CHALMERS
UNIVERSITY OF TECHNOLOGY

A Second Look at the Impact of Passive Voice Requirements on Domain Modeling: Bayesian Reanalysis of an Experiment

Downloaded from: https://research.chalmers.se, 2025-11-10 04:01 UTC

Citation for the original published paper (version of record):
Frattini, J., Fucci, D., Torkar, R. et al (2024). A Second Look at the Impact of Passive Voice Requirements on Domain Modeling: Bayesian
Reanalysis of an Experiment. PROCEEDINGS OF THE 2024 IEEE/ACM INTERNATIONAL WORKSHOP ON METHODOLOGICAL ISSUES WITH EMPIRICAL STUDIES IN SOFTWARE ENGINEERING, WSESE 2024: 27-33. http://dx.doi.org/10.1145/3643664.3648211

N.B. When citing this work, cite the original published paper.

© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, or reuse of any copyrighted component of this work in other works.

(article starts on next page)

A Second Look at the Impact of Passive Voice Requirements on Domain Modeling: Bayesian Reanalysis of an Experiment

Julian Frattini
Davide Fucci
{firstname}.{lastname}@bth.se
Blekinge Institute of Technology
Karlskrona, Sweden

Richard Torkar
richard.torkar@gu.se
Chalmers and University of Gothenburg
Göteborg, Sweden
Stellenbosch Institute for Advanced Study (STIAS)
Stellenbosch, South Africa

Daniel Mendez
daniel.mendez@bth.se
Blekinge Institute of Technology
Karlskrona, Sweden
fortiss GmbH
Munich, Germany

ABSTRACT

The quality of requirements specifications may impact subsequent, dependent software engineering (SE) activities. However, empirical evidence of this impact remains scarce and too often superficial as studies abstract from the phenomena under investigation too much. Two of these abstractions are caused by the lack of frameworks for causal inference and frequentist methods which reduce complex data to binary results. In this study, we aim to demonstrate (1) the use of a causal framework and (2) contrast frequentist methods with more sophisticated Bayesian statistics for causal inference. To this end, we reanalyze the only known controlled experiment investigating the impact of passive voice on the subsequent activity of domain modeling. We follow a framework for statistical causal inference and employ Bayesian data analysis methods to re-investigate the hypotheses of the original study. Our results reveal that the effects observed by the original authors turned out to be much less significant than previously assumed. This study supports the recent call to action in SE research to adopt Bayesian data analysis, including causal frameworks and Bayesian statistics, for more sophisticated causal inference.

CCS CONCEPTS

• **Software and its engineering → Requirements analysis**; • **Mathematics of computing → Bayesian computation**.

KEYWORDS

Requirements Engineering, Requirements Quality, Controlled experiment, Bayesian Data Analysis

ACM Reference Format:
Julian Frattini, Davide Fucci, Richard Torkar, and Daniel Mendez. 2024. A Second Look at the Impact of Passive Voice Requirements on Domain Modeling: Bayesian Reanalysis of an Experiment. In *International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE '24), April 16, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3643664.3648211

This work licensed under Creative Commons Attribution International 4.0 License.

WSESE '24, April 16, 2024, Lisbon, Portugal
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0567-0/24/04
https://doi.org/10.1145/3643664.3648211

1 INTRODUCTION

Requirements specifications serve as input to several subsequent software engineering (SE) activities [36]. Consequently, the quality of requirements specifications impacts the performance of these dependent activities [26]. For example, ambiguous or incomplete requirements specifications may result in incorrect or missing features when implementing the requirements. Because the cost for remediating these defects scales the longer they remain in the development process [4], organizations are interested in detecting and removing requirements quality defects as soon as possible.

The requirements quality research domain aims to meet this need [29]. However, while requirements quality research abounds with normative rules about requirements quality [15], it lacks empirical evidence that supports the relevance of these rules [14, 29]. Moreover, the few studies contributing empirical evidence are often confounded, too abstract, and their inference reduces complex, context-sensitive data to binary results, for example, through the use of frequentist methods [24]. The insufficient quantity and quality of evidence impede the adoption of requirements quality research in practice [13].

With this study, we aim to demonstrate how more sophisticated inference methods than frequentist approaches derive deeper insights from an empirical study and may even revise frequentist claims. This paper makes the following contributions:

(1) A recovery of the analysis of one of the only controlled experiments on requirements quality known to us [12].
(2) A reanalysis of the hypothesis of this experiment using more sophisticated statistical methods.

Data Availability

We disclose all supplementary material, including the data, figures, and analysis scripts, in our replication package.[1]

2 RELATED WORK

2.1 Requirements Quality

Requirements quality research is a sub-domain within requirements engineering (RE) research dedicated to the assessment and improvement of requirements artifacts and processes [29]. Given the importance of RE to the software development life cycle, the quality of its artifacts and processes plays a major role in project success

[1] https://zenodo.org/doi/10.5281/zenodo.10283010

or failure [26, 36]. For requirements artifacts like (systematic) requirements specifications, use cases, user stories, and others [27], a popular concept to identify quality defects is the *requirements quality factor* [15]. A requirements quality factor is a normative metric that maps a requirements artifact onto some level of quality based on defined criteria [15]. One commonly researched requirements quality factor is *passive voice* [12, 22], which associates the use of passive voice in a natural language (NL) requirements sentence with bad quality since it potentially omits the semantic agent of the sentence [12]. For example, the requirements specification "If the settings *are changed*, ..." obscures the agent of the requirement. An active formulation of this specification, "If an administrator *changes* the settings, ..." makes the agent explicit.

Recent research has identified a major shortcoming of requirements quality factors, namely their relevance [14]. The requirements quality research domain abounds with publications proposing new quality factors and tools to detect violations against them but lacks empirical evidence for the implied causal relationship, i.e., that the violation causes an actual impact on subsequent SE activities [1]. A previous literature survey has revealed that among 57 primary studies proposing requirements quality factors, only 40 discuss their impact at all, and of these, only 11 provide some sort of empirical evidence [14]. Without empirical evidence of the impact of a requirements quality factor on subsequent activities, these factors do not reliably identify requirements quality defects that matter. Practitioners rightfully harbor skepticism toward requirements quality research given this lack of evidence which impedes research adoption in practice [10, 13, 31].

For example, while several sources advise against the use of passive voice as described above [11, 20, 22, 32] only two publications known to the authors investigate its actual impact on subsequent activities. Krisch et al. conducted a document study in which domain experts classified active and passive requirements sentences as either problematic or unproblematic [23]. The results indicate that passive voice is generally unproblematic as adjacent text often compensates for the information omitted due to the passive voice. Femmer et al. conducted a controlled experiment with university students to assess how passive voice in requirements sentences impacts the domain modeling activity [12]. The authors conclude that passive voice requirements increase the number of missing associations with statistical significance but not the number of missing actors or domain objects.

2.2 Inferential Statistics

Most statistical methods applied in SE beyond descriptive statistics are limited to frequentist inferential statistics. These usually take the form of null hypothesis significance testing (NHST), which stratifies the distribution of a dependent response variable by one or more independent variables and compares their mean. We assume that the popularity of these methods stems from the established guidelines [39], the availability of tools to perform them, and their acceptance in the community.

However, frequentist methods like NHST have several shortcomings. From a research design perspective, they overemphasize the variables involved in an alleged, causal relationship without a systematic approach for addressing confounders [30]. From a data analysis perspective, common issues like the multiple-hypothesis problem [3] and the unscientific practice of fishing for significant test results below an arbitrary significance level [2] are well-known, yet still occur in practice [28]. Moreover, NHST reduces complex, context-sensitive data down to binary answers (i.e., whether there is a significant difference in the distributions' mean or not), which leads to superficial and overly abstracted research results that are void of any uncertainty that the data originally encoded [17].

The recent rise of Bayesian data analysis (BDA) aims to mitigate these shortcomings [24, 25] by (1) embedding inferential statistics in causal reasoning frameworks [30, 33] and (2) applying Bayesian statistics, i.e., encoding the uncertainty of the impact that independent variables have on dependent variables in probability distributions [25]. Prior to any data analysis, involved variables and their causal relationship are made explicit. During the data analysis, explicit prior assumptions are updated in light of the observed data using Bayes' Theorem. As a result, BDA produces uncertainty-preserving statistical inferences with explicit causal assumptions. Recently, SE researchers have started to advocate for the adoption of BDA methods [17, 18, 34] but they still remain to be niche [33].

3 METHOD

In this study, we aim to demonstrate how frameworks for causal inference and Bayesian statistics provide more sophisticated insights which reduce issues of drawing inappropriate conclusions from empirical studies. To this end, we reanalyzed the data of a previous controlled experiment using BDA. Section 3.1.1 presents the design of the original experiment and Section 3.1.2 elaborates on the issues with the experiment. Section 3.2 then presents the reanalysis performed in the scope of this study.

3.1 Original Experiment

The original experiment by Femmer et al. aims to understand the impact of passive voice in requirements on domain modeling [12] by asking the following research questions:

- RQ1.1: Is the use of passive sentences in requirements harmful for finding actors?
- RQ1.2: Is the use of passive sentences in requirements harmful for identifying domain objects?
- RQ1.3: Is the use of passive sentences in requirements harmful for identifying associations?

3.1.1 Design. The experimental task was to create a domain model based on a single-sentence NL requirements specification. The domain model consisted of the following three types of elements: *actors*, which represent human participants in the requirement, *domain objects*, which represent any non-human entities in the requirement, and *associations*, which connect elements that have a relationship according to the requirement. Figure 1 visualizes a domain model for the requirements specification "The system shall be capable of returning the search results latest 30 seconds after the user has entered the search criteria." [12]

The authors of the original study conducted a controlled experiment with independent measures, i.e., every participant is assigned to only one treatment [39]. The authors recruited $n_p = 15$ participants for the experiment. The participants consisted of two Bachelor students, eight Master students, four Ph.D. students, and

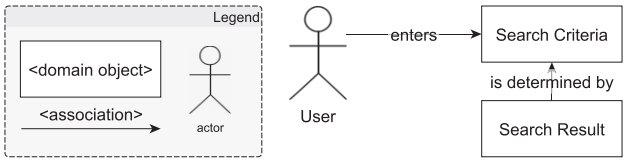

Figure 1: Domain model example

Table 1: Results of the original study [12]. P-values indicating a statistically significant difference with $\alpha = 0.05$ are prefixed with an asterisk (*)

Element	Mean (A)	Mean (P)	Median (A)	Median (P)	P-value	Conf. Int.	Cliff's δ
Actors	0.43	1.00	0	1	0.10	(0; ∞)	0.39
Objects	1.29	2.00	1	1	0.25	(-1; ∞)	0.25
Associations	4.14	7.88	3	8	*0.02	(1; ∞)	0.75

one student with an unknown background. In addition to the participants' study program, the authors also recorded their age group as well as their industrial and academic experience in SE, RE, and programming on an ordinal scale.

To enable independent measures, seven participants were assigned to the control group (A) and eight to the treatment group (P). The control group received the requirements specifications in active formulation. The treatment group received semantically similar requirements specifications in passive formulation. For example, the authors transformed the aforementioned active requirements sentence to the following passive formulation for the treatment group: "The search results *shall be returned* no later 30 seconds after the user has entered the search criteria." [12].

After assessing the general SE and RE knowledge in a quiz, the participants conducted the experimental task. Every participant received $n_r = 7$ requirements specifications, such that the experiment produced $n_p \times n_r = 105$ observations. The authors then compared the 105 domain models with the sample solution and counted the number of missing actors, domain objects, and associations. To evaluate the hypotheses implied by the research questions, the authors summed up these numbers for all seven requirements sentences of each participant. Each participant was associated with a total number of missed actors, domain objects, and associations throughout all seven requirements. Then, the authors calculated the mean and median number of missing elements for the control and treatment groups and conducted a Mann-Whitney test with a 95% confidence interval to determine whether there was a statistically significant difference between the two groups.

Table 1 shows the results of the original study [12]. With a significance level of $\alpha = 0.05$, the NHST rejects only the null hypothesis implied by RQ1.3 ($p = 0.02 < \alpha$). The authors conclude that the use of passive voice does not have a statistically significant impact on the number of actors and domain objects missing from resulting domain models, but it does have an impact on the number of missing associations.

3.1.2 Issues. The original experiment by Femmer et al. [12] suffers from at least the following issues.

Issues with reproduction. The authors originally disclosed their experiment data at http://goo.gl/WlTPE5, which was forwarded to https://www.in.tum.de/i04/~femmer/data/passives_experiment.zip. However, this link does no longer resolve given that institutional websites commonly discontinue hosting resources of members that change their affiliation [19, 38]. Thankfully, the authors of the original paper were able to recover the lost replication package [16] and archived it via Zenodo.[2] Still, the replication package contains only the study protocol and obtained data, but not the script to reproduce the evaluation. The lack of reproducibility impedes our goal of comparing methods of statistical inference.

Issues with drawing appropriate conclusions. The employed research design and analysis risks drawing inappropriate conclusions in two regards. Firstly, the significance test investigates the isolated impact of passive voice on the three dependent variables. Possible confounders, like the experience of participants, were recorded but not considered in the evaluation. Secondly, frequentist NHSTs reduce the data to single, binary results, omitting any uncertainty [17] and comparing point estimates, which are unreasonably precise.

Issues selecting an appropriate study design. The selected experimental design introduced one more potential confounder. Because the authors of the original study used an *independent measures* design [39] the evaluation does not account for between-subject variance [35]. In other words: the evaluation does not consider that the observed differences in the dependent variables are caused by the treatment or by other factors like the individual skill of each participant.

3.2 Reanalysis

We address the first of the three issues by reproducing the original evaluation and disclosing it for future replication. For this, we extracted the experimental results from the original study and performed the evaluation according to the information in the manuscript [12]. The reproduced evaluation script is contained in our replication package.

To address the second and third issue, we reanalyze the data generated by the experiment using an established framework for causal inference and Bayesian instead of frequentist methods. The framework allows us to (1) revise and extend the causal assumptions of the original experiment and (2) consider potential confounders in the analysis, while the use of BDA allows us to (3) generate more sophisticated inferences that preserve the uncertainty of the causal influences.

We employ the framework for statistical causal inference that was developed by Siebert [33]. This framework is based on Pearl's original model of causal inference [30] and consists of the three major steps modeling, identification, and estimation. The following paragraphs briefly summarize each of these steps and are further elaborated in our replication package. For a gentler introduction to frameworks for statistical causal inference, we refer the interested reader to appropriate literature [30, 33]. For a gentler introduction to

[2]Now available at https://zenodo.org/records/7499290

BDA, we refer the interested reader to appropriate textbooks [25] or descriptive demonstrations of the application of BDA in SE research [9, 17, 18, 34].

3.2.1 Modeling. In the first step, we make our causal assumptions of the phenomenon under investigation explicit [33]. These causal assumptions are specified in a directed acyclic graph (DAG), in which nodes represent variables and directed edges between them represent assumed causal effects of one variable on another [8]. In our reanalysis, the eligible variables are limited to the variables collected during the original experiment [12].

3.2.2 Identification. In the second step, we select all variables that form the so-called adjustment set. Four causal criteria inform this selection and prevent variable bias like colliders or backdoors [25], mitigating that non-causal correlations do not influence the causal relation of interest. The selection of the adjustment set mitigates the second issue mentioned in Section 3.1.2.

3.2.3 Estimation. In the third and final step, we derive a regression model from the adjustment set of eligible variables. We first select an appropriate probability distribution type to represent each of the three response variables based on the maximum entropy criterion [21] and ontological assumptions. All three variables are whole numbers bounded by the number of expected actors, domain objects, and associations. Consequently, we model all response variables with Binomial distributions.

We model the parameter p—which defines the shape of the Binomial distribution—in dependency of all eligible independent variables, called the predictors. Each predictor is multiplied with a coefficient that represents the strength and direction of the influence that the predictor has on the response variable. To begin, we assign an uninformative prior distribution to each of these coefficients, i.e., a normal distribution centered around $\mu = 0$ with a standard deviation of $\sigma = 1$. This represents our prior belief of the causal relationship between the predictors and response variables, which are yet unknown. We confirm the appropriateness of the selected priors via prior predictive checks [37].

The predictors of each response variable consist of the independent variables selected during the identification step. Further, we include the following variables as predictors:

- Intercept: The global average of missing any element of the domain model. This represents the general challenge of creating a domain model from an NL requirements specification, independent of any predictor values.
- Participant-specific intercept: The participant-specific average of missing any element of the domain model. This represents the general skill of a participant.
- Requirement-specific intercept: The requirement-specific average of missing any element of the domain model. This represents the general complexity of a requirement.

While involving a global intercept is a general best practice [25], the two group-specific intercepts retain local variance in the model [9]. The resulting hierarchical model makes use of partial pooling, which is understood to outperform purely global or local models [9, 25]. The inclusion of a participant-specific intercept mitigates the third issue mentioned in Section 3.1.2, as it represents between-subject variance in the statistical evaluation.

Table 2: Results of the strict reproduction

Element	Mean (A)	Mean (P)	Median (A)	Median (P)	P-value	Conf. Int.	Cliff's δ
Actors	0.43	1.00	0	1	0.19	(0; 1)	0.38
Objects	1.29	2.00	1	1	0.50	(-1; 3)	0.22
Associations	4.14	7.88	3	8	*0.03	(1; 7)	0.68

Given the selected probability distribution and predictors, we train one Bayesian model for each of the three response variables with the experimental data gathered during the original experiment [12]. We conduct this step using the `brms` library [6] in *R*. During the training process, Hamiltonian Monte Carlo Markov Chains update the prior distributions of the predictor coefficients to better reflect the impact of the predictors in light of the observed data [5]. This produces the posterior distributions of the predictor coefficients, which then represent the updated belief of the model about the strength and direction of the influence with which a predictor impacts a response variable. The standard deviation of each coefficient reflects the uncertainty of the impact of its associated predictor. This further mitigates the second issue mentioned in Section 3.1.2 by retaining the uncertainty of each impact.

We confirm that the model was trained appropriately by inspecting the Markov Chains [25] and by performing posterior predictive checks [37]. Finally, we evaluate the trained models by plotting the marginal effects of relevant predictors, mainly the use of passive voice. The marginal plots show the distribution of the response variable for all levels of the selected predictor while keeping all other predictors at representative levels. The resulting mean predictions and confidence intervals visualize the difference that the chosen predictor has on the response variable. This visualization represents the isolated effect of that predictor on the outcome.

4 RESULTS

4.1 Reproduction of the original evaluation

Table 2 shows the strict reproduction of the experimental results using the same frequentist methods as the original study [12]. The mean and median values match exactly. The calculated p-values differ (0.10 vs. 0.19, 0.25 vs. 0.50, 0.02 vs. 0.03), but using the same significance level $\alpha = 0.05$ would result in the same hypotheses being rejected (i.e., only the hypothesis implied by RQ1.3). Similarly, the effect size calculated via Cliff's δ matches with a margin of 0.07. Only one extreme end of every confidence interval could not be reproduced. We assume this to be due to incorrect calculation or reporting in the original study.

4.2 Reanalysis of the data using BDA

Figure 2 visualizes the DAG that makes the causal assumptions of the phenomenon under investigation explicit. The DAG is populated with all variables recorded during the original experiment [12] and connected with all causal relationships that we assume based on our prior knowledge. The causal relationships between the main factor (red node) and the three dependent response variables (turquoise

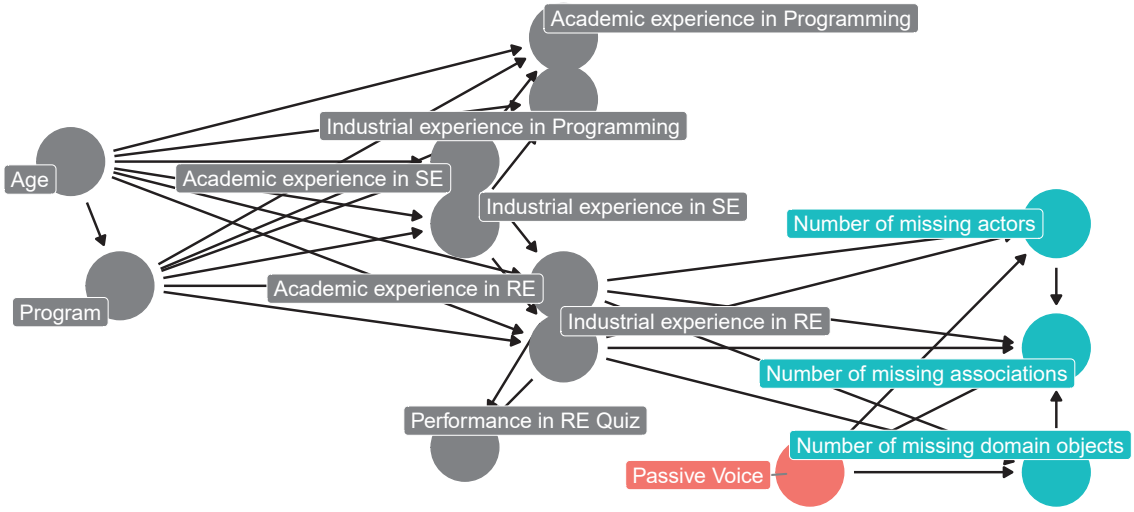

Figure 2: Full DAG visualizing the causal assumptions (red: exposure/main factor, turquoise: response/dependent variables)

nodes) were already assumed in the original study [12] and are the main relationships of interest. We assume additional relationships, for example:

- Age → Program: The older a participant, the more likely it is that they have progressed further in their studies.
- Program → Academic experience in RE: The more advanced the study program, the higher the academic experience that a student has collected in RE.
- Academic/industrial experience in RE → number of missing actors/domain objects/associations: The higher the experience in RE, the fewer mistakes a student makes during domain modeling.
- Number of missing actors/domain objects → Number of missing associations: Missing an actor or domain object leads to missing an association, as one of the two nodes connected through an expected association is unavailable.

All other causal assumptions and their justification can be found in our replication package. Figure 3 visualizes the reduced DAG resulting from the identification step. This DAG contains only variables included in the adjustment set, i.e., all variables relevant for the causal analysis. The causal effect of all excluded variables passes through these remaining variables. Hence, they suffice to model the causal influence on the response variables.

Figure 4 visualizes the marginal effects of the main factor (passive voice) on the three response variables. All plots show that the use of passive voice slightly raises the mean of the response variable distribution, i.e., the use of passive voice increases the likelihood of missing more actors, domain objects, and associations. However, the confidence intervals of the main factor overlap in all three cases, meaning that this difference is not significant. The chance that the use of passive voice results in equal or even fewer missing actors, domain objects, and even associations remains.

Figure 5 shows the marginal effects of the number of missing actors and missing domain objects on the likelihood of missing an

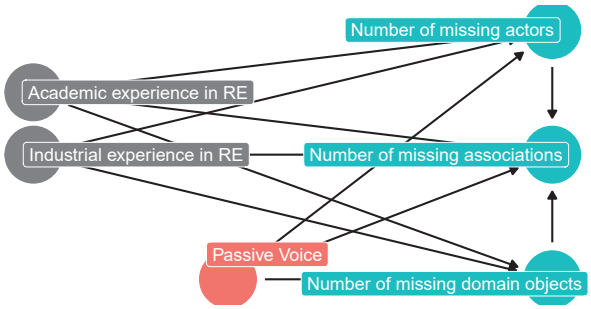

Figure 3: Reduced DAG including all variables eligible for the regression model

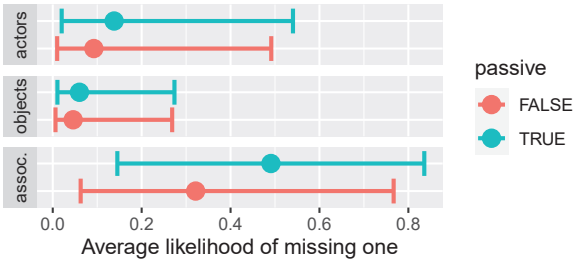

Figure 4: Isolated impact of passive voice on the likelihood of missing an actor, object, or association ("assoc.")

association. The plot shows that missing an actor or domain model increases the likelihood of missing an association, which confirms the causal assumption represented in our DAG. The average and confidence interval for the number of missing actors (red in Figure 5) is only defined for 0 and 1 because the experiment data did not

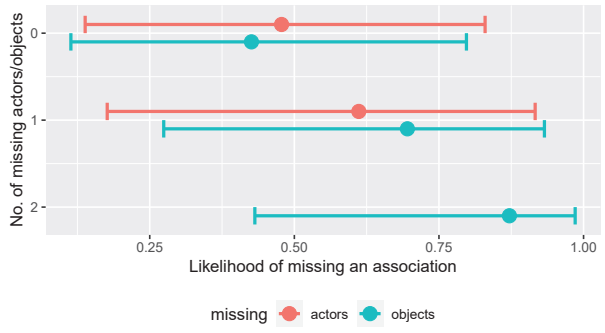

Figure 5: Impact of the number of missing actors and objects on the likelihood of missing an association

contain any observation with more than one missing actor per domain model.

5 DISCUSSION

Finally, we discuss the implications of the results in Section 5.1 and address remaining threats to validity in Section 5.2.

5.1 Implications

Issues of reproduction can be overcome as long as the authors of the original work preserve their replication package. This encounter supports the observation by Gabelica et al. [19] and Winter et al. [38] that replication packages hosted on institutional websites are prone to become inaccessible over time. We strongly advise hosting replication packages via services that committed to a long-term retention policy, like Zenodo[3] or figshare.[4]

More importantly, the reanalysis presented in this study shows that the lack of a framework for causal inference as well as frequentist methods may cause issues with drawing appropriate conclusions. The results of the reanalysis revealed that the use of passive voice does not have a significant impact on the number of missing associations in resulting domain models as claimed in the original study [12]. Instead, the use of a framework for causal inference showed that this impact is confounded by the number of missing actors and domain objects, which also do not experience a significant impact by the main factor of interest. Additionally, the use of Bayesian statistics highlighted that the remaining difference in the response variables is uncertain and not significantly different.

These insights imply two recommendations for future research. For research design, the use of an explicit framework for causal inference provides a systematic approach for dealing with potential confounders [18, 30]. For data analysis, the use of Bayesian statistics retains uncertainty and allows transparent inferences from empirical data [17, 25, 34].

5.2 Threats to validity

The reanalysis continues to suffer from threats to validity. We discuss these according to the classification by Cook et al. [7].

Construct validity. The construct validity suffers from *inadequate preoperational explication of constructs* for all variables concerning experience [7]. In the experiment, industrial and academic experience in RE—two of the predictors with an impact on the three response variables—are measured on an ordinal scale with four levels: no experience, up to 6 months, 6 to 12 months, and more than 12 months [12]. Whether these variables adequately reflect experience remains questionable.

Internal validity. The internal validity suffers from potential *confounders*. The reanalysis could only involve the variables recorded during the original study and was, therefore, constrained to the variables listed in Figure 2. Other variables with a potential causal impact on the response variables—like domain knowledge or prior training in domain modeling—were not available. The internal validity further suffers from an unknown *interaction with selection* due to the design of the experiment. Given the independent measures design, each participant was exposed to only one treatment [35, 39]. This produced the risk of an interaction effect between the participant and the treatment, i.e., participants of one group could excel with their respective treatment for unknown reasons.

External validity. The external validity suffers from an *interaction of selection and treatment*, i.e., the experiment participants are potentially not a representative sample of the intended target population. The study only involved university students of different programs. Hence, there is no evidence that the conclusions are generalizable to SE practitioners.

6 CONCLUSION

This study reanalyses the only controlled experiment investigating the impact of passive voice in requirements specifications [12] by employing a framework for statistical causal inference [33] and using Bayesian in contrast to frequentist data analysis methods [17]. We could show that the results of the original study are much less significant than suggested by the frequentist analysis and that passive voice has, in consequence, a much smaller impact in the studied context than the original study had assumed.

Needless to say, our aim is not to criticize the original study [12] itself. In fact, we would like to acknowledge the authors' contributions to the requirements quality research domain, especially as controlled experiments were, and still are, rare in this domain [14]. Instead, our intention is to critically reflect upon frequentist analysis that still constitutes the prevalent choice in the empirical software engineering community with little to no attention to its limitations.

Our reanalysis continues to suffer from several threats to validity. For example, the experimental design made it impossible to identify whether some participants performed particularly well or badly given their assignment to the control or treatment group. Using a crossover design in which all treatments are applied to all subjects could mitigate this threat [35].

One hope that we associate with our study is to raise awareness of the shortcomings of frequentist analyses, especially when applied as a universal tool. We especially hope that our short demonstration, as well as our replication package, will caution fellow SE researchers to use out-of-the-box frequentist approaches and,

[3] https://zenodo.org/
[4] https://figshare.com/

instead, encourage them to consider Bayesian data analysis approaches [25], which include (1) proper frameworks for statistical causal inference [30, 33] and (2) Bayesian statistics [17, 18]. These approaches ensure that experimental designs are informed by explicit causal assumptions, and their execution produces more sophisticated inferences preserving uncertainty, in turn enriching scientific contributions to be more reflected and insightful.

ACKNOWLEDGMENTS

This work was supported by the KKS foundation through the S.E.R.T. Research Profile project at Blekinge Institute of Technology. We particularly thank Henning Femmer, representing the authors of the original study, for his support and the recovery of the data, which made this reanalysis possible in the first place.

REFERENCES

[1] Muneera Bano. 2015. Addressing the challenges of requirements ambiguity: A review of empirical literature. In *2015 IEEE Fifth International Workshop on Empirical Requirements Engineering (EmpiRE)*. IEEE, 21–24.
[2] JC Barnes and Shannon J Linning. 2021. Statistical Power, P-Values, and the Positive Predictive Value. *The Encyclopedia of Research Methods in Criminology and Criminal Justice* 1 (2021), 337–343.
[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.
[4] Barry W Boehm and Philip N. Papaccio. 1988. Understanding and controlling software costs. *IEEE transactions on software engineering* 14, 10 (1988), 1462–1477.
[5] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. 2011. *Handbook of markov chain monte carlo*. CRC press.
[6] Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80 (2017), 1–28.
[7] Thomas D Cook, Donald Thomas Campbell, and Arles Day. 1979. *Quasi-experimentation: Design & analysis issues for field settings*. Vol. 351. Houghton Mifflin Boston.
[8] Felix Elwert. 2013. Graphical causal models. In *Handbook of causal analysis for social research*. Springer, 245–273.
[9] Neil A Ernst. 2018. Bayesian hierarchical modelling for tailoring metric thresholds. In *Proceedings of the 15th international conference on mining software repositories*. 587–591.
[10] Henning Femmer. 2018. Requirements Quality Defect Detection with the Qualicen Requirements Scout.. In *REFSQ Workshops*.
[11] Henning Femmer, Daniel Méndez Fernández, Stefan Wagner, and Sebastian Eder. 2017. Rapid quality assurance with requirements smells. *Journal of Systems and Software* 123 (2017), 190–213.
[12] Henning Femmer, Jan Kučera, and Antonio Vetrò. 2014. On the impact of passive voice requirements on domain modelling. In *Proceedings of the 8th ACM/IEEE international symposium on empirical software engineering and measurement*. 1–4.
[13] Xavier Franch, Daniel Mendez, Andreas Vogelsang, Rogardt Heldal, Eric Knauss, Marc Oriol, Guilherme Travassos, Jeffrey Clark Carver, and Thomas Zimmermann. 2020. How do Practitioners Perceive the Relevance of Requirements Engineering Research? *IEEE Transactions on Software Engineering* (2020).
[14] Julian Frattini, Lloyd Montgomery, Jannik Fischbach, Daniel Mendez, Davide Fucci, and Michael Unterkalmsteiner. 2023. Requirements Quality Research: a harmonized Theory, Evaluation, and Roadmap. *Requirements engineering* (2023).
[15] Julian Frattini, Lloyd Montgomery, Jannik Fischbach, Michael Unterkalmsteiner, Daniel Mendez, and Davide Fucci. 2022. A live extensible ontology of quality factors for textual requirements. In *2022 IEEE 30th International Requirements Engineering Conference (RE)*. IEEE, 274–280.
[16] Julian Frattini, Lloyd Montgomery, Davide Fucci, Jannik Fischbach, Michael Unterkalmsteiner, and Daniel Mendez. 2023. Let's Stop Building at the Feet of Giants: Recovering unavailable Requirements Quality Artifacts. *arXiv preprint arXiv:2304.04670* (2023).
[17] Carlo A Furia, Robert Feldt, and Richard Torkar. 2019. Bayesian data analysis in empirical software engineering research. *IEEE Transactions on Software Engineering* 47, 9 (2019), 1786–1810.
[18] Carlo A Furia, Richard Torkar, and Robert Feldt. 2022. Applying Bayesian analysis guidelines to empirical software engineering data: The case of programming languages and code quality. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 3 (2022), 1–38.
[19] Mirko Gabelica, Ružica Bojčić, and Livia Puljak. 2022. Many researchers were not compliant with their published data sharing statement: mixed-methods study. *Journal of Clinical Epidemiology* (2022).
[20] Gonzalo Génova, José M Fuentes, Juan Llorens, Omar Hurtado, and Valentín Moreno. 2013. A framework to measure and improve the quality of textual requirements. *Requirements engineering* 18 (2013), 25–41.
[21] E. T. Jaynes. 2003. *Probability theory: The logic of science*. Cambridge University Press, Cambridge.
[22] Leonid Kof. 2007. Treatment of passive voice and conjunctions in use case documents. In *Natural Language Processing and Information Systems: 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France, June 27-29, 2007. Proceedings 12*. Springer, 181–192.
[23] Jennifer Krisch and Frank Houdek. 2015. The myth of bad passive voice and weak words an empirical investigation in the automotive industry. In *2015 IEEE 23rd International Requirements Engineering Conference (RE)*. IEEE, 344–351.
[24] J Jack Lee. 2011. Demystify statistical significance—time to move on from the p value to Bayesian analysis. , 2–3 pages.
[25] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.
[26] Daniel Méndez, Stefan Wagner, Marcos Kalinowski, Michael Felderer, Priscilla Mafra, Antonio Vetrò, Tayana Conte, M-T Christiansson, Des Greer, Casper Lassenius, et al. 2017. Naming the pain in requirements engineering: Contemporary problems, causes, and effects in practice. *Empirical software engineering* 22 (2017), 2298–2338.
[27] Daniel Méndez Fernández and Birgit Penzenstadler. 2015. Artefact-based requirements engineering: the AMDiRE approach. *Requirements Engineering* 20 (2015), 405–434.
[28] Tim Menzies and Martin Shepperd. 2019. "Bad smells" in software analytics papers. *Information and software technology* 112 (2019), 35–47.
[29] Lloyd Montgomery, Davide Fucci, Abir Bouraffa, Lisa Scholz, and Walid Maalej. 2022. Empirical research on requirements quality: a systematic mapping study. *Requirements Engineering* 27, 2 (2022), 183–209.
[30] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
[31] Keith Thomas Phalp, Jonathan Vincent, and Karl Cox. 2007. Assessing the quality of use case descriptions. *Software Quality Journal* 15, 1 (2007), 69–97.
[32] Klaus Pohl. 2016. *Requirements engineering fundamentals: a study guide for the certified professional for requirements engineering exam-foundation level-IREB compliant*. Rocky Nook, Inc.
[33] Julien Siebert. 2023. Applications of statistical causal inference in software engineering. *Information and Software Technology* (2023), 107198.
[34] Richard Torkar, Robert Feldt, and Carlo A Furia. 2020. Bayesian data analysis in empirical software engineering: The case of missing data. *Contemporary Empirical Methods in Software Engineering* (2020), 289–324.
[35] Sira Vegas, Cecilia Apa, and Natalia Juristo. 2015. Crossover designs in software engineering experiments: Benefits and perils. *IEEE Transactions on Software Engineering* 42, 2 (2015), 120–135.
[36] Stefan Wagner, Daniel Méndez Fernández, Michael Felderer, Antonio Vetrò, Marcos Kalinowski, Roel Wieringa, Dietmar Pfahl, Tayana Conte, Marie-Therese Christiansson, Desmond Greer, et al. 2019. Status quo in requirements engineering: A theory and a global family of surveys. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 28, 2 (2019), 1–48.
[37] Jeff S Wesner and Justin PF Pomeranz. 2021. Choosing priors in Bayesian ecological models by simulating from the prior predictive distribution. *Ecosphere* 12, 9 (2021), e03739.
[38] Stefan Winter, Christopher S Timperley, Ben Hermann, Jürgen Cito, Jonathan Bell, Michael Hilton, and Dirk Beyer. 2022. A retrospective study of one decade of artifact evaluations. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 145–156.
[39] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer Science & Business Media.