Statistical Journal of the IAOS 32 (2016) 11–19
DOI 10.3233/SJI-160967
IOS Press

Helping authors and reviewers ask the right questions: The InfoQ framework for reviewing applied research

Ron S. Kenett[a,*] and Galit Shmueli[b]
[a] *University of Turin, Italy and KPA Group, Raanana, Israel*
[b] *Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan*

Abstract. Reviewers play a critical role in the publication process, the hallmark of scientific advancement. Yet, in many journals, determining the contribution of a paper is left to the reviewer's experience and good sense without providing structured guidelines. This lack of guidance to authors and reviewers increases uncertainty and variability in the usefulness of reviews. We propose an approach, based on the Information Quality (InfoQ) framework, that provides guideline scaffolding for the review process of applied research papers submitted for publication in scientific journals.

Keywords: Information quality, publication, empirical study, data analysis, reviewing guidelines

1. Introduction

Publication of research in academic journals is an important component of developing science as well as a contribution to personal professional development. In the publication process, reviewers play a critical role. They are the main advisors to the gate keepers (the editor) and they also provide feedback to the authors that can be valuable in improving the work. A good reviewer is one who is able to see the contribution of the paper and judge its level relative to the standard expected from the particular journal. Yet, this process is usually carried out in an unstructured way with inherent variability between reviewers and even, within the same reviewer. Quoting Gewin [1]: "many graduate and postdoctoral students, were never taught how to review a manuscript; most peer reviewers learn journals' needs and the reviewer's role only through trial and error. Editors' expectations differ according to their fields, but most agree that simply writing thorough, respectful and helpful reviews is the best way for early-career scientists to find their footing and avoid mistakes."

As a context to this paper we consider here the many statisticians in academia and in industry who regularly review papers for journals without a general framework to help them guide the review process. As in other disciplines, it is typically left to the reviewer's experience and good sense to determine the contribution of a paper. Together with the Associate Editor and Editor's opinions, one assumes that the "wisdom of the review team" will uncover the value of the paper in a reliable and reproducible way. Moreover, in recent years, there has also been a concerted effort by many journals to expedite the reviewing cycle in order to make new knowledge available in a timely fashion. For example, *The Journal of Business and Economic Statistics* instituted the following policy: "The Journal of Business and Economic Statistics has a policy that after the first round of revisions, papers must be either rejected or accepted subject to specific minor revisions."

The requirement to reduce the number of review cycles of a paper creates an even more urgent need to improve the first-round assessment of the contribution of submitted manuscripts.

[*] Corresponding author: Ron S. Kenett, University of Turin, Italy and KPA Group, PO Box 2525, Raanana 43100, Israel. E-mail: ron@kpa-group.com.

The information quality concept (InfoQ) introduced in [2] is a general framework that can be used for reviewing papers and partially fill the gap outlined above. The approach can be used by reviewers as well as by authors themselves as a basis for self-assessment for improving their paper prior to submission. A similar InfoQ assessment has been applied to the review of research proposals, see Chapter 12 in [3]. InfoQ consist of a framework that ties the goal, data, analysis and utility of an empirical study. It can be deconstructed into eight dimensions that are useful for assessment purposes. The approach has been implemented in assessing official statistics reports and in student projects and applied research in the area of risk management, statistical process control, data mining and ecommerce analytics. In this work, we highlight potential benefits of using the InfoQ dimensions and framework in the review process of papers submitted for publication in scientific journals.

InfoQ is defined as "the utility of a particular data set for achieving a given analysis goal by employing statistical analysis or data mining" [2]. InfoQ is affected by the data (X), the data analysis (f) and the analysis goal (g), as well as by the relationships between them. Utility is measured using specific metric(s) (U). By examining each of the components and their relationships, we can learn about the contribution of a given project, study or paper. For example, the contribution can be a new research question, and/or a novel dataset, and/or a new analysis method or approach. Journals typically focus on contributions along one of these three directions. Generally speaking, methodological or theoretical statistics and data mining journals are interested in contributions to f and U, while applied journals (in statistics or other scientific areas) are more interested in novelties in g and X. For illustration, here are the guidelines for authors for *JASA Applications & Case Studies*:

The Applications and Case Studies section publishes original articles that do one or more of the following:

1. For real datasets, present analyses that are statistically innovative as well as scientifically and practically relevant.
2. Contribute substantially to a scientific field through the use of sound statistical methods.
3. Present new and useful data, such as a new life table for a segment of the population or a new social or economical indicator.
4. Using empirical tests, examine or illustrate for an important application the utility of a valuable statistical technique.
5. Evaluate the quality of important data sources.

And for *Science* magazine: "Research Articles should report a major breakthrough in a particular field. They should be in the top 20% of the papers that Science publishes and be of strong interdisciplinary interest or unusual interest to the specialist." In comparison, the guidelines for *JASA Theory & Methods* state: "The Theory and Methods section publishes articles that make original contributions to the foundations, theoretical development, and methodology of statistics and probability."

Because InfoQ is an abstraction, measuring it requires operationalization. A common approach in such cases is to deconstruct the concept into multiple dimensions that are easier to measure. Kenett and Shmueli [2] break down the InfoQ construct into eight dimensions: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Generalizability, Chronology of Data and Goal, Operationalization, and Communication. The eight dimensions can be used for developing streamlined evaluation metrics of InfoQ. In particular [4], describe two studies where InfoQ was integrated into research methods courses, guiding students in evaluating InfoQ of prospective and retrospective studies. For examples using InfoQ in the design of data science and business analytics programs see Chapter 13 in [3]. The results and feedback received so far indicate the importance and usefulness of InfoQ and its eight dimensions for evaluating empirical studies.

The remainder of the paper is organized as follows: We start in Section 2 by listing guidelines for reviewers from several journals in Statistics and Data Mining in order to motivate the need for a more guided paper reviewing process framework. Section 3 is about the application of InfoQ dimensions to the review process. The goal is to help reviewers, associate editors and editors assess the contribution of a paper, its suitability for the journal, and the potential of improving the work sufficiently as to request a revision. Reviewing papers is mostly done by volunteers so that our proposal should be considered a suggestion for those interested in achieving clarity in reviews. Associate editors and editors could consider these suggestions as a structured approach to ensure some homogeneity in the quality the review process. We conclude with a final section discussing the ideas presented in this paper and evidence for the poor quality of current review practices.

Table 1
List of journals published by different societies and their referee guidelines (accessed 7-July-2014)

Journal	Reviewer guidelines URL	Main criteria
Journals of the American Statistical Association:		
	From communications with the editorial staff at ASA, reviewer guidelines are not available online. Providing specific guidelines are left at the discretion of the AE, who would need to modify the generic email sent to reviewers.	
Journals of the American Society for Quality		
	No reviewer guidelines online	
Journals of the Institute of Mathematical Statistics		
Annals of Statistics	www.imstat.org/aos/referee.html	1. Interest and importance and novelty as a scientific contribution. 2. Quality of writing and presentation. 3. Technical correctness.
Annals of Applied Statistics	www.imstat.org/aoas/referee.html	1. Does the paper genuinely concern applied statistics? 2. Is the paper clearly written? 3. Is the paper correct? 4. Is the paper interesting?
Annals of Probability	imstat.org/aop/referee.htm	Is the presentation clear and well organized? Are the results new and interesting? Are the proofs correct and given in adequate detail, or can they be substantially simplified? Do the introduction and abstract give an adequate summary?
Annals of Applied Probability	imstat.org/aap/referee.html	Is the presentation clear and well organized? Are the results new and interesting? Are the proofs correct and given in adequate detail, or can they be substantially simplified? Do the introduction and abstract give an adequate summary?
Statistical Science	www.imstat.org/sts/referee.html	Is the presentation clear and well organized? Are the results new and interesting? Are the proofs correct and given in adequate detail, or can they be substantially simplified? Do the introduction and abstract give an adequate summary? Can paper be streamlined?
Journals of the Royal Statistical Society		

JRSS Series A: Statistics in Society	www.rss.org.uk/article905.asp	(1) suitability for Series A;
		(2) importance;
		(3) interest;
		(4) originality;
		(5) correctness

JRSS Series B: Statistical Methodology	www.rss.org.uk/article906.asp	(1) suitability for Series B;
		(2) importance;
		(3) interest;
		(4) originality;
		(5) correctness

JRSS Series C: Series C: Applied Statistics	www.rss.org.uk/article907.asp	Contents 1. Are the facts, arguments and conclusions in the paper technically valid and accurate? 2. Is previous work adequately referenced and integrated with the new results? 3. Is the paper of practical importance? Structure Exposition
Machine Learning Journals		
Journal of Machine Learning Research	jmlr.org/reviewer-guide.html	Goals: What are research goals and learning task? Description: Is the description adequately detailed for others to replicate the work?... Evaluation: Do the authors evaluate their work in an adequate way (theoretically and/or empirically)? Significance: Does the paper constitute a significant, technically correct contribution to the field that is appropriate for JMLR? Related Work and Discussion: Are strength and limitations and generality of the research adequately discussed . . . Clarity

Table 1, continued

Journal	Reviewer guidelines URL	Main criteria
Machine Learning	Instructions for Authors: www.springer.com/computer/ai/journal/10994	What is the main claim of the paper? Why is this an important contribution to the machine learning literature? What is the evidence you provide to support your claim? What papers by other authors make the most closely related contributions, and how is your paper related to them? Have you published parts of your paper before
Reviewing Guidelines for Conference Papers		
KDD 2012	kdd2012.sigkdd.org/reviewing_criteria.shtml	Novelty: This is arguably the single most important criterion for selecting papers for the conference. Technical Quality: Are the results sound? Potential Impact and Significance: Is this really a significant advance in the state of the art? Clarity of Writing
Scientific Journals/Magazines		
Science Magazine	www.sciencemag.org/site/feature/contribinfo/review.xhtml and www.sciencemag.org/site/feature/contribinfo/RAinstr13.pdf	Technical Rigor: Evaluate whether, or to what extent, the data and methods substantiate the conclusions and interpretations. If appropriate, indicate what additional data and information are needed to do so. Novelty: Indicate in your review if the conclusions are novel or are too similar to work already published.
Nature Journals	www.nature.com/authors/policies/peer_review.html	Who will be interested in reading the paper, and why? What are the main claims of the paper and how significant are they? Is the paper likely to be one of the five most significant papers published in the discipline this year? How does the paper stand out from others in its field? Are the claims novel? If not, which published papers compromise novelty? Are the claims convincing? If not, what further evidence is needed? Are there other experiments or work that would strengthen the paper further? How much would further work improve it, and how difficult would this be? Would it take a long time? Are the claims appropriately discussed in the context of previous literature? If the manuscript is unacceptable, is the study sufficiently promising to encourage the authors to resubmit? If the manuscript is unacceptable but promising, what specific work is needed to make it acceptable?

2. Current guidelines in applied journals

Results from a search of the websites of applied journals published by the leading statistical societies (ASA, IMS and the Royal Statistical Society as well as *Science* and *Nature* magazines) are summarized in Table 1. While journal guidelines range in terms of detail provided to reviewers, the key criteria for judging novelty and importance are typically generally defined (e.g., “Interest and importance and novelty as a scientific contribution”, “Is the paper of practical importance?”). For example, Science requests reviewers to comment on two aspects:

> *Technical Rigor*: Evaluate whether, or to what extent, the data and methods substantiate the conclusions and interpretations. If appropriate, indicate what additional data and information are needed to do so.
> *Novelty*: Indicate in your review if the conclusions are novel or are too similar to work already published.

JRSS-C (Applied Statistics) asks reviewers to consider the following point: “Are the facts, arguments and conclusions in the paper technically valid and accurate?”

These guidelines need to be operationalized. The eight InfoQ dimensions presented in the next section provide a way to do that.

3. InfoQ guidelines for reviewers

We present here the eight InfoQ dimensions and propose a few specific guiding questions for reviewers interested in assessing the level of information quality (InfoQ) of a manuscript submitted for publication which involves data analysis and applied research.

3.1. Data resolution

Data resolution refers to the measurement scale and aggregation level of the data. The measurement scale of the data should be carefully evaluated in terms of its suitability to the stated goal, the analysis methods used, and the required resolution of the research utility. Questions that a reviewer should ask to figure out the strength of this dimension:

- Is the data scale used aligned with the stated goal?
- How reliable and precise are the measuring devices or data sources?
- Is the data analysis suitable for the data aggregation level?

A low rating on data resolution can be indicative of low trust in the usefulness of the study's findings.

As an example, consider Google's ability to predict the prevalence of flu on the basis of the type and extent of internet searches. These predictions match quite well the official figures published by the Centers for Disease Control and Prevention (CDC). The point is that Google's tracking has only a day's delay, compared with the week or more it takes for the CDC to assemble a picture based on reports from doctors' surgeries. Google is faster because it is tracking the outbreak by finding a correlation between what people search for online and whether they have flu symptoms. Another example is provided by the RISCOSS project (www.riscoss.eu) that developed a risk management methodology for adopters of open source software (OSS). The data used to generate risk indicators combines online data of the OSS community such as time to fix bugs and social network analysis of the OSS community and expert opinions. The community data can be collected online with continuous updates. Risk management is based on evaluating risk indicators on a weekly or monthly basis so that the OSS data in RISCOSS is aggregated on a weekly or monthly basis to match the needs of the adopter's risk management activity [5]. Data collected on a minute by minute basis or on a yearly basis would not have the proper resolution.

3.2. Data structure

Data structure relates to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, network data) as well as unstructured, non-numerical data (e.g., text, text with hyperlinks, audio, video, and semantic data). The InfoQ level of a certain data type depends on the goal at hand. Questions that a reviewer should ask to figure out the strength of this dimension:

1. Is the type of the data used aligned with the stated goal?
2. Are data integrity details (corrupted/missing values) described and handled appropriately?
3. Are the analysis methods suitable for the data structure?

A low rating on data structure can be indicative of poor data coverage in terms of the project goals. For example, using a cross-sectional analysis method to analyze a time series warrants special attention when the goal is parameter inference, but is of less concern if the goal is forecasting future values. Another example is removing records with missing data, when missingness might not be random. A paper analyzing on line transactions with the objective to evaluate actual behavior versus declared behavior also needs data on declared behavior through focused queries or questionnaires. Without that, the structure of the data will not provide adequate information quality.

3.3. Data integration

With the variety of data source and data types available today, studies sometimes integrate data from multiple sources and/or types to create new knowledge regarding the goal at hand. Such integration can increase InfoQ, but in other cases it can reduce InfoQ, for example by creating privacy breaches. Questions that a reviewer should ask to figure out the strength of this dimension:

- Are the data integrated from multiple sources? If so, what is the credibility of each source?
- How is the integration done? Are there linkage issues that lead to dropping crucial information?
- Does the data integration add value in terms of the stated goal?
- Does the data integration cause any privacy or confidentiality concerns?

A low rating on data integration can be indicative of missed potential in data analysis. A prime example of data integration is the fusion feature in google. In the RISCOSS methodology, aggregated quantitative data captured from OSS communities is integrated with qualitative expert opinion through an assessment of risk scenarios to derive risk indicators using Bayesian networks [4]. Other examples of data integration include the combination of structured and unstructured semantic data using ETL methods [6] or the calibration of organizational data with official statistics data using copulas [7].

3.4. Temporal relevance

The process of deriving knowledge from data can be put on a time line that includes the data collection, data analysis, and results' usage periods as well as the temporal gaps between these three stages. The different durations and gaps can each affect InfoQ. The data collection duration can increase or decrease InfoQ, depending on the study goal, e.g., studying longitudinal effects vs. a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal. Questions that a reviewer should ask to figure out the strength of this dimension:

- Considering the data collection, data analysis and deployment stages, is any of them time-sensitive?
- Does the time gap between data collection and analysis cause any concern?
- Is the time gap between the data collection and analysis and the intended use of the model (e.g., in terms of policy recommendations) of any concern?

A low rating on temporal relevance can be indicative of an analysis with low relevance to decision makers due to data collected in a different contextual condition. This can happen in economic studies with policy implications that are based on old data.

3.5. Chronology of data and goal

The choice of variables to collect, the temporal relationship between them, and their meaning in the context of the goal at hand affects InfoQ. Questions that a reviewer should ask to figure out the strength of this dimension:

- If the stated goal is predictive, are all the predictor variables expected to be available at the time of prediction?
- If the stated goal is causal, do the causal variables precede the effects?
- In a causal study, are there issues of endogeneity (reverse-causation)?

A low rating on chronology of data and goal can be indicative of low relevance of a specific data analysis due to misaligned timing. A customer satisfaction survey that was designed to be used as input to the annual budget planning cycle becomes irrelevant if its results are communicated after the annual budget is finalized [8]. Reporting of air quality indicators to help allergic patients take proactive actions needs to be reported before the potentially dangerous conditions arise (for such an indicator see the environmental protection agency's air quality index). In another example, in the context of online auctions, classic auction theory dictates that the number of bidders is an important driver of auction price. Models based on this theory are useful for explaining the effect of the number of bidders on price. However, for the purpose of predicting the price of ongoing online auctions, where the number of bidders is unknown until the auction end, the variable "number of bidders", even if available in the data, is useless. Hence, the level of InfoQ contained in number of bidders for models of auction price depends on the goal at hand.

3.6. Generalizability

The utility of $f\ (X|\ g)$ is dependent on the ability to generalize f to the appropriate population. Two types of generalizability are statistical generalizability and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern/model f to other populations, or applying f estimated from one population to predict individual observations in other populations. Determining the level of generalizability requires careful characterization of g. Generalizability is related to the concepts of reproducibility, repeatability and replicability. Reproducibility is the ability to replicate the scientific conclusions and insights, while repeatability is the ability to replicate the exact same numerical results [9]. Replicability (used mainly in biostatistics) refers to replicating results under different conditions, i.e., it is related to scientific generalization. Questions that a reviewer should ask to figure out the strength of this dimension:

- Is the stated goal statistical or scientific generalizability?
- For statistical generalizability in the case of inference, does the paper answer the question "What population does the sample represent?"
- For generalizability in the case of a stated predictive goal (predicting the values of new observations; forecasting future values), are the results generalizable to the to-be-predicted data?
- Does the paper provide sufficient detail for the type of needed reproducibility and/or repeatability, and/or replicability?

Table 2
InfoQ questionnaire for reviewing an empirical research paper or study

Dimension	Questions
1. Data Resolution	1.1 Is the data scale used aligned with the stated goal? 1.2 How reliable and precise are the measuring devices or data sources? 1.3 Is the data analysis suitable for the data aggregation level?
2. Data Structure	2.1 Is the type of the data used aligned with the stated goal? 2.2 Are data integrity details (corrupted/missing values) described and handled appropriately? 2.3 Are the analysis methods suitable for the data structure?
3. Data Integration	3.1 Are the data integrated from multiple sources? If so, what is the credibility of each source? 3.2 How is the integration done? Are there linkage issues that lead to dropping crucial information? 3.3 Does the data integration add value in terms of the stated goal? 3.4 Does the data integration cause any privacy or confidentiality concerns?
4. Temporal Relevance	4.1 Considering the data collection, data analysis and deployment stages, is any of them time-sensitive? 4.2 Does the time gap between data collection and analysis cause any concern? 4.3 Is the time gap between the data collection and analysis and the intended use of the model (e.g., in terms of policy recommendations) of any concern?
5. Chronology of Data & Goal	5.1 If the stated goal is predictive, are all the predictor variables expected to be available at the time of prediction? 5.2 If the stated goal is causal, do the causal variables precede the effects? 5.3 In a causal study, are there issues of endogeneity (reverse-causation)?
6. Generalizability	6.1 Is the stated goal statistical or scientific generalizability? 6.2 For statistical generalizability in the case of inference, does the paper answer the question "What population does the sample represent?" 6.3 For generalizability in the case of a stated predictive goal (predicting the values of new observations; forecasting future values), are the results generalizable to the to-be-predicted data? 6.4 Does the paper provide sufficient detail for the type of needed reproducibility and/or repeatability, and/or replicability?
7. Operationalization	Construct operationalization: 7.1 Are the measured variables themselves of interest to the study goal, or is their underlying construct? 7.2 What are the justifications for the choice of variables? Strength of operationalizing results: 7.3 Who can be affected (positively or negatively) by the research findings? 7.4 What can the affected parties do about it?
8. Communication	8.1 Is the exposition of the goal, data and analysis clear? 8.2 Is the exposition level appropriate for the readership of this journal? 8.3 Are there any confusing details or statements that might lead to confusion or misunderstanding?

A low rating on generalizability reflects a study with relatively low impact. Pearl [10] stated that "Science is about generalization, and generalization requires transportability. Conclusions that are obtained in a laboratory setting are transported and applied elsewhere, in an environment that differs in many aspects from that of the laboratory." Rasch [11,12] used the term specific objectivity to describe that case essential to measurement in which comparisons between individuals become independent of which particular instruments – tests or items or other stimuli – have been used. Symmetrically, it thought to be possible to compare stimuli belonging to the same class – measuring the same thing – independent of which particular individuals, within a class considered, were instrumental for comparison.“ The term general objectivity is reserved for the case in which absolute measures (i.e., amounts) are independent of which instrument (within a class considered) is employed, and no other object is required. By "absolute" we mean the measure "is not dependent on, or without reference to, anything else; not relative". In reviewing a paper, one should assess the contribution of the paper also in terms of its generalization.

3.7. Operationalization

Two types of operationalization are considered: Construct operationalization and action operationalization of the analysis results. Constructs are abstractions that describe a phenomenon of theoretical interest. Measurable data are an operationalization of underlying constructs. The relationship between the underlying construct and its operationalization can vary, and its level relative to the goal is another important aspect of InfoQ. The role of construct operationalization is dependent on the goal, and especially on whether the goal is explanatory, predictive, or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as the data are assumed to measure the construct, the variable is considered adequate. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variable is critical. Action operationalizing results refers to three questions posed by Deming [13]:

- What do you want to accomplish?
- By what method will you accomplish it?
- How will you know when you have accomplished it?

Questions that a reviewer should ask to figure out the strength of construct operationalization:

- Are the measured variables themselves of interest to the study goal, or is their underlying construct?
- What are the justifications for the choice of variables?

Questions that a reviewer should ask to figure out the strength of operationalizing results:

- Who can be affected (positively or negatively) by the research findings?
- What can the affected parties do about it?

A low rating on operationalization indicates that the research might have academic value but, in fact, has no practical impact.

3.8. Communication

Effective communication of the analysis and its utility directly impacts InfoQ. There are plenty of examples where miscommunication of valid results has led to disasters, such as the NASA shuttle Challenger disaster. This is the dimension that typically sees the most detail in reviewer guidelines. Questions that a reviewer should ask to figure out the strength of this dimension:

- Is the exposition of the goal, data and analysis clear?
- Is the exposition level appropriate for the readership of this journal?
- Are there any confusing details or statements that might lead to confusion or misunderstanding?

A low rating on communication can be indicative that poor communication might cover the true value of the analysis and, thereby, dump the value of the information provided by the analysis. For a description of the communication issues in the NASA shuttle program see [14].

4. Discussion

The role of Associate Editors and Editors is to aggregate information from reviewers in order to form an informed opinion regarding a specific submission. An unstructured approach to the review process leads to inconsistencies and high variability in the depth and breadth of reviews. Another aspect of this condition is the range of feedback provided to authors. From very brief and hardly informative reviews to extremely detailed reviews which almost rewrite the paper. This leads to a sense of arbitrariness in the process, which is sometimes justified but sometimes not. In a recent controversial paper, eventually retracted from Nature, one of the commenters wrote: “I feel that it would be great if Nature would find a way to publish the reviewers’ comments on this manuscript as well as the editorial procedure. As long as the reviewers agree, it could be very beneficial. For instance, it might enable the community to see where things went awry in finding the concerns that have been discussed since the publication of these studies. It has been the common practice in EMBO since 2009, (http://bit.ly/1iAVP5i). As it might be problematic to apply this policy retrospectively, this case enhances the need to adopt this model going forward.”

In another recent publication [15], the author analyzes data from an experiment design to assess the effect of variability in the review process. The paper described an experiment where 10% of submitted manuscripts (166 items) submitted for publication in a conference proceedings went through the review process twice. Arbitrariness was measured as the conditional probability for an accepted submission to get rejected if examined by the second committee. This number was equal to 60%, for a total acceptance rate equal to 22.5%. The author applies a Bayesian analysis to these two numbers, by introducing a hidden parameter which measures the probability that a submission meets basic quality criteria. The standard quality criteria considered in this study include novelty, clarity, reproducibility, correctness and no form of misconduct. These were met by a large proportions of submitted items. The Bayesian estimate for the hidden parameter was equal to 56% (95%CI: I = (0.34, 0.83)). As a result of this analysis the author suggests that the total acceptance rate should be increased in order to decrease arbitrariness estimates in future review processes.

In considering official guidelines for review, we found a general lack of clarity in guidelines for reviewers. With the proliferation of journals and communication channels, the good journals can further distinguish themselves by establishing structured guidelines for reviewers and perhaps even training for new reviewers. The general assumption is that experienced authors can instantly become good reviewers is not necessarily valid. Specifically, the *Statistical Journal of the International Association of Official Statistics* (IAOS) states: *The main aim of the journal is to support the IAOS mission by publishing articles to promote the understanding and advancement of official statistics and to foster the development of effective and efficient official statistical services on a global basis*. We propose that InfoQ can provide a structured framework for the review process of applied research papers in order to meet the aims of journals such as the IAOS journal. Table 2 provides a checklist designed to help guide a reviewer of a manuscript on applied research, by pointing out key questions that relate to each of the InfoQ dimensions. We suggest adopting this structure, either formally or informally, in the reviewing process of applied journals.

Acknowledgements

We thank the editor Fritz Scheuren and the six reviewers for their insightful comments and suggestions. In this work, Ron Kenett was partially supported by the RISCOSS project, funded by the EC 7th Framework Programme FP7/2007-2013, agreement number 318249 and Galit Shmueli was supported in part by grant 104-2410-H-007-001-MY2 from the Ministry of Science and Technology in Taiwan.

References

[1] V. Gewin, What the novice peer reviewer needs to know before combing through a submission, *Nature* **478** (2011), 275–277.

[2] R.S. Kenett and G. Shmueli, On information quality. *Journal of the Royal Statistical Society*, Series A (Statistics in Society), **177**(1) (2014), 3–27.

[3] R.S. Kenett and G. Shmueli, *Information quality (InfoQ): The Potential of Data and Analytics to Generate Knowledge*, John Wiley and Sons, 2016.

[4] G. Shmueli and R.S. Kenett, An Information Quality (InfoQ) Framework for Ex-Ante and Ex-Post Evaluation of Empirical Studies, *The 3rd International Workshop on Intelligent Data Analysis and Management (IDAM)*, Kaohsiung, Taiwan, Springer Verlag, 1–13, 2013.

[5] R.S. Kenett, X. Franch, A. Susi and N. Galanis, Adoption of Free Libre Open Source Software (FLOSS): A Risk Management Perspective, *Proc. of the 38th Annual IEEE International Computer Software and Applications Conference (COMPSAC)*, Västerås, Sweden, 2014.

[6] R.S. Kenett and Y. Raanan, *Operational Risk Management: A practical approach to intelligent data analysis* John Wiley and Sons, 2010.

[7] L. DallaValle, Official statistics data integration using vines and non parametric Bayesian belief nets, *Quality Technology and Quantitative Management* **11**(1) (2014), 111–131.

[8] R.S. Kenett and S. Salini, *Modern Analysis of Customer Surveys: With applications using R* (John Wiley and Sons, 2012).

[9] C. Drummond, Replicability is not Reproducibility: Nor is it Good Science, *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, Canada, 2009.

[10] J. Pearl, Transportability across studies: A formal approach, (R-372, UCLA Cognitive Science Laboratory, 2013).

[11] G. Rasch, On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements, *The Danish Yearbook of Philosophy* **14** (1977), 58–93.

[12] G. Rasch, On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symp on Mathematical Statistics and Probability*, **IV** (pp. 321–334) Berkeley: Univ of Chicago Press, 1980.

[13] W.E. Deming, *Quality, Productivity, and Competitive Position*, (Massachusetts Institute of Technology, Center for Advanced Engineering Study, 1982).

[14] R.S. Kenett and P. Thyregod, Aspects of statistical consulting not taught by academia, *Statistica Neerlandica*, special issue on Industrial Statistics, **60**(3) (2006), 396–412.

[15] O. Francois, Arbitrariness of peer review: A Bayesian analysis of the NIPS experiment, http://arxiv.org/abs/1507.06411.