University of Groningen

Numerical Data Quality in Simulation Research

Dong, John Qi

Published in:
Decision Support Systems

DOI:
[10.1016/j.dss.2019.113134](https://doi.org/10.1016/j.dss.2019.113134)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Dong, J. Q. (2019). Numerical Data Quality in Simulation Research: A Reflection and Epistemic Implications. *Decision Support Systems*, 126(1), [113134]. <https://doi.org/10.1016/j.dss.2019.113134>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Numerical data quality in simulation research: A reflection and epistemic implications

John Qi Dong

Faculty of Economics and Business, University of Groningen, 9747 AE Groningen, The Netherlands

ARTICLE INFO

Keywords:

Data quality
Simulation research
Research methodology

ABSTRACT

Marsden and Pingry (2018) discuss different aspects of numerical data that are used in empirical research, with an aim to improve data quality and facilitate research replication in the IS field. However, the implications for numerical data from simulation research also need to be discussed. This article tries to join the discussion by using Marsden and Pingry's (2018) *What, When, Where, How, Who, Which, and Why* aspects to discuss how we can create the necessary, but not sufficient, conditions for data accuracy, validity and reproducibility in simulation research.

1. Introduction

Marsden and Pingry [1] recently point to the issue of data quality in IS research and discuss their implications for replication of empirical findings and accumulation of scientific knowledge. They list seven data types that are used in empirical research — interviews, surveys, field experiments, quasi-experiments, controlled laboratory experiments, empirical observed data, and third-party fee-for-service data — and consider how each of these data types fares on *What, When, Where, How, Who, Which, and Why* aspects. These seven aspects describe what is captured in the data, the time at which the data is collected, the virtual or real location of the data collection, the precise process of data collection, the individuals or agents involved in the data collection, the instruments or artifacts used in collecting the data, and the reasons or goals for collecting the data ([1], p. A2). They provide useful thresholds for IS scholars to improve data accuracy, validity and producibility in empirical research. While common data types for empirical research have all been discussed, numerical data from simulation research need to be further considered. The aim of this article is, therefore, to join the conversation on the data quality issue in IS research with a particular focus on numerical data from simulation research.

Simulation is a useful research method that has been increasingly used by IS scholars (see Beese et al. [2] and Xie [3] for literature reviews). It is a scientific and rigorous method that is not only useful for social science research but also commonly applied in natural science fields, such as physics, chemistry, biology, and pharmacy [4]. Simulation research takes advantage of both analytical and statistical techniques by combining computational modeling and quantitative analysis [5,6]. In this article, I follow *What, When, Where, How, Who, Which, and*

Why aspects from Marsden and Pingry [1] to discuss the necessary, but not sufficient, conditions under which numerical data used in simulation research could be accurate and valid, and more importantly, reproducible. I believe that this is an important complement to Marsden and Pingry [1], because the meaning of some aspects is very different for numerical data from simulation research. I will start my discussion with briefing the importance of simulation method in IS research, and then discuss the data quality issue in simulation research.

2. The importance of simulation method in information systems research

Simulation research is defined as “a method of using computer software to model the operation of real-world processes, systems, or events” ([5], p. 481). This method allows IS scholars to isolate and vary the key parameters of computational models in controlled experiments, while producing massive amount of numerical data for statistical analysis [2]. Notable simulation research efforts have been carried out to advance IS theory (e.g., [7–12]). For example, Xie [3] conduct a survey on simulation research published in the AIS basket of eight journals and *Information and Management*, revealing 143 publications. More recently, Beese et al. [2] search simulation IS literature extensively and find 697 publications from 23 major IS journals and conferences, with an increasing number of publications over years. Many different simulation techniques have been used in IS research, including agent-based modeling, stochastic processes, system dynamics, artificial neural networks, and genetic algorithms, among others. Various IS topics are studied by using simulation techniques, such as e-commerce, IS management, artificial intelligence, knowledge management, IS use, IS security, social

E-mail address: john.dong@rug.nl.

<https://doi.org/10.1016/j.dss.2019.113134>

Received 15 February 2019; Received in revised form 16 May 2019; Accepted 16 August 2019

Available online 16 August 2019

0167-9236/ © 2019 Elsevier B.V. All rights reserved.

networks, online communities, decision support systems, enterprise systems, IS value, IS planning and implementation, IS outsourcing, and software development [2,3].

3. Numerical data quality in simulation research

Next, I discuss how to assess the data quality in the simulation research setting, based on Marsden and Pingry's [1] framework entailing *What*, *When*, *Where*, *How*, *Who*, *Which*, and *Why* aspects. Again, it is important to note that as a result of the difference between simulation research and empirical research, the meaning of some aspects is very different for numerical data from simulation research.

What: Like all other data types, simulation research needs to clarify what exactly is captured in the data. Though other data types in empirical research are gathered from the real world, numerical data are generated by computer algorithms in simulation research. Because of this unique data gathering practice, simulation research needs to provide a more detailed description of what is simulated in what way and the numerical data for what parameters are generated in simulation experiments.

When: With regard to the time at which the data are collected, it has different meaning for numerical data from simulation research. Since simulation researchers need to generate numerical data from computer algorithms based on a computational model, it is often irrelevant to know at what time the computer algorithms generate the data. Instead, simulation research has the advantage to simulate dynamic interactions over time (e.g., decision making based on the inputs from different types of agents in a number of periods) and thanks to the advance of computers, recent simulation research can typically generate a large volume of numerical data up to thousands of periods. Such longitudinal feature is rarely possible for collecting data in empirical research. While the pattern of simulation results can become rather stable over time periods, it is not always the case. Therefore, the choice of time periods in simulation design could matter for what numerical data are generated and therefore needs to be carefully decided and explained.

Where: Marsden and Pingry [1] refer the *Where* question to the virtual or real location of the data collection. In simulation research, numerical data are always generated in a virtual environment based on computer algorithms. Such virtual environments are often generalizable across different programming languages, software systems, and computers. However, changes of software and hardware may sometimes affect the data generating process. While simulation researchers are often very brief about their virtual environments, it could be useful for replication efforts by providing all details about what programming language is used to develop the algorithms, what software is used to run the algorithms and the version number of the software, as well as what computer specifications are used to generate the data.

How: The precise details of computer algorithms are key for knowing how numerical data are generated in simulation research. The computational model often has a complex structure for specifying the decision making and interacting processes of simulated agents. Sometimes the data generated and the simulation results could be different only if the sequence of decision and interaction rules are exchanged in the same model. Therefore, it is important that simulation researchers carefully explain how numerical data are generated *step by step*, rather than simply describing the model in use. A flowchart, pseudo code or full code in an appendix can considerably facilitate replication of simulation data and results.

Who: Different from empirical research, simulation research does not collect data from individuals or organizations in the real world. Having said that, agents are typically modeled in simulation to represent individuals or organizations in the real world. Simulation researchers should clearly specify and justify who the agents are in

the model and how they are connected to real-world individuals or organizations. This requires detailed specifications and logical explanations about the decision and interaction rules on which agents behave in the model. Otherwise, it is unclear what agents are actually modeled and if they are modeled appropriately. It is not uncommon that the decision and interaction rules in simulation are specified based on prior empirical findings, which provide support for justifying the modeling choices on the one hand and restrict what can be modeled on the other.¹ When extent empirical findings are used to set up the model, simulation researchers need to carefully check the quality of empirical data from which the findings were derived. Only valid findings based on empirical data meeting the thresholds of Marsden and Pingry [1] can provide a sound foundation for simulation.

Which: While simulation research does not often apply empirical instruments, it does employ artifacts in generating the data. Similar to empirical research, how primary parameters capture and map to the reality being modeled need to be well explained and justified. To increase the accuracy of simulation models, techniques that can be used by IS scholars include 1) building on rationalism and existing models, 2) sensitivity analysis by running models with varying parameters and extreme values in particular, 3) replication of simulation results multiple times, 4) displaying patterns of simulation results graphically over time, and 5) comparing simulation results with empirical data and findings [2,3]. When any of these techniques are used in simulation research, they need to be clarified. For example, if existing simulation models are used, an informative description is needed about how the existing models were developed and why the chosen models are appropriate and suitable. In particular, it is not uncommon that IS scholars gather empirical data to check the validity of parameters and outcomes used in their simulation research [2,3]. If any empirical data are used in simulation research, the data quality needs to meet the thresholds of Marsden and Pingry [1].

Why: Marsden and Pingry [1] suggest that the *Why* question is less a concern in empirical research, because high quality empirical data gathered for one purpose may turn out to have other unanticipated but valuable purposes. In simulation research, however, this situation rarely happens. Simulation researchers formulate specific computational models to generate numerical data for specific research purposes. The generated numerical data are often not used for other purposes. Thus, simulation researchers are expected to not only explain why the numerical data are generated in certain ways, but also explain why the use of simulation method is appropriate to conduct the research. Simulation method can be used when empirical research is not possible or suffers from crucial limitations, and where research purposes are to understand complex decisions and interactions unfold over time [5]. If the choice of simulation method is appropriate, simulation researchers need to further explain why they model the reality in certain ways (i.e., the accuracy of model) and why they choose specific experimental design to generate numerical data. Given that controlled experiments are commonly used to generate data in simulation research, Marsden and Pingry's [1] discussion on data from controlled laboratory experiments provides useful guidance for justifying experimental design in simulation research.

4. Concluding remarks

Replication of simulation results has long been suggested to be important for examining the internal and external validity of research findings [5]. To do that, information about *What*, *When*, *Where*, *How*, *Who*, *Which*, and *Why* numerical data are generated in simulation

¹ I thank one anonymous reviewer who suggested this point.

research is necessary. Though it is not sufficient to guarantee the validity of findings if all these aspects are covered in the explanations of simulation design and the reporting of simulation results, replication efforts could be largely facilitated and the assessment of data quality is possible. I hope this article provides some useful thoughts on the data quality issue in simulation research, as a meaningful addition to the useful conversation initiated by Marsden and Pingry [1]. As I believe, this is an important discussion to promote quality IS research that may be empirical, simulation-based, or both.

References

- [1] J.R. Marsden, D.E. Pingry, Numerical data quality in IS research and the implications for replication, *Decision Support Systems* 115 (1) (2018) A1–A7.
- [2] J. Beese, M.K. Haki, S. Aier, R. Winter, Simulation-based research in information systems, *Business and Information Systems Engineering* 61 (4) (2019) 503–521.
- [3] Y. Xie, A survey of simulation research in information systems discipline, *Proceedings of Australasian Conference on Information Systems*, 2017, pp. 1–12.
- [4] C.S. Taber, R.J. Timpone, *Computational Modeling*, Sage, Thousand Oaks, CA, 1996.
- [5] J.P. Davis, K.M. Eisenhardt, C.B. Bingham, Developing theory through simulation methods, *Academy of Management Review* 32 (2) (2007) 480–499.
- [6] J.R. Harrison, Z. Lin, G.R. Carroll, K.M. Carley, Simulation modeling in organization and management research, *Academy of Management Review* 32 (4) (2007) 1229–1245.
- [7] T.K. Abdel-Hamid, The economics of software quality assurance: a simulation-based case study, *MIS Quarterly* 12 (3) (1988) 395–411.
- [8] R.M. Chang, W. Oh, A. Pinsonneault, D. Kwon, A network perspective of digital competition in online advertising industries: a simulation-based approach, *Information Systems Research* 21 (3) (2010) 571–593.
- [9] G.C. Kane, M. Alavi, Information technology and organizational learning: an investigation of exploration and exploitation processes, *Organization Science* (2007) 749–883.
- [10] N. Nan, Capturing bottom-up information technology use processes: a complex adaptive systems model, *MIS Quarterly* 35 (2) (2011) 505–532.
- [11] W. Oh, J.Y. Moon, J. Hahn, T. Kim, Leader influence on sustained participation in online collaborative work communities: a simulation-based approach, *Information Systems Research* 27 (2) (2016) 383–402.
- [12] T.S. Raghu, B. Jayaraman, H.R. Rao, Toward an integration of agent- and activity-centric approaches in organizational process modeling: incorporating incentive mechanisms, *Information Systems Research* 15 (4) (2004) 316–335.

John Qi Dong is an Associate Professor of Strategy and Organization at Faculty of Economics and Business, University of Groningen in the Netherlands. He holds a PhD degree in information systems from Hong Kong University of Science and Technology. He also received a master degree and a bachelor degree, both in management, from Renmin University of China. His research interests include digital innovation, digital transformation, innovation management, and strategic management. His work has been published or forthcoming in *MIS Quarterly*, *Journal of Management*, *Journal of the Association for Information Systems*, *Journal of Product Innovation Management*, *Journal of Strategic Information Systems*, *European Journal of Information Systems*, *Decision Support Systems*, *Information and Management*, *Long Range Planning*, *Technological Forecasting and Social Change*, *Journal of Business Research*, and *Drug Discovery Today*, among others. He currently serves as the senior editor for *Information Technology and People*, the associate editor for *Information and Management*, and the editorial board member for *Journal of Strategic Information Systems*. He was the associate editor for *Decision Support Systems* special issue on omnichannel business.