinal Trends in the Performance of Scientific Peer Reviewers. *Ann Emerg Med*. 2011 Feb;57(2):141–8.
86. Yang Y. Effects of Training Reviewers on Quality of Peer Review: A Before-and-After Study (Abstract) [Internet]. 2009. Available from: https://peerreviewcongress.org/abstracts_2009.html
87. Prechelt L. Review quality collector [Internet]. Available from: <https://reviewqualitycollector.org/static/pdf/rqdef-example.pdf>
88. Callaham ML, Schriger DL. Effect of structured workshop training on subsequent performance of journal peer reviewers. *Ann Emerg Med*. 2002 Sep;40(3):323–8.
89. Ćurković M, Košec A. Bubble effect: including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility. *BMC Med Res Methodol*. 2018 Dec;18(1):130.
90. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. V edition. Oxford University Press; 2015.
91. SurveyMonkey Inc. SurveyMonkey. In San Mateo, California, USA; Available from: www.surveymonkey.com
92. Putton, M. Purposeful sampling. In: *Qualitative evaluation and research methods*. Sage; p. 169–86.
93. Sharp MK, Tokalić R, Gómez G, Wager E, Altman DG, Hren D. A cross-sectional bibliometric study showed suboptimal journal endorsement rates of STROBE and its extensions. *J Clin Epidemiol*. 2019 Mar;107:42–50.
94. EASE Blog. Biomedical editors survey on peer review [Internet]. 2018. Available from: <https://ese-bookshelf.blogspot.com/2018/11/biomedical-editors-survey-on-peer-review.html>
95. MiRoR. Biomedical editors survey on peer review [Internet]. 2018. Available from: <http://mirror-ejd.eu/2018/11/12/biomedical-editors-survey-on-peer-review/>
96. Fantini, Damiano. Search and Retrieve Scientific Publication Records from PubMed [Internet]. 2019. Available from: https://www.data-pulse.com/dev_site/easypubmed/
97. Jolliffe, I.T. *Principal Component Analysis*. 2nd ed. Springer; 2002.

98. Husson F, Josse J, Le S, Mazet J. FactoMineR: A Package for Multivariate Analysis. *J Stat Softw.* 2008;25(1):1–18.
99. Thomas, David R. A general inductive approach for qualitative data analysis. *AJE.* 2013;27.
100. QSR International. NVivo [Internet]. Available from: <https://www.qsrinternational.com/nvivo/home>
101. Murad MH, Katabi A, Benkhadra R, Montori VM. External validity, generalisability, applicability and directness: a brief primer. *BMJ Evid-Based Med.* 2018 Feb;23(1):17–9.
102. EQUATOR Network. Welcome to our toolkit for peer reviewing health research! [Internet]. Available from: <https://www.equator-network.org/toolkits/peer-reviewing-research/>
103. Fan W, Yan Z. Factors affecting response rates of the web survey: A systematic review. *Comput Hum Behav.* 2010 Mar;26(2):132–9.
104. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, et al. External Validation of a Measurement Tool to Assess Systematic Reviews (AMSTAR). Gagnier J, editor. *PLoS ONE.* 2007 Dec 26;2(12):e1350.
105. Nunnally J, Bernstein I. *Psychometric theory.* Third. New York: McGraw-Hill; 1994.
106. Fleiss J, Levin B, Paik M. *Statistical Methods for Rates and Proportions.* 2nd ed. New York: John Wiley; 1981.
107. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med.* 2016 Jun;15(2):155–63.
108. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76(5):378–82.
109. Mukaka MM. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. *Malawi Med J.* 2012;24(3):69–71.
110. Sourial N, Wolfson C, Zhu B, Quail J, Fletcher J, Karunanathan S, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *J Clin Epidemiol.* 2010 Jun;63(6):638–46.
111. *BMJ Open.* *BMJ Open* is a medical journal addressing research questions in clinical

- medicine, public health and epidemiology [Internet]. Available from: <https://bmjopen.bmj.com/>
112. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–333.
 113. Revelle W. *Procedures for Psychological, Psychometric, and Personality Research* [Internet]. 2018. Available from: <https://personality-project.org/r/psych>
 114. Doran H, Bates D, Bliese P, Dowling M. Estimating the Multilevel Rasch Model: With the lme4 Package. *J Stat Softw* [Internet]. 2007 [cited 2020 Jan 7];20(2). Available from: <http://www.jstatsoft.org/v20/i02/>
 115. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960 Apr;20(1):37–46.
 116. Chang W. Package “Shiny: Web Application Framework for R”. 2020. Available from: <https://shiny.rstudio.com/>
 117. Gamer M, Fellows I, Singh P. Package “irr” 2019. Available from: <https://cran.r-project.org/web/packages/irr/irr.pdf>
 118. Greenland S, O’Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*. 2001;2(4):463–71.
 119. Jüni P, Witschi A, Bloch R. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;282(11):1054–60.
 120. Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. *BMJ*. 2011 Oct 18;343(oct18 2):d5928–d5928.
 121. Schroter S, Price A, Flemyng E, Demaine A, Elliot J, Harmston RR, et al. Perspectives on involvement in the peer-review process: surveys of patient and public reviewers at two journals. *BMJ Open*. 2018 Sep;8(9):e023357.
 122. Bates C, Gordon L, Travis E, Chatterjee A, Chaudron L, Fivush B, et al. Striving for Gender Equity in Academic Medicine Careers: A Call to Action. *Acad Med*. 2016 Aug;91(8):1050–2.
 123. Filardo G, da Graca B, Sass DM, Pollock BD, Smith EB, Martinez MA-M. Trends and

comparison of female first authorship in high impact medical journals: observational study (1994-2014). *BMJ*. 2016 Mar 2;i847.

124. Gayet-Ageron A, Poncet A, Perneger T. Comparison of the contributions of female and male authors to medical research in 2000 and 2015: a cross-sectional study. *BMJ Open* 2019;9:e024436.

125. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol*. 1990 Jan;43(6):551–8.

APPENDIX

Appendix 1. Search strategies

Pubmed (n=2474) (last search: October 27, 2017)

("Peer Review, Research/methods"[Majr] OR "Peer Review, Research/standards"[Majr] OR "Peer Review /methods"[Majr] OR "Peer Review /standards"[Majr] OR peer review*[tiab]) AND (manuscript*[tiab] OR article*[tiab] OR publication*[tiab] OR report [tiab] OR reports [tiab] OR submission*[tiab] OR review*[tiab]) AND ("Quality Control"[Mesh] OR quality [ti] OR validity [ti] OR measure*[ti] OR instrument*[ti] OR scale*[ti] OR score*[ti] OR assessment*[ti] OR apprais*[ti] OR tool [ti])

Cochrane Library (n=340) (last search: October 27, 2017)

ID Search

#1 MeSH descriptor: [Peer Review, Research] explode all trees and with qualifier(s): [Methods - MT]

#2 MeSH descriptor: [Peer Review, Research] explode all trees and with qualifier(s): [Standards - ST]

#3 MeSH descriptor: [Peer Review] explode all trees and with qualifier(s): [Methods - MT]

#4 MeSH descriptor: [Peer Review] explode all trees and with qualifier(s): [Standards - ST]

#5 "peer review*":ti,ab,kw (Word variations have been searched)

#6 #1 or #2 or #3 or #4 or #5

#7 "manuscript*":ti,ab,kw (Word variations have been searched)

#8 "article*":ti,ab,kw (Word variations have been searched)

#9 "publication*":ti,ab,kw (Word variations have been searched)

#10 "report":ti,ab,kw (Word variations have been searched)

#11 reports:ti,ab,kw (Word variations have been searched)

#12 "submission*":ti,ab,kw (Word variations have been searched)

#13 "review*":ti,ab,kw (Word variations have been searched)

#14 #7 or #8 or #9 or #10 or #11 or #12 or #13

#15 MeSH descriptor: [Quality Control] explode all trees

#16 "quality":ti (Word variations have been searched)

#17 "validity":ti (Word variations have been searched)

#18 "measure*":ti (Word variations have been searched)

#19 "instrument*":ti (Word variations have been searched)

#20 "scale*":ti (Word variations have been searched)

#21 "score*":ti (Word variations have been searched)

#22 "assessment*":ti (Word variations have been searched)

#23 "apprais*":ti (Word variations have been searched)

#24 "tool":ti (Word variations have been searched)

#25 #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24

#26 #6 and #14 and #25

EMBASE (n=3698) (last search: October 27, 2017)

(peer review/exp OR 'peer review*') AND ('methodology'/exp OR 'research'/exp OR 'standards'/exp) AND (manuscript*:ab,ti OR article*:ab,ti OR publication*:ab,ti OR report:ab,ti OR reports:ab,ti OR submission*:ab,ti OR review*:ab,ti) AND ('quality control'/exp OR quality:ti OR validity:ti OR measure*:ti OR instrument*:ti OR scale*:ti OR score*:ti OR assessment*:ti OR apprais*:ti OR tool:ti) AND [embase]/lim

Google® Search (last search: October 20, 2017)

It was conducted using the following terms: peer review, report and quality. The first 200 links were investigated.

Appendix 2. Excluded studies and reasons for exclusion

| References* | Reasons for exclusion |
|---|--|
| 1. Ammenwerth E, Knaup P, Ulmer H, Wolff AC, Haux R. Developing and evaluating criteria to help reviewers of biomedical informatics manuscripts. <i>Informatik Biometrie und Epidemiologie in Medizin und Biologie</i> . 2003;10(5):512-14. | No tool of interest (Criteria for reviewers to support an objective high-quality review) |
| 2. Baxt WG, Waeckerle JF, Berlin JA, Callaham ML. Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. <i>Annals of emergency medicine</i> . 1998;32(3):310-7. | No tool of interest (Number of errors oppositely introduced by the editors) |
| 3. Blank RM. The effects of double-blind versus single-blind reviewing: Experimental evidence from the <i>American Economic Review</i> . <i>The American Economic Review</i> . 1991;81(5):1041-67. | No outcome of interest (Paper acceptance rate) |
| 4. Bornmann L, Daniel HD. Do author-suggested reviewers rate submissions more favorably than editor-suggested reviewers? a study on atmospheric chemistry and physics. <i>PLoS ONE</i> . 2010;5(10):1-8. | No tool of interest (Assessment of manuscript) |
| 5. Callaham M. Training of peer reviewers: validation of a 5-point rating scale. <i>PLoS medicine</i> . 2007;4:e166. | Type of reference (Note of the author to his published manuscript) |
| 6. Cohen IT, Patel K. Peer review interrater concordance of scientific abstracts: A study of anesthesiology subspecialty and component societies. <i>Anesthesia and Analgesia</i> . 2006;102(5):1501-3. | No tool of interest (Assessment of abstract) |
| 7. Cummings P. Effects of differences between peer reviewers suggested by authors and by editors. <i>JAMA</i> . 2006;296(10):1231-2. | No tool involved |
| 8. Das Sinha S, Sahni P, Nundy S. The effect of informing referees that their comments would be exchanged on the quality of their reviews (abstract) [Internet]. 1997 Available from: https://peerreviewcongress.org/abstracts_1997.html#review | Abstract of an included study |
| 9. Earnshaw JJ, Farndon JR, Guillou PJ, Johnson CD, Murie JA, Murray GD. A comparison of reports from referees chosen by authors or journal editors in the peer review process. <i>Annals of the Royal College of Surgeons of England</i> . 2000;82(4 Suppl):133-5. | No tool of interest (Assessment of manuscript) |
| 10. Fisher M, Friedman SB, Strauss B. The effects of blinding on acceptance of research papers by peer review. <i>Journal of the American Medical Association</i> . 1994;272(2):143-6. | No tool of interest (Assessment of manuscript) |
| 11. Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. <i>JAMA</i> . 1998;280(3):237-40. | No tool of interest (Number of weaknesses oppositely introduced by the editors) |
| 12. Godlee F, Gale CR, Martyn CN. The effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial [abstract] [Internet]. 1997 Available from: https://peerreviewcongress.org/abstracts_1997.html#thtr | No tool of interest (Number of weaknesses oppositely introduced by the editors) |
| 13. Green SM, Callaham ML. Implementation of a journal peer reviewer stratification system based on quality and | No tool of interest (Peer Reviewer Stratification System) |

| | |
|--|---|
| reliability. <i>Annals of emergency medicine</i> . 2011;57(2):149-52. | |
| 14. Groves T. Best practice in peer review and editing, ensuring article quality. <i>Notfall und Rettungsmedizin</i> . 2010;13(1):6-8. | No tool involved |
| 15. Helton M, Balistreri W. Assessment of reviewers recommended by authors vs editors: is there bias? (abstract) [Internet]. 2009 Available from: https://peerreviewcongress.org/abstracts_2009.html#81 | No outcome of interest |
| 16. Hwang K, Hwang SH. Is Double-Blinded Peer Review Necessary? The Effect of Blinding on Review Quality. <i>Plastic and reconstructive surgery</i> . 2016;138(1):161e-2e. | No tool involved |
| 17. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary?. <i>Controlled clinical trials</i> . 1996;17(1):1-2. | No tool of interest (Assessment of RCTs report) |
| 18. Janke KK, Bzowyckj AS, Traynor AP. Editors' perspectives on enhancing manuscript quality and editorial decisions through peer review and reviewer development. <i>American Journal of Pharmaceutical Education</i> . 2017;81(4):73. | No outcome of interest (Manuscript quality) |
| 19. Jurkat-Rott K, Lehmann-Horn F. Reviewing in science requires quality criteria and professional reviewers. <i>European journal of cell biology</i> . 2004;83(3):93-5. | No tool involved |
| 20. Lee SS. How to be a great reviewer: an editor's view. <i>Liver International</i> . 2008;28(2):158-9. | No tool involved |
| 21. Marchionini G. Rating reviewers. <i>Science</i> . 2008;319(5868):1335-6. | No tool involved |
| 22. McNutt R, Glass RM. Peer reviewer recommendations and ratings of manuscript quality for accepted and rejected manuscripts (abstract) [Internet]. 2001. Available from: https://peerreviewcongress.org/abstracts_2001.html#rejected | No tool involved |
| 23. Moore A. What's in a peer review report?. <i>Bioessays</i> . 2013;35(2):77-. | No tool involved |
| 24. Okike K, Hug KT, Kocher MS, Leopold SS. Single-blind vs double-blind peer review in the setting of author prestige. <i>JAMA</i> . 2016;316(12):1315-6. | No tool of interest (Number of errors oppositely introduced by the editors) |
| 25. Open peer review is feasible and does not reduce quality of reviews. <i>BMJ</i> . 1999;318:d. | Type of reference (Part of the introductory page "This week in the BMJ") |
| 26. Parikh L, Benner RS, Riggs TW, Chescheir NC. Factors influencing review quality and reviewer recommendation for a high-impact ob-gyn journal. <i>Obstetrics and Gynecology</i> . 2016;127:139S. | No tool involved |
| 27. Polak JF. The role of the manuscript reviewer in the peer review process. <i>AJR. American journal of roentgenology</i> . 1995;165(3):685-8. | No tool of interest (Monitor reviewer's performance) |
| 28. Resnik DB, Elmore SA. Ensuring the Quality, Fairness, and Integrity of Journal Peer Review: A Possible Role of Editors. <i>Science and engineering ethics</i> . 2016;22(1):169-88. | No tool involved |
| 29. Richards D. Little evidence to support the use of editorial peer review to ensure quality of published research. <i>Evidence-based dentistry</i> . 2007;8(3):88-9. | No outcome of interest (Manuscript quality) |
| 30. Rogers LF. Peer reviewers: reviewing manuscripts for the AJR. (editorial) <i>AJR</i> 2002;178(5):1051-1052 | No tool of interest (Assessment of manuscript) |

| | |
|--|--|
| 31. Shauver MJ, Chung KC. Reply: Is Double-Blinded Peer Review Necessary? The Effect of Blinding on Review Quality. <i>Plastic and reconstructive surgery</i> . 2016;138(1):162e-3e. | No tool involved |
| 32. Silobrčić V. Relative scales and their possible use in evaluation of scientific research in a small scientific community. <i>Acta Medica Croatica</i> . 2004;58(3):173-6. | No tool of interest (Assessment of manuscript) |
| 33. Szekely T, Kruger O, Krause ET. Errors in science: the role of reviewers. <i>Trends in ecology & evolution</i> . 2014;29(7):371-3. | No tool involved |
| 34. Tonks A. Reviewers chosen by authors. May be better than reviewers chosen by editors. <i>British Medical Journal</i> . 1995;311(6999):210. | No tool of interest (Evaluation of journal's review process) |

*In alphabetical order

Appendix 3. Included studies

1. Almquist M, Von Allmen RS, Carradice D, Oosterling SJ, McFarlane K, Wijnhoven B. A prospective study on an innovative online forum for peer reviewing of surgical science. *PLoS One*. 2017;12(6):1–13.
2. Berquist T. Improving your reviewer score: It's not that difficult. *AJR*. 2017;209:711–2.
3. Bingham CM, Higgins G, Coleman R, Van Der Weyden MB. The Medical Journal of Australia internet peer-review study. *The Lancet*. 1998;352(9126):441-5.
4. Black N, Van Rooyen S, Godlee F, Smith R, Evans S. What makes a good reviewer and a good review for a general medical journal?. *JAMA*. 1998;280(3):231-3.
5. Callaham ML, Baxt WG, Waeckerle JF, Wears RL. Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *JAMA*. 1998;280(3):229-31.
6. Callaham ML, Knopp R, John Gallagher E. Effect of written feedback by editors on quality of reviews: Two randomized trials. *JAMA*. 2002;287(21):2781–3.
7. Callaham ML, McCulloch C. Longitudinal trends in the performance of scientific peer reviewers. *Ann Emerg Med*. 2011;57(2):141-8.
8. Callaham ML, Schriger DL. Effect of Structured Workshop Training on Subsequent Performance of Journal Peer Reviewers. *Ann Emerg Med*. 2002;40(3):323–8.
9. Callaham ML, Tercier J. The Relationship of Previous Training and Experience of Journal Peer Reviewers to Subsequent Review Quality. 2007;4(1):0032–40.
10. Callaham ML, Wears RL, Waeckerle JF. Effect of attendance at a training session on peer reviewer quality and performance. *Ann Emerg Med*. 1998;32(3):318-22.
11. Chung KC, Shauver MJ, Malay S, Zhong L, Weinstein A, Rohrich RJ. Is Double-Blinded Peer Review Necessary? the Effect of Blinding on Review Quality. *Plast Reconstr Surg*. 2015;136(6):1369–77.
12. Das Sinha S., Sahni P, Nundy S. Does exchanging comments of Indian and non-Indian reviewers improve the quality of manuscript reviews ? *Natl Med J India*. 1999;12(5):210–3.
13. Evans AT, McNutt RA, Fletcher SW, Fletcher RH. The characteristics of peer reviewers who produce good-quality reviews. *J Gen Intern Med*. 1993;8(8):422-8.
14. Feurer ID, Becker GJ, Picus D, Ramirez E, Darcy MD, Hicks ME. Evaluating peer reviews. *JAMA*. 1994;272(2):98-100.
15. Friedman DP. Manuscript peer review at the *AJR*: facts, figures, and quality assessment. *AJR*. 1995;164(4):1007-9.
16. Hettyey A, Griggio M, Mann M, Raveh S, Schaedelin FC, Thonhauser KE, et al. Peerage of Science: Will it work? *Trends Ecol Evol*. 2012;27(4):189–90.
17. Henly SJ, Bennett JA, Dougherty MC. Scientific and statistical reviews of manuscripts submitted to *Nursing Research*: Comparison of completeness, quality, and usefulness. *Nurs Outlook*. 2010;58(4):188-99.
18. Henly SJ, Dougherty MC. Quality of manuscript reviews in nursing research. *Nurs Outlook*. 2009 Jan 1;57(1):18-26.

19. Houry D, Green S, Callaham M. Does mentoring new peer reviewers improve review quality? A randomized trial. *BMC Med Educ.* 2012;12(83):1–7
20. Jawaid SA, Jawaid M, Jafary MH. Characteristics of reviewers and quality of reviews: a retrospective study of reviewers at Pakistan Journal of Medical Sciences. *Pak J Med Sci.* 2006;22(2):101-6.
21. Justice AC, Cho MK, Winker MA, Berlin JA, Rennie D, Peer Investigators. Does masking author identity improve peer review quality?: A randomized controlled trial. *JAMA.* 1998;280(3):240-2.
22. Kliever MA, Freed KS, DeLong DM, Pickhardt PJ, Provenzale JM. Reviewing the reviewers: comparison of review quality and reviewer characteristics at the American Journal of Roentgenology. *AJR.* 2005;184(6):1731-5.
23. Kowalczyk MK, Dudbridge F, Nanda S, Harriman SL, Patel J, Moylan EC. Retrospective analysis of the quality of reports by author-suggested and non-author-suggested reviewers in journals operating on open or single-blind peer review models. *BMJ Open.* 2015;5(e008707):1–10
24. Landkroon AP, Euser AM, Veeken H, Hart W, Overbeke AJ. Quality assessment of reviewers' reports using a simple instrument. *Obstet Gynecol.* 2006;108(4):979-85.
25. Moore A, Jones R. Supporting and enhancing peer review in the BJGP. *Br J Gen Pract.* 2014;64(624):e459-61.
26. McNutt RA, Evans AT, Fletcher RH, Fletcher SW. The effects of blinding on the quality of peer review: a randomized trial. *JAMA.* 1990;263(10):1371-6.
27. Rajesh A, Cloud G, Harisinghani MG. Improving the quality of manuscript reviews: Impact of introducing a structured electronic template to submit reviews. *AJR.* 2013;200(1):20-3.
28. Pitkin RM, Burmeister LF. Identifying manuscript reviewers: randomized comparison of asking first or just sending. *JAMA.* 2002;287(21):2795-6.
29. Prechelt L. Review Quality Collector [Internet]. Available from: <https://reviewqualitycollector.org/static/pdf/rqdef-example.pdf>
30. Publons. Publons for Editors : Overview [Internet]. Available from: <https://static1.squarespace.com/static/576fcda2e4fcb5ab5152b4d8/t/58e21609d482e9ebf98163be/14912117>
31. Rivara FP, Cummings P, Ringold S, Bergman AB, Joffe A, Christakis DA. A Comparison of Reviewers Selected by Editors and Reviewers Suggested by Authors. *J Pediatr.* 2007;151(2):202–5.
32. Schroter S, Black N, Evans S, Carpenter J, Godlee F, Smith R. Effects of training on quality of peer review: randomised controlled trial. *BMJ.* 2004;328(7441):673.
33. Schroter S, Tite L, Hutchings A, Black N. Differences in review quality and recommendations for publication between peer reviewers suggested by authors or by editors. *JAMA.* 2006;295(3):314-7
34. Schroter S, Black N, Evans S, Godlee F, Osorio L, Smith R. What errors do peer reviewers detect, and does training improve their ability to detect them? *J R Soc Med.* 2008;101(10):507–14.
35. Shattell MM, Chinn P, Thomas SP, Cowling WR. Authors' and Editors' Perspectives on Peer Review Quality in Three Scholarly Nursing Journals. *J Nurs Scholarsh.* 2010;42(1):58–65.
36. Stossel TP. Reviewer status and review quality. *N Engl J Med.* 1985;312(10):658–9.
37. Thompson SR, Agel J, Losina E. The JBS Peer-Review Scoring Scale : A valid, reliable instrument for measuring the quality of peer review reports. *Learn Publ.* 2016;29:23–5.

38. Van Rooyen S, Black N, Godlee F. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *J Clin Epidemiol*. 1999;52(7):625-9.
39. Van Rooyen S, Delamothe T, Evans SJ. Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *BMJ*. 2010;341:c5729.
40. Van Rooyen S, Godlee F, Evans S, Black N, Smith R. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ*. 1999;318(7175):23-7.
41. Van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of Blinding and Unmasking on the Quality of Peer Review. *JAMA*. 1998;280(3):234-7.
42. Vinther S, Nielsen OH, Rosenberg J, Keiding N, Schroeder T V. Same review quality in open versus blinded peer review in ' Ugeskrift for Læger '. *Dan Med J*. 2012;59(8):1-5.
43. Wager E, Parkin EC, Tamber PS. Are reviewers suggested by authors as good as those chosen by editors? Results of a rater-blinded, retrospective study. *BMC Med*. 2006;32(3):61-4.
44. Walsh E, Rooney M, Appleby L, Wilkinson G. Open peer review : a randomised controlled trial. *Br J Psychiatry*. 2000;176:47-51.
45. Weber EJ, Katz PP, Waeckerle JF, Callaham ML. Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA*. 2002;287(21):2790-3.
46. Yang Y. Effects of Training Reviewers on Quality of Peer Review: A Before-and-After Study (Abstract) [Internet]. 2009. Available from: https://peerreviewcongress.org/abstracts_2009.html

Appendix 4. Classification of peer review report quality components

| Peer Review Report Quality Components | N. of tools |
|--|-------------|
| <p>1. Relevance of the study</p> <ul style="list-style-type: none"> • Did the reviewer discuss the importance of the research question? (1,2) • Importance of question: Did the reviewer pay appropriate attention to the importance of the research question? (3) • Did the reviewer give appropriate attention to the importance of the question? (4) • Usefulness of the study (5) • How well the review addressed the clinical or research importance of the study? (6) • Discussion: measures the merit of identifying and discussing the importance, implications or improvements of the research (7) • Theoretical framework (8,9) • Literature review/ references (8,9) | 9 |
| <p>2. Originality of the study</p> <ul style="list-style-type: none"> • Did the reviewer discuss the originality of the paper? (1,2) • Problem statement/originality (8,9) • Originality of manuscripts;(5) | 5 |
| <p>3. Interpretation of the results</p> <ul style="list-style-type: none"> • Did the reviewer comment on the author's interpretation of the results? (1,2) • The reviewer commented accurately and productively on the quality of the author's interpretation of the data, including acknowledgment of the data's limitations. (10) • Discussion/ Interpretation of results (8,9) • Interpretation of results.(5) | 6 |
| <p>4. Strengths and weaknesses of the study</p> | |
| <p>4.1 General</p> <ul style="list-style-type: none"> • Comments enhance the merits and relevance of the work (11) • Strong and weak points (5) • How well it identified the study's strengths and weaknesses? (6) • Merits: measures how accurate and justified the review is in identifying manuscript strengths (7) • Critique: measures how accurate the and justified the review is in identifying manuscript weaknesses (7) • Shortcomings identified (12) • Target key issues: Did the reviewer concisely identify the key issues that either make or break the manuscript (from the editor's perspective)? (3) • Did the reviewer target key issues? (4) | 7 |

| | |
|--|---|
| <p>4.2 Methods</p> <ul style="list-style-type: none"> • Did the reviewer clearly identify the strengths and weaknesses of the methods (study design, data collection and data analysis)? (1,2) • The reviewer identified and commented on major strengths and weaknesses of study design and methods. (10) • Methods: Did the reviewers clearly identify the strengths and weaknesses of the a. study design; b. data collection and c. data analysis (3) • Did the reviewer clearly identify strengths and weaknesses in the study's methods? (4) • Design/methods (8,9) | 7 |
| <p>4.3 Statistical methods</p> <ul style="list-style-type: none"> • Persons (9) • Occasions (9) • Variables and measurement (9) • Procedure (9) • Data quality (9) • Model/estimator/assumptions (9) • Confidence intervals/p-values/overall fit (9) • Statistical interpretation (9) • Tables (9) • Graphs (9) | 1 |
| <p>5. Presentation and organization of the manuscript</p> <ul style="list-style-type: none"> • Did the reviewer make specific useful comments on the writing, organisation, tables and figures of the manuscript? (1,2) • The reviewer commented on mayor strengths and weaknesses of the manuscript as a written communication, independent of the design, methods, results, and interpretation of the study.(10) • Presentation: Are there any constructive suggestions on improvement of a. writing; b. data presentation and c. interpretation (3) • Did the reviewer make a constructive comments about the quality of writing and presentation of data? (4) • Data analysis/Presentation (8,9) • Organization/writing (8,9) • Tables and figures (5) | 8 |
| <p>6. Structure of reviewer's comments</p> <ul style="list-style-type: none"> • Sophisticated detailed comments to the author by section with line and page references (11) • Concise well-organized comments to the editor (11) • Section-by-section review (13) • Structure (14) • Consistent with journal's review criteria (12) | 4 |
| <p>7. Characteristics of reviewers' comments</p> <p>7.1 Clarity</p> <ul style="list-style-type: none"> • Clear (8,9) • How clear was this review? The review was easily read and interpreted by the editor and authors. (17) | 3 |

| | |
|---|----|
| <p>7.2 Constructiveness</p> <ul style="list-style-type: none"> • Were the reviewer's comments constructive? (1,2) • The reviewer's comments to author were constructive and professional. (10) • Constructive; (8,9) • Constructiveness (4,14) • How helpful was this review? Comments were constructive, relevant, and realistic.(17) | 9 |
| <p>7.3 Detail/ Thoroughness</p> <ul style="list-style-type: none"> • The amount of detail;(16) • Level of sophistication of the commentary; (16) • Detail of commentary (12) • General: Was the reviewer a. thorough; (3) • Thoroughness (4) • Precise; (8,9) • How thorough was this review? The review gave adequate consideration to all aspects of the paper including methodology, figures, interpretation and presentation of results, ethics, relevance, etc. (17) • Did the reviewer supply appropriate evidence using examples from the paper to substantiate their comments? (1,2) • Offering supporting references (13) • Did reviewers supply evidence to support their statements? (6) • Logical; (8,9) | 11 |
| <p>7.4 Fairness</p> <ul style="list-style-type: none"> • Fair; (3) • Fairness (4) • Balanced/fair;(8,9) • Objectivity;(14) | 5 |
| <p>7.5 Knowledgeable</p> <ul style="list-style-type: none"> • Knowledgeable (3) • Knowledge of the manuscript's content area. (4) • Knowledgeable/substantiated;(8,9) • Understands content (12) | 5 |
| <p>7.6 Tone</p> <ul style="list-style-type: none"> • How would you rate the tone of the review? (2) • Etiquette (13) • Courteous (3) • Courteousness (4) • Overall tone of the reviewers was also assessed as harsh or courteous. (5) • Were reviewers courteous? (6) • Constructive tone (12) | 7 |
| <p>8. Timeliness of the review report</p> <ul style="list-style-type: none"> • Timely (14 days) or early review completion.(11) • Timeliness (13) • Aspect: Timeliness (15) • Punctuality of the review (16) • Turnaround time (14) • How timely was this review? The review assignment was completed within the time limits established by the editor.(17) • Time taken to review (<4 weeks) (18) | 7 |
| <p>9. Usefulness of the review report</p> <p>9.1 Decision making</p> <ul style="list-style-type: none"> • Grade sheet (13) | 6 |

| | |
|--|---|
| <ul style="list-style-type: none"> • Summary and/or recommendation (13) • The reviewer provided the editor with the proper context and perspective to make a decision about acceptance or revision of the manuscript. (10) • Summary grade (4) • Aspect: Helpfulness for Decision (weight 27): This aspect should be evaluated regardless of how useful the review will be as feedback to the authors. (15) • Usefulness to editor (8,9) | |
| <p>9.2 Manuscript improvement</p> <ul style="list-style-type: none"> • Aspect: Helpfulness for Authors (weight 19): This aspect is solely interested in how well the review aids the authors for improving their work and/or writing. Whether the review makes a good judgment regarding acceptance of the submission plays no role here whatsoever. (15) • Perceived Usefulness to authors (8,9) • ≥ 300 words or more than 4 suggestions for improvement (18) • Suggestions to correct errors (18) • Specific errors identified (18) • Better references (18) • The reviewer provided the author with useful suggestions for improvement of the manuscript. (10) • The quality of the suggestions for manuscript improvement; (16) • Specific suggestions (12) • Insight; (14) • New insights/perspectives (2/15%) (13) | 9 |

1. Black N, Van Rooyen S, Godlee F, Smith R, Evans S. What makes a good reviewer and a good review for a general medical journal? *JAMA*. 1998;280(3):231–3.
2. Van Rooyen S, Godlee F, Evans S, Black N, Smith R. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ*. 1999;318(7175):23–7.
3. Das Sinha S., Sahni P, Nundy S. Does exchanging comments of Indian and non-Indian reviewers improve the quality of manuscript reviews? *Natl Med J India*. 1999;12(5):210–3.
4. McNutt RA, Evans AT, Fletcher RH, Fletcher SW. The Effects of Blinding on the Quality of Peer Review. *JAMA*. 1990;263(10):1371–6.
5. Jawaid SA, Jawaid M, Jafary MH. Characteristics of reviewers and quality of reviews: A retrospective study of reviewers at Pakistan Journal of Medical Sciences. *Pakistan J Med Sci*. 2006;22(2):101–6.
6. Justice AC, Cho MK, Winker MA, Berlin JA. Does Masking Author Identity Improve Peer Review Quality? A Randomized Controlled Trial. *JAMA*. 1998;280(3):240–2.
7. Hettyey A, Griggio M, Mann M, Raveh S, Schaedelin FC, Thonhauser KE, et al. Peerage of Science: Will it work? *Trends Ecol Evol*. 2012;27(4):189–90.
8. Henly SJ, Dougherty MC. Quality of manuscript reviews in nursing research. *Nurs Outlook*. 2009;57(1):18–26.
9. Henly SJ, Bennett JA, Dougherty MC. Scientific and statistical reviews of manuscripts submitted to *Nursing Research*: Comparison of completeness, quality, and usefulness. *Nurs Outlook*. 2010;58(4):188–99.

10. Callaham M, Knopp R, John Gallagher E. Effect of written feedback by editors on quality of reviews: Two randomized trials. *JAMA*. 2002;287(21):2781–3.
11. Berquist T. Improving your reviewer score: It's not that difficult. *AJR*. 2017;209(4):711–2.
12. Shattell MM, Chinn P, Thomas SP, Cowling WR. Authors' and Editors' Perspectives on Peer Review Quality in Three Scholarly Nursing Journals. *J Nurs Scholarsh*. 2010;42(1):58–65.
13. Feurer I, Becker G, Picus D, Ramirez E, Darcy M, Hicks M. Evaluating peer reviews: Pilot testing of a grading instrument. *JAMA*. 1994;272(2):98–100.
14. Landkroon AP, Euser AM, Veeken H. Quality Assessment of Reviewers' Reports Using a Simple Instrument. *Obstet Gynecol*. 2006;108(4):979–85.
15. Prechelt L. Review Quality Collector [Internet]. Available from: <https://reviewqualitycollector.org/static/pdf/rqdef-example.pdf>
16. Kliever MA, Freed KS, DeLong DM, Pickhardt PJ, Provenzale JM. Reviewing the reviewers: Comparison of review quality and reviewer characteristics at the American Journal of Roentgenology. *AJR*. 2005;184(6):1731–5.
17. Publons. Publons for Editors: Overview [Internet]. Available from: <https://static1.squarespace.com/static/576fcda2e4fcb5ab5152b4d8/t/58e21609d482e9ebf98163be/14912117>
18. Yang Y. Effects of Training Reviewers on Quality of Peer Review: A Before-and-After Study (Abstract) [Internet]. 2009. Available from: https://peerreviewcongress.org/abstracts_2009.html

Appendix 5. Survey questionnaire



Welcome to the survey!

Although the peer review process plays a key role in research dissemination, only limited research has been conducted so far in this field.

The objective of this survey is to investigate the perspectives of biomedical editors and authors towards the **quality of peer review reports**. We hope this work will help us to develop a new tool to assess the quality of a peer review report in biomedical research.

Knowing your expertise, we would be very grateful if you could answer a few questions and share your opinion. The survey will take approximately **10 minutes** to complete. Your participation in this study is completely voluntary. If you decide to participate, all your answers will be de-identified and stored in a secured repository at Universitat Politècnica de Catalunya, Barcelona-Tech (Spain). The de-identified data from this study will be shared on Zenodo repository. In case you opt out of sharing your data, you will still be able to participate in the study.

This survey has received ethics approval from the Research Ethics Committee of the Universitat Politècnica de Catalunya, Barcelona-Tech (Spain).

This study is part of the **Methods in Research on Research (MiRoR)** project, a joint doctoral training programme in the field of clinical research funded by Marie Skłodowska-Curie Action <http://miror-ejd.eu/>. The objective of MiRoR project is to train future generations of scientists in Research on Research, a new discipline aiming to promote research integrity increasing research value and reducing waste in health research.

This study is conducted by **Cecilia Superchi**, a PhD student at Universitat Politècnica de Catalunya, Barcelona-Tech and Université Paris Descartes, Sorbonne Paris Cité in collaboration with [Prof. Darko Hren](#) (University of Split), [Prof. José Antonio Gonzalez](#) (Universitat Politècnica de Catalunya) and [Prof. Isabelle Boutron](#) (Université Paris Descartes).

If you have any questions about this study or your rights as a participant, you may contact by email Cecilia Superchi, cecilia.superchi@upc.edu or Darko Hren, dhren@ffst.hr

Do you agree to take part in the study?

- Yes, I agree
- No, I do not agree

Do you agree to share your de-identified data?

- Yes, I agree
- No, I do not agree
-

Definition of peer review report quality

The **quality of a peer review report** could be defined as "to what extent the peer review report helps editors to make a fair decision and authors to improve the quality of the submitted manuscript"

Do you agree with this definition?

- Yes
- No
- Partially

Please add your comments and ideas on how to improve the definition

Importance of the items to assess peer review report quality

The following items have been identified in a systematic review as possible quality components of a peer review report.

We are interested to know your opinion on the importance of these items, particularly whether the item should be included in a new tool assessing the quality of a peer review report.

Please rate the **IMPORTANCE** of each item in assessing the quality of a peer review report from 1 (not important) to 5 (very important).

We expect that for some items it will not be easy for you to make a clear decision about the importance of the item. In those cases we still invite you to offer your rating but you can elaborate on your decision. Furthermore we invite you to suggest potential improvements in wording of the items.

The reviewer's comments are **clear and easy to read**

Not important Slightly important Moderately important Important Very important
1 2 3 4 5

Please add any comments about your decision and/or wording of this item (not a mandatory field)

The reviewer knows and understands correctly **the content of the manuscript**

Not important Slightly important Moderately important Important Very important
1 2 3 4 5

Please add any comments about your decision and/or wording of this item (not a mandatory field)

The reviewer's comments are **constructive**

Not important Slightly important Moderately important Important Very important
1 2 3 4 5

Please add any comments about your decision and/or wording of this item (not a mandatory field)

The reviewer's comments are **detailed and thorough**

Not important Slightly important Moderately important Important Very important
1 2 3 4 5

Please add any comments about your decision and/or wording of this item (not a mandatory field)

The reviewer uses a **courteous tone**

Not important Slightly important Moderately important Important Very important
1 2 3 4 5

Please add any comments about your decision and/or wording of this item (not a mandatory field)

The reviewer comments on the **relevance of the study**

| Not important 1 | Slightly important 2 | Moderately important 3 | Important 4 | Very important 5 |
|-----------------------|-------------------------|---------------------------|-----------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Please add any comments about your decision and/or wording of this item (not a mandatory field)

New items to assess peer review report quality

Are there any other items to assess the quality of a peer review report that you think should be included?

Please list them.

Demographic characteristics

What is your gender?

- Woman
- Man
- Prefer not to answer
- Other (please specify)

What is your age?

What is the highest level of education obtained?

- Bachelor Degree
- Master Degree
- PhD
- MD or equivalent
- Prefer not to answer
- Other (please specify)

Author's characteristics

What is your job title at your institution?

- Researcher
- Assistant Professor
- Associate Professor
- Professor
- Other (please specify)

What type of institution are you affiliated at?

- Private University
- Public University
- Research Centre
- Other (please specify)

Where is the institution located?

- Europe
- North America
- South America
- Africa
- Asia
- Australia

Author's characteristics

How long have you been publishing scientific papers?

- <5 years
- 6-10 years
- 11-15 years
- 16-20 years
- >20 years

Do you also work as biomedical editor?

- Yes
- No

Are you involved in making decisions on the manuscripts received by your journal?

- Yes
 - No
-

Editor's characteristics

What is your job title at your journal?

- Editor in chief
- Associate editor
- Academic editor
- Section editor
- Deputy editor
- Other (please specify)

Are you involved in making decisions on the manuscripts received by your journal?

- Yes
- No

At what type of journal do you currently working as editor?

- General Journal
- Specialty Journal

Editor's characteristics

Where is the journal located?

- Europe
- North America
- South America
- Africa
- Asia
- Australia

How long have you been working as editor?

- <5 years
- 6-10 years
- 11-15 years
- 16-20 years
- >20 years

Does your work inside or outside the journal include authoring scientific papers?

- Yes
- No

Study results and next step

Please check which of the following options you would be interested in

- I would be interested in receiving the results of the present study
- I would be interested in participating in the validation study of a new tool for assessing the quality of a peer-review report

Please write down your name and email address. Your data will be **EXCLUSIVELY** used for the option(s) which you have previously chosen.

Name

Email address

Appendix 6. Invitation email

- **Invitation email for corresponding authors**

From:
Cc:
To:
Subject: Academic Survey on Peer Review

Dear researcher,

As corresponding author of the article recently published in [CUSTOM 1], we would like to invite you to participate in an **academic survey**.

The objective of this survey is to investigate the perspectives of biomedical editors and authors on the **quality of peer-review reports**. We hope this work will help us to develop a new tool to assess the quality of a peer-review report in biomedical research.

The survey will take approximately **10 minutes to complete**. Participation in this study is completely **voluntary** and you may withdraw at any time.

This study is part of the **Methods in Research on Research** (MiRoR) project, a joint doctoral training programme in the field of clinical research funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 676207

We would be very grateful if you would take the time to complete our survey. **Your insights** as an author are **essential** to us.

If you have any questions, comments or queries, please do not hesitate to contact us at cecilia.superchi@upc.edu or dhren@ffst.hr

We kindly thank you for your time, attention, and cooperation.

Sincerely,

Cecilia Superchi, PhD Student at Universitat Politècnica de Catalunya & Université Paris Descartes

Darko Hren, PhD, Prof. at University of Split

José Antonio Gonzalez, PhD, Prof. at Universitat Politècnica de Catalunya

Isabelle Boutron, MD, PhD, Prof. at Université Paris Descartes

- **Invitation email for biomedical editors**

From:
Cc:
To:
Subject: Academic Survey on Peer Review

Dear [Name] [Surname],

As [CUSTOM 1] at [CUSTOM 2], we would like to invite you to participate in an **academic survey on peer review**.

The objective of this survey is to investigate the perspectives of biomedical editors and authors on the **quality of peer-review reports**. We hope this work will help us to develop a new tool to assess the quality of a peer-review report in biomedical research.

The survey will take approximately **10 minutes to complete**. Participation in this study is completely **voluntary** and you may withdraw at any time.

You are also encouraged to **forward the link** of the survey to your colleagues who may be interested in participating in this study https://www.surveymonkey.com/r/REPORT_QUALITY_EDITORS

This study is part of the **Methods in Research on Research (MiRoR)** project, a joint doctoral training programme in the field of clinical research funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 676207 <http://miror-ejd.eu/>

We would be very grateful if you would take the time to complete our survey. **Your insights** as a biomedical editor are **essential** to us.

If you have any questions, comments or queries, please do not hesitate to contact us at cecilia.superchi@upc.edu or dhren@ffst.hr

We thank you kindly for your time, attention, and cooperation.

Sincerely,

Cecilia Superchi, PhD Student at Universitat Politècnica de Catalunya & Université Paris Descartes

Darko Hren, PhD, Prof. at University of Split

José Antonio Gonzalez, PhD, Prof. at Universitat Politècnica de Catalunya

Isabelle Boutron, MD, PhD, Prof. at Université Paris Descartes

Appendix 7. Top 30-biomedical journals with the highest impact factors

| Full Journal Title** | IF |
|--|------|
| New England Journal Of Medicine* | 79.3 |
| Lancet* | 53.3 |
| JAMA-Journal Of The American Medical Association | 47.7 |
| BMJ-British Medical Journal* | 23.3 |
| JAMA Internal Medicine | 20.0 |
| Annals Of Internal Medicine | 19.4 |
| Nature Reviews Disease Primers | 16.1 |
| Journal Of Cachexia Sarcopenia And Muscle | 12.5 |
| Plos Medicine | 11.7 |
| Bmc Medicine* | 9.1 |
| Mayo Clinic Proceedings* | 7.2 |
| Cochrane Database Of Systematic Reviews | 6.8 |
| Journal Of Internal Medicine | 6.8 |
| Canadian Medical Association Journal* | 6.2 |
| Journal Of Clinical Medicine* | 5.6 |
| American Journal Of Medicine* | 5.1 |
| Translational Research* | 4.9 |
| Annals Of Family Medicine* | 4.5 |
| Medical Journal Of Australia* | 4.2 |
| American Journal Of Preventive Medicine* | 4.1 |
| Amyloid-Journal Of Protein Folding Disorders | 4.0 |
| Journal Of General Internal Medicine* | 4.0 |
| Deutsches Arzteblatt International | 3.9 |
| Palliative Medicine | 3.8 |
| Preventive Medicine* | 3.5 |
| British Medical Bulletin | 3.4 |
| European Journal Of Internal Medicine* | 3.3 |
| British Journal Of General Practice* | 3.3 |
| Journal Of Pain And Symptom Management* | 3.2 |
| Qjm-An International Journal Of Medicine | 3.2 |

* Journal reporting the corresponding author in the PubMed abstract.

**Source: InCites Journal Citation Reports 2017 under the category "Medicine, general and internal".

Appendix 8. Complete participants characteristics

| Characteristics | Editors N=165 |
|--|------------------|
| Journal Role | |
| Editor-in-Chief | 50 (30.3%) |
| Associate Editor | 63 (38.2%) |
| Academic Editor | 7 (4.2%) |
| Section Editor | 6 (3.6%) |
| Deputy Editor | 12 (7.3%) |
| Other (e.g. Statistical Editor, Patient Editor) | 27 (16.4%) |
| Involvement in making decisions on the manuscript | |
| Yes | 144 (87.3%) |
| No | 21 (12.7%) |
| Type of Journal | |
| General Journal | 39 (23.6%) |
| Specialty Journal | 126 (76.4%) |
| Journal location | |
| Europe | 132 (80.0%) |
| North America | 23 (13.9%) |
| South America | 2 (1.2%) |
| Africa | 1 (0.6%) |
| Asia | 3 (1.8%) |
| Australia | 4 (2.4%) |
| Number of years of experience as editor | |
| <5 years | 74 (44.8%) |
| 6-10 years | 46 (27.9%) |
| 11-15 years | 27 (16.4%) |
| 16-20 years | 7 (4.2%) |
| >20 years | 11 (6.7%) |
| Authorship of scientific papers | |
| Yes | 141 (85.5%) |
| No | 24 (14.5%) |

| Characteristics | Authors N=224 |
|--|------------------|
| Occupation | |
| Professor | 63 (28.1%) |
| Associate Professor | 31 (13.8%) |
| Assistant Professor | 34 (15.2%) |
| Researcher | 47 (21.0%) |
| Other (e.g. Lecturer, Postdoc, PhD) | 49 (21.9%) |
| Type of Institution | |
| Public University | 134 (59.8%) |
| Private University | 33 (14.7%) |
| Research Centre | 17 (7.6%) |
| Other (e.g. Hospital) | 40 (17.9%) |
| Institution location | |
| Europe | 87 (38.8%) |
| North America | 95 (42.4%) |
| South America | 5 (2.2%) |
| Africa | 1 (0.4%) |
| Asia | 11 (4.9%) |
| Australia | 25 (11.2%) |
| Number of years of experience as author | |
| <5 years | 36 (16.1%) |
| 6-10 years | 51 (22.8%) |
| 11-15 years | 34 (15.2%) |
| 16-20 years | 19 (8.5%) |
| >20 years | 84 (37.5%) |
| Employment as biomedical editor | |
| Yes | 63 (28.1%) |
| No | 161 (71.9%) |
| Involvement in making decisions on the manuscript | |
| Yes | 56 (88.9%) |
| No | 7 (11.1%) |

Appendix 9. Survey questionnaire

A video explaining how to fill in the questionnaire is available at <https://www.youtube.com/watch?v=c1119EB8-oM&feature=youtu.be>

PSYCHOMETRIC TESTING OF ARCADIA

Intro Instructions 1. Quality assessment 2. ARCADIA validation 3. Demographic questions

WELCOME

Thank you again for taking part in the first phase of our research. Based on your valuable response we were able to develop a new tool to measure the quality of peer review reports, the **ARCADIA** (Assessment of Review reports with a Checklist Available to eDitors and Authors). We would now like to invite you to take part in a subsequent survey which will enable us to evaluate its psychometric properties.

The survey will take approximately **20 minutes** to complete.

Your participation in this study is completely voluntary. If you decide to participate, all your answers will be de-identified and stored in a secured repository at Universitat Politècnica de Catalunya, Barcelona-Tech (Spain). The de-identified data from this study will be shared on Zenodo repository. If you opt out of sharing your data, you will still be able to participate in the study.

This survey has received ethics approval from the Research Ethics Committee of the Universitat Politècnica de Catalunya, Barcelona-Tech (Spain).

If you have any questions about this study or your rights as a participant, please email [Cecilia Superchi](mailto:Cecilia.Superchi).

Research team

This study is conducted by **Cecilia Superchi**, a PhD student at Universitat Politècnica de Catalunya, Barcelona-Tech and Université de Paris in collaboration with:

- Ketelevan Glonti, PhD student at University of Split and Université de Paris
- Sara Schroter, Senior researcher at BMJ
- Alessandro Recchioni, Senior editor at BMC Medicine
- Josep Antoni Sánchez, Assistant prof. at Universitat Politècnica de Catalunya
- Darko Hren, Prof. at University of Split
- Isabelle Boutron, Prof. at Université de Paris
- José Antonio Gonzalez, Prof. at Universitat Politècnica de Catalunya

Related research:

1. Superchi C, Hren D, Blanco D, Rius R, Recchioni A, Boutron I, González JA. The development of ARCADIA: a tool for assessing the quality of peer review reports in biomedical research. Submitted.
2. Superchi C, González JA, Solà J, Cobo E, Hren D, Boutron I. Tools used to assess the quality of peer review reports: a methodological systematic review. BMC Med Res Methodol. 2019;19(48):1-14. DOI: <https://doi.org/10.1186/s12874-019-0688-x>

Do you agree to take part in the study?

- Yes, I agree
- No, I do not agree

Do you agree to share your de-identified data?

- Yes, I agree
- No, I do not agree

[Start the survey](#)

The MiRoR Project



This study is part of the **Methods in Research on Research (MiRoR)** project, a joint doctoral training programme in the field of clinical research funded by Marie Skłodowska-Curie Action.

PSYCHOMETRIC TESTING OF ARCADIA

Intro Instructions 1. Quality assessment 2. ARCADIA validation 3. Demographic questions

[Save](#) [Exit](#) [Submit](#)

How is the questionnaire structured?

The questionnaire is divided into **three parts** (1. Quality assessment; 2. ARCADIA validation and 3. Demographic questions). The approximate time to complete each part is indicated in brackets in the box below. You can **stop and re-enter the questionnaire at any moment**.

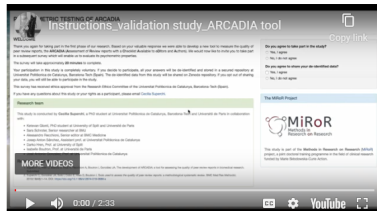
1. Quality assessment: you are asked to assess the quality of a **peer review report** using: 1) a subjective scale from 0 (extremely poor peer review report) to 100 (excellent peer review report) and 2) the ARCADIA tool. You are also asked to indicate the **time** you have spent to complete the assessment using ARCADIA. To facilitate this task, a **chronometer** is provided to you (a **🕒** button appears in the page, when you assess the peer review report using ARCADIA). ([Task 1, approx. 15 min](#))
2. ARCADIA validation: you are asked to indicate which ARCADIA items need to be revised and how they could be improved. ([Task 2, approx. 4 min](#))
3. Demographic questions: you are asked some questions related to your demographic characteristics. ([Task 3, approx. 1 min](#))

Tips

Please consider the following tips before starting the questionnaire:

- Use the **Save** button to make sure your answers are always saved
- Avoid quitting the questionnaire abruptly, since unsaved data can be lost. If you need to interrupt the questionnaire, click on **Save** and then **Exit** button. You can restart the questionnaire right where you left off the last time.
- Use the **Submit** button when you have completed the questionnaire.
- How to use the chronometer?
 - Click on the **🕒** button to show or hide the chronometer.
 - You can drag the chronometer through the screen using the mouse.
 - Use the **Start** button (**▶**) when you begin the quality assessment of the peer review report using ARCADIA.
 - Use the **Reset** button (**⏪**) to set the clock to zero;
 - Use the **Pause** button (**⏸**) to interrupt the chronometer.
 - Keep in mind that even if the chronometer is hidden, it continues to run.

Instructions for filling in the questionnaire (2:33 min.)



PSYCHOMETRIC TESTING OF ARCADIA

Intro Instructions 1. Quality assessment 2. ARCADIA validation 3. Demographic questions

Save Exit Submit


Review

Thank you for sending me this paper to review. It is on an important subject, and is generally written in a clear style. However, I regret to say that it does not add usefully to existing knowledge.

Major points

1. Originality - the paper adds little to existing literature which is very thoroughly reviewed and cited. I suggest that fewer than 40 references are needed for this type of paper, and that older historical papers need not be mentioned. Despite the problems of confidentiality, the greater use today of fax, email and computerised patient records surely make some of the conditions assumed in older papers obsolete, and even some conditions of the trial which is 4-5 years old. Although the trial design is perhaps uniquely original, it raises problems discussed later.
2. Importance of the work to general readers. The topic is important to general readers but the research reported is not well enough designed to be useful to them.

Scientific reliability



Quality assessment 1: Subjective scale Quality assessment 2: The ARCADIA tool Items explanation

Please assess the quality of the report using the ARCADIA tool. To know how to use ARCADIA, please click on the button.

Hide or show the chronometer using the button below:

Item: < 1a >

Domain: Importance of the study

In the peer review report, did the reviewer comment on the contribution of the study to scientific knowledge?

Yes

Completed



PSYCHOMETRIC TESTING OF ARCADIA

Intro Instructions 1. Quality assessment 2. ARCADIA validation 3. Demographic questions

Save Exit Submit

The ARCADIA tool

In the peer review report, did the reviewer comment on...

| | | |
|---|--|--|
| Domain 1. Importance of the study | the contribution of the study to scientific knowledge? (item 1a) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | whether the relevant literature was accurately reviewed? (item 1b) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| Domain 2. Robustness of the study methods | the soundness of the study methods (e.g., study design, outcome measures, risk of bias)? (item 2a) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | the suitability of the statistical methods? (item 2b) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| Domain 3. Interpretation and discussion of the study results | whether the study conclusions answer the research question(s) and correctly summarize the study results? (item 3a) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | whether the study limitations are acknowledged? (item 3b) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |
| | the applicability and generalizability (external validity) of the study results? (item 3c) | <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> NA |

1. Please indicate if you feel any ARCADIA items need revising (by ticking them in the drop-down list below).
2. If so, write down how they may be improved.
3. When you are done or if you believe changes are not needed, jump to the next tab, "3. Demographic questions"

Nothing selected

Suggestions for improvement

Enter a text ...

Save text

Remember to save the text after completing your suggestions. Should you wish to change them, just modify the text and save it again.

PSYCHOMETRIC TESTING OF ARCADIA

Intro Instructions 1. Quality assessment 2. ARCADIA validation 3. Demographic questions

Save Exit Submit

What is your gender?

Choose ...

What is your age?

Choose ...

Where are you located?

Choose ...

What is your highest level of education?

Choose ...

How long have you been working as author and/or editor in the biomedical field?

Choose ...

Have you ever received training on how to review a manuscript?

Choose ...

How many manuscripts do you usually publish in a year?

Choose ...

How many manuscripts do you usually review in a year?

Choose ...

Appendix 10. Invitation email for biomedical editors and authors who agreed to take part in the validation study

From:
Cc:
To:
Subject: Validation study_ ARCADIA tool

Dear [Name] [Surname],

We are writing to you because between November 2018 and February 2019 you participated in an online survey about the development of a new tool for assessing the quality of peer review reports. In the survey, you kindly agreed to take part in the subsequent validation study of the new tool.

We are now inviting you to participate in the validation study of our newly developed tool, for assessing peer review report quality (ARCADIA).

The objective of this survey is to field test the tool and to evaluate its acceptability, reliability and validity. The survey will take approximately 20 minutes to complete

If you have any questions, comments or queries, please do not hesitate to contact Cecilia at cecilia.superchi@upc.edu

Participation in this study is completely voluntary and you may withdraw at any time.

Sincere thanks,

Cecilia Superchi, PhD Student at Universitat Politècnica de Catalunya and Université de Paris

Ketevan Glonti, PhD Student at University of Split and Université de Paris

Sara Schroter, Senior researcher at BMJ

Alessandro Recchioni, Senior editor at BMC Medicine

Darko Hren, Prof. at University of Split

Isabelle Boutron, Prof at. Université de Paris

José Antonio Gonzalez, Prof. at Universitat Politècnica de Catalunya

This study is part of the **Methods in Research on Research (MiRoR)** project, a joint doctoral training programme in the field of clinical research funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 676207 <http://miror-ejd.eu/>

Appendix 11. Participants' feedback on ARCADIA

| | |
|----|--|
| 1 | I think the scoring tool need to consider the suitability of the review report in light of the pertinent journal (e.g., a peer review report for BMJ cannot be the same as one for a tiny circulation subspecialty journal). |
| 2 | Add more precise comments |
| 3 | Helpful questions, though not sure always imperative to think of all of the questions. If something wasn't mentioned, such as limitations, I assume that there was nothing noteworthy to mention there. |
| 4 | For the editor, but also for the authors, the comments should be numbered. In order for the comments to be constructive, suggestions for improvement should be provided. A summary of the major strengths and weaknesses of the manuscript should be requested from reviewers. |
| 5 | This "question" 3a has two questions embedded in it: whether the study answers the research question and whether it summarizes study results. These should be two separate questions. The same happens in 4b: asking the reviewer to judge (generally) if the study is reproducible is one thing. Asking them to verify reporting guideline adherence is another (huge) task. |
| 6 | In general, the items are too concise and comments are not provided for all. There would be a need for an Explanation appendix. |
| 7 | No suggestions. The tool seems useful, although some questions may require creative interpretation when applied to any given review. This is to say that while attempting objectivity in specifying criteria for review, the questionnaire forces subjectivity in interpretation of individual criteria, which kind of defeats the original goal. |
| 8 | Off course, since I do not have the manuscript I cannot comment on the correctness of the reviews comment. So this seems to be a test of my awareness of the what has been written in this case,. The relevance of this is not obvious to me |
| 9 | I appreciate the intent of the binary YES/NO responses, but in practice many of these questions could have partial answers. |
| 10 | 4b) acronym "RG" used in the online phrasing: unclear 4d) the suggestion was to tick NA if the reviewer didn't mention it. Unclear why there was a "no" choice provided as well. |
| 11 | This only seems relevant for RCT. What about observational studies? |
| 12 | I was surprised to find "study protocol" because I had not understood that ARCADIA was only for clinical research. There can only be deviations from a study protocol that has been approved, authorized, and registered before the study begins, as is done in clinical research. In basic research, unexpected discoveries and dead ends are normal phenomena, so the study design mutates as the research unfolds: one cannot speak of deviating from a "protocol" in every research field. This question will be difficult to answer. Not all research fits into a category for which a reporting guideline has been written. The question should be reworded to ask either about adherence to reporting guidelines whenever available, or about completeness of reporting for reproducibility. For presentation and organization, it is difficult to answer yes/no because a reviewer may make a few superficial comments. Here, a scale of 1 to 5 would be easier to use. Clarity, too, is difficult to dichotomize as yes/no. |
| 13 | I did not know what the abbreviation RG was in your questions. Please do not use abbreviations. I would suggest that it is important for a reviewer to flag up areas that they have not got expertise to assess - for example this reviewer implied that the statistical methods should be checked by an expert. It is not appropriate to expect peer reviewers to cover all areas. |

| | |
|----|--|
| 14 | Overall, I found the questions tended to put me a box. In many cases, the answers were yes, somewhat. For example, this reviewer did a great job on bias, but for literature review, commented on the accuracy of the articles he read, but not on whether it was comprehensive. He also noted where apparently a limitation was stated, but didn't go into specifically what should be added to the limitations. Anytime a question had multiple things to consider, it was hard to say yes or no (e.g. organization, no, but tables, yes). I also think that a question that is pertinent is whether the reviewer adequately summarized the paper - this is very helpful to both the author and editor reassuring them that the reviewer took care with the paper. I think its awn essential part of a good review. Specific questions: Relevant literature - This question is a bit unclear - it asks if the reviewer commented on whether the literature was "accurately" reviewed, but this isn't the same as "comprehensively reviewed" which is equally important as we want to know they are aware of any other similar studies. Suitability of statistical methods - its a bit vague - the reviewer cites the randomization scheme under statistics. Plus this is not necessarily a requirement of many journals. May be better to ask if the reviewer commented on his ability to assess statistical methods and if so, suitability. Constructiveness - As defined by the tool, it seems to be more aimed at the handling editor but I think of constructiveness for the author. Again, there was some yes and some no - the review is polite, detailed, and points out major flaws that need to be addressed. For the Editor though I would prefer a clearer "bottom line" on whether the paper can be saved with acknowledgment of limitations. But I realize that can't be in the part the author sees. I don't think of objectivity as a critical piece, for one thing, or the need for references with the exception of telling the author about missing literature, or statistical methods, etc. |
| 15 | "Acknowledged" could be eliminated, its mention in the manuscript could break the anonymity of the revision. |
| 16 | Should completeness of reporting be explained this way? I am not sure it should be the task of the reviewer to judge the data availability. In section 5 I miss a question about concrete suggestions for improvement of the manuscript. |
| 17 | My biggest issue here is the dichotomization of the items. For me, it is not just about Yes and No, but the degree to which or quality of the item. For example, A reviewer could mention an item but do a terrible job at reviewing/giving feedback on the item. |
| 18 | Study protocol- might be useful to include an ethics section as reviewers are often best placed to judge the nuances of research ethics in a study. Presentation and organisation- it would be good to include a question about grammar and usage, especially for journals that may not have in-house staff to check this. Statistical methods- this is really important but often difficult as reviewers are not in a position to comment on the statistics due to lack of appropriate expertise but also for study types that do not typically have statistical results ie qualitative data. It would be good to have an addendum that states "please consider if the submission contains qualitative data and as such may not contain statistical methods". This sounds obvious, but its the number one complaint from authors who do mixed methods studies or qualitative work. |
| 19 | I like the tool as is - it is thorough yet concise. |
| 20 | The 3 category "yes, no, NA" is ok. Consider a Yes, No, "Could be improved" category or a 5 point ranking. Some of these areas cover a lot of territory. |
| 21 | Peer review should address clear presentation of study goals, clinical questions and objectives. Study funding, ethic approval, protocol registration and COI should be assessed. Peer review should address study hypotheses and should be aligned with quality tools, e.g. reporting quality and risk of bias. Domain 3 should be after domain 4 following recommended reporting of epidemiological studies, e.g., the reporting of the results should be followed by the discussion of study limitations. Peer review should address recommended reporting of statistical results, e.g. minimum data sets, access to shared data, models and variables. Peer review should provide precise recommendations on the improvement of the quality of the manuscripts. The results of the registered completed or terminated studies should be published with objective assessment of the quality of studies. Peer review should ensure the results availability from all conducted studies with objective assessment of study quality. |

| | |
|----|---|
| 22 | The exact meaning and applicability of "... deviations from the study protocol..." are not enough clear."... if opportune, supported by evidence." - the term "evidence" needs clear characterization about its type & content. |
| 23 | The 'helpful' test going along with this is unclear. "the availability of study data and materials?" # The reviewer explicitly comments on their presence or their absence --> YES # The reviewer does not explicitly comment on their presence or their absence --> ??? No or NA ??? the help text doesn't assist me in deciding (but implied to me an 'NA' was the right output) |
| 24 | I think you should include the following additional criteria: Did the review begin with an overall comment on the quality, interest and implications of the study? Did the review divide its comments into Major and Minor? Did the review follow the typical IMRAD structure of the paper? Did the review include clear recommendations for modification rather than vague suggestions? I think it's important to remember that the purpose of peer review is not only to provide an evaluation on the basis of which journal editors make decisions, but also to provide constructive, when possible, feedback to authors so that they can improve their paper, either for resubmission to the same Journal , or for submission to another journal. Although reviewers should be discouraged from making comments about suitability for publication in the Comments to Authors section, reviews should always be sufficiently robust to support editorial 'reject' decisions. |
| 25 | The YES-NO NA is in my opinion unfit for the survey. Many items are PARTLY answered by the reviewer. such a manichean survey does not seem fully covering the problem |
| 26 | Is the manuscript in scope for this journal? |
| 27 | This is a bit vague- the statement should be more specific in relation to scientific knowledge- citing a systematic review and effect size in terms of benefits/harms |
| 28 | Not all studies are quantitative. I would need to adapt ARCADIA if I am to adopt it in my journal, because sometimes we receive qualitative research, as well as theoretical and methodological essays (in the same journal section). I could probably use ARCADIA as-is for reviews in the quality improvement section, but I'm not sure how much adaptation I would need for reviews in sections "clinical review" and "perspective" (commentary, opinion). |
| 29 | Perhaps to add "suspected" plagiarism, duplication, simultaneous submission... |
| 30 | no |
| 31 | I found the choice of yes/no problematic. Particularly for Domains 2-4 would it be better to offer more options: a. Fully, b. Partially, c. Not at all Or a scale 0-3? The use of the tool may also depend on the instructions offered to the peer reviewer on the issues they should address. If they were sent the tool's list of questions to consider it is more likely that they will provide valuable answers. For Domain 4, presentation might need to be split into 2 questions: a. quality of writing (clear, logical and well organised), b. quality of data presentation (including figures, tables and diagrams). For Domain 5, alternative questions: Instead of "clear?", it would be better to ask "were the reviewer's comments relevant and easy to understand? Instead of "constructive?", it would be better to ask "did the reviewer suggest ways to improve the manuscript?" It is difficult to assess objectivity without knowing more about the reviewer and any COI they may have. Perhaps better to ask "Did the reviewer support their comments with appropriate evidence/literature?" How would this tool operate for a non-research article (e.g., a review article)? |
| 32 | 4b The wording is complicated and needs to be simpler. 4d not entirely clear what this means - I assume data means access to the raw data, but what about materials? |

Appendix 12. Codebook of participants' comments (N=32)

| Theme | Definition | Code | Sub-code | Example | N references |
|---------------|--|-------------------------|---------------------------|---|--------------|
| ARCADIA items | Statements on how to improve the ARCADIA items | 1a. Contribution | Wording | This is a bit vague- the statement should be more specific in relation to scientific knowledge- citing a systematic review and effect size in terms of benefits/harms | 1 |
| | | 1b. Relevant literature | Wording | Relevant literature - This question is a bit unclear - it asks if the reviewer commented on whether the literature was "accurately" reviewed, but this isn't the same as "comprehensively reviewed" which is equally important as we want to know they are aware of any other similar studies. | 1 |
| | | 2b. Statistical methods | Wording | Suitability of statistical methods - its a bit vague - the reviewer cites the randomization scheme under statistics. Plus this is not necessarily a requirement of many journals. May be better to ask if the reviewer commented on his ability to assess statistical methods and if so, suitability. | 2 |
| | | 3a. Study conclusions | Double-barrelled question | This "question" 3a has two questions embedded in it: whether the study answers the research question and whether it summarizes study results. These should be two separate questions. | 1 |
| | | 3b. Study limitations | Wording | "Acknowledged" could be eliminated, its mention in the manuscript could break the | 1 |

| | | | | | |
|--|--|-----------------------------------|---------------------------|---|---|
| | | | | anonymity of the revision. | |
| | | 4a. Study protocol | Ethics section | Study protocol- might be useful to include an ethics section as reviewers are often best placed to judge the nuances of research ethics in a study. | 3 |
| | | | Wording | the exact meaning and applicability of "... deviations from the study protocol..." are not enough clear." | |
| | | 4b. Reporting | Double-barrelled question | These should be two separate questions. The same happens in 4b: asking the reviewer to judge (generally) if the study is reproducible is one thing. Asking them to verify reporting guideline adherence is another (huge) task. | 5 |
| | | | Use of acronym | 4b) acronym "RG" used in the online phrasing : unclear | |
| | | | Wording | 4b The wording is complicated and needs to be simpler. | |
| | | 4c. Presentation and organization | Dichotomous answer | For presentation and organization, it is difficult to answer yes/no because a reviewer may make a few superficial comments. Here, a scale of 1 to 5 would be easier to use. | 4 |
| | | | Grammar and usage section | Presentation and organisation- it would be good to include a question about grammar and usage, especially for journals that may not have in-house staff to check this. | |
| | | | Numbered comments | For the editor, but also for the authors, the comments should be numbered. | |
| | | | Double-barrelled question | For Domain 4, presentation might need to be split into 2 questions: a. quality of writing (clear, logical and well organised), b. quality of data | |

| | | | | | |
|--------------|--------------------------------|-----------------------|-----------------------------------|---|---|
| | | | | presentation (including figures, tables and diagrams). | |
| | | 4d. Data availability | Unclear reviewer's responsibility | I am not sure it should be the task of the reviewer to judge the data availability. | 4 |
| | | | Wording | 4d not entirely clear what this means - I assume data means access to the raw data, but what about materials? | |
| | | 5a. Clarity | Dichotomous answer | Clarity, too, is difficult to dichotomize as yes/no. | 2 |
| | | | Wording | For Domain 5, alternative questions: Instead of "clear?", it would be better to ask "were the reviewer's comments relevant and easy to understand?" | |
| | | 5b. Constructiveness | Suggestions for improvement | In order for the comments to be constructive, suggestions for improvement should be provided | 4 |
| | | | Wording | Instead of "constructive?", it would be better to ask "did the reviewer suggest ways to improve the manuscript?" | |
| | | 5c. Objectivity | Relevance of the item | I don't think of objectivity as a critical piece, for one thing, or the need for references with the exception of telling the author about missing literature, or statistical methods, etc. | 3 |
| | | | Wording | if opportune, supported by evidence." - the term "evidence" needs clear characterization about its type & content. | |
| ARCADIA tool | Statements on the ARCADIA tool | Response type | NA | My biggest issue here is the dichotomization of the items. For me, it is not just about Yes and No, but the degree to which or quality of | 8 |

| | | | | | |
|--|--|--|----|--|---|
| | | | | the item. For example, A reviewer could mention an item but do a terrible job at reviewing/giving feedback on the item | |
| | | Applicability of the tool | NA | This only seems relevant for RCT. What about observational studies? | 2 |
| | | Journal suitability | NA | I think the scoring tool need to consider the suitability of the review report in light of the pertinent journal (eg a peer review report for BMJ cannot be the same as one for a tiny circulation subspecialty journal). | 2 |
| | | Related to instructions for peer reviewers | NA | The use of the tool may also depend on the instructions offered to the peer reviewer on the issues they should address. If they were sent the tool's list of questions to consider it is more likely that they will provide valuable answers. | 1 |
| | | Relevance of the tool | NA | such a manichean survey does not seem fully covering the problem | 3 |
| | | Structure of the tool | NA | Domain 3 should be after domain 4 following recommended reporting of epidemiological studies, e.g., the reporting of the results should be followed by the discussion of study limitations. | 1 |
| | | Subjective interpretation | NA | The tool seems useful, although some questions may require creative interpretation when applied to any given review. This is to say that while attempting objectivity in specifying criteria for review, the questionnaire forces subjectivity in interpretation of individual criteria, | 1 |

| | | | | | |
|---------------|--|------------------------------|----|---|---|
| | | | | which kind of defeats the original goal. | |
| | | Too concise items | NA | In general, the items are too concise and comments are not provided for all. | 2 |
| Missing items | Statements on some missing items in the ARCADIA tool | Clear recommendation | NA | Did the review include clear recommendations for modification rather than vague suggestions? | 1 |
| | | General comment on the study | NA | Did the review begin with an overall comment on the quality, interest and implications of the study? | 1 |
| | | Plagiarism | NA | Perhaps to add "suspected" plagiarism, duplication, simultaneous submission... | 1 |
| | | Reviewer expertise | NA | I would suggest that it is important for a reviewer to flag up areas that they have not got expertise to assess - for example this reviewer implied that the statistical methods should be checked by an expert. It is not appropriate to expect peer reviewers to cover all areas. | 1 |
| | | Structure of the review | NA | Did the review divide its comments into Major and Minor? Did the review follow the typical IMRAD structure of the paper? | 1 |
| | | Summary of S&W | NA | A summary of the major strengths and weaknesses of the manuscript should be requested from reviewers | 1 |

| | | | | | |
|---------------------|---|----------------------|----|--|---|
| | | Summary of the paper | NA | I also think that a question that is pertinent is whether the reviewer adequately summarized the paper - this is very helpful to both the author and editor reassuring them that the reviewer took care with the paper. I think its awn essential part of a good review. | 1 |
| Peer review process | Statements on the aims of the peer review process | NA | NA | I think it's important to remember that the purpose of peer review is not only to provide an evaluation on the basis of which journal editors make decisions, but also to provide constructive, when possible, feedback to authors so that they can improve their paper, either for resubmission to the same Journal , or for submission to another journal. Although reviewers should be discouraged from making comments about suitability for publication in the Comments to Authors section, reviews should always be sufficiently robust to support editorial 'reject' decisions. | 2 |
| Unclear comments | Unclear statements | NA | NA | Add more precise comments | 1 |

