Reflecting on the Use of the Policy-Process-Product Theory in Empirical Software Engineering

Kelechi G. Kalu
Purdue University, IN, USA
kalu@purdue.edu

Taylor R. Schorlemmer
Purdue University, IN, USA
tschorle@purdue.edu

Sophie Chen
University of Michigan, MI, USA
sophie.cy.chen@gmail.com

Kyle Robinson
Purdue University, IN, USA
robin489@purdue.edu

Erik Kocinare
Purdue University, IN, USA
ekocinar@purdue.edu

James C. Davis
Purdue University, IN, USA
davisjam@purdue.edu

ABSTRACT

The primary theory of software engineering is that an organization's Policies and Processes influence the quality of its Products. We call this the *PPP Theory*. Although empirical software engineering research has grown common, it is unclear whether researchers are trying to evaluate the PPP Theory. To assess this, we analyzed half (33) of the empirical works published over the last two years in three prominent software engineering conferences. In this sample, 70% focus on policies/processes or products, not both. Only 33% provided measurements relating policy/process and products. We make four recommendations: (1) Use PPP Theory in study design; (2) Study feedback relationships; (3) Diversify the studied feedforward relationships; and (4) Disentangle policy and process. Let us remember that research results are in the context of, and with respect to, the relationship between software products, processes, and policies.

CCS CONCEPTS

• **General and reference** → **Empirical studies**; • **Software and its engineering**;

KEYWORDS

Empirical Software Engineering, Software Process and Policy

ACM Reference Format:
Kelechi G. Kalu, Taylor R. Schorlemmer, Sophie Chen, Kyle Robinson, Erik Kocinare, and James C. Davis. 2023. Reflecting on the Use of the Policy-Process-Product Theory in Empirical Software Engineering. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '23), December 3–9, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3611643.3613075

1 INTRODUCTION

Empirical software engineering research analyzes data to improve software products and engineering processes [1, 2]. International standards organizations [3], industry consortia [4], and professional organizations [5] all assert that the *Policies* and *Processes* of software engineering influence the quality of the software *Product* (the *PPP Theory*). Various studies support some of the relationships predicted by the PPP Theory [6, 7]. Nevertheless, it remains unclear which policies and processes are most effective in achieving high-quality products, and how these vary by context [8]. To address this, experts have recommended that empirical software engineering researchers incorporate the PPP Theory, either as contextual information in case studies or as part of a controlled experiment [9, 10]. Doing so could resolve widely remarked-upon challenges with the generalizability and replicability of empirical software engineering research results [7, 9, 11–16]. However, the extent to which the research community has taken this advice is unclear.

This reflection paper examines whether empirical software engineering researchers are considering the relationship between policies, processes, and software products. To achieve this, we reviewed empirical software research works published in 3 software engineering venues (ICSE, ESEC/FSE, and ASE) in 2021 and 2022. We identified the primary aspects of the PPP Theory considered by each work, and the extent to which the PPP Theory was incorporated into the work. We report that empirical studies consider a subset of the PPP Theory and are usually focused on individual theoretical concepts rather than the relationships of the theory. We challenge the Empirical Software Engineering research community to consciously consider the PPP Theory in their study designs.

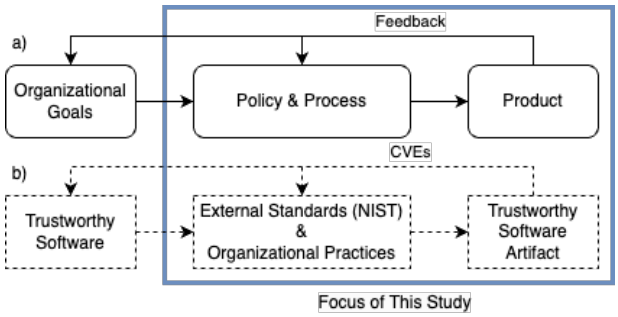

Figure 1: (a) Policy-Process-Product (PPP) Theory. Organizational goals influence the policies and processes adopted by software engineers. Policies and process influence product development. Feedback may modify policies, processes, or the original goals. We treat the (often overlapped) concepts of Policy and Process as a single entity. (b) Example of the PPP Theory for the goal of producing trustworthy software.

ESEC/FSE '23, December 3–9, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0327-0/23/12.
https://doi.org/10.1145/3611643.3613075

2 BACKGROUND: THE PPP THEORY

2.1 Theoretical Constructs

Policy: Policy has many meanings, including processes, artifacts, discourses, and bodies of knowledge about a field [17–19]. In the software engineering literature, policy means both organizational strategies [20–22], and technical system behaviors [23–25]. For PPP Theory, we define **policy** as *an official statement of an organization's software engineering practices, derived from the organization's goals.*

Process: A process consists of the steps followed to accomplish a task, *e.g.,* performing code review or implementing a new feature [6, 20]. For PPP Theory, we define **process** as *the methods used by software engineers to accomplish their tasks.*

In the software engineering literature, we found that *process* and *policy* typically have overlapping definitions. We lump them together into a single **process/policy** construct as shown in Figure 1.

Product: A software product is a set of software and associated documentation, designed and developed to meet a specific set of user needs [6, 26, 27]. For PPP Theory, we define a **product** as *the artifacts produced by a software engineering process.* What comprises a product is context-dependent; some teams produce libraries, others web services, others mobile applications, and so on.

2.2 Policy-Process-Product Relationship

Figure 1 shows the PPP Theory: these constructs and the relationships between them. Organizational goals are iteratively refined into policies, processes, and finally products. This theory is propounded by documents from international standards organizations [3], industry consortia [4], professional organizations [5], governments [20], and the academic literature [6, 7, 21–23, 28, 29].

The PPP Theory predicts bi-directional relationships between the three constructs. A software team's policy informs how its processes are defined, and a team's process influences the quality of the product. In the reverse direction, retrospectives and postmortems provide feedback to modify processes and policies.

An example of the PPP Theory is demonstrated in Figure 1(b). An organization has the goal of securing its artifact's supply chain [30]. Organizational leaders create a policy: "Follow NIST standards" [31]. Engineering teams comply through several process elements, such as code review (for code vulnerability inspection) and using provenance certification tools (*e.g.,* Sigstore [32]). The desired product quality, a secure supply chain, is assessed: defects (*e.g.,* CVEs) provide feedback to improve the process.

Some seminal works explore the relationships between the PPP theory constructs [33–36]. For example, Humphrey *et al.* [33] and Wohlin *et al.* [34] demonstrated the impact of the Personal Software Process (PSP) on the software product (forward direction). In a follow-up study, Wohlin *et al.* showed software defects can be utilized in the Feedback direction to improve the PSP [35].

3 QUESTION AND METHODS

We ask: *To what extent does the PPP Theory inform modern empirical software engineering research?* To answer this question, we assessed 33 papers from top software engineering research venues. This section describes the selection of those papers, the initial assessment approach used in our pilot study, and our revised assessment approach. Our final methodology is summarized in Figure 2.

3.1 Paper selection

We gathered recent empirical software engineering papers (2021-2022) from all tracks of three prominent conferences (ICSE, ESEC/FSE, and ASE). We retrieved full-length papers, totaling 65, that included the term "empirical" in their title or keywords. Initially, we used the DBLP database for the title match and later cross-verified our findings and expanded our search using the ACM digital library, considering both the title and author's keywords. For analysis, we randomly selected 50% of the collected papers.

3.2 Analysis process

Our goal was to assess the presence of PPP relationships in our selected papers. We iteratively refined an analysis instrument through a pilot study. Ultimately, we assessed two distinct aspects of each work: its *construct focus* and its *relationship prevalence.*

3.2.1 Pilot study. In our pilot study, we established a complex classification scheme to rate the appearance of PPP Theory in empirical research papers. This includes a set of rating metrics and a method for scoring each paper on those metrics.

Metrics: Our initial metrics attempted to measure the occurrence of process-product and policy-product relationships in each paper. Each of these section-relationship combinations was evaluated on a four-point scale: (1) *Silent* (no mention of relationships), (2) *Implicit* (acknowledges relationship without discussion of impact), (3) *Descriptive* (describes extensively the relationship between process/policy/product), and (4) *Experimental* (describes and controls for these relationships in their experiment).

Analysis Process: In our initial approach, raters focused on the Methodology, Results/Discussion, and Threats to Validity sections — we thought any use of PPP Theory would be documented here. After reading through a paper, raters classified the section-relationship combinations according to our four-point scale.

Refinements: We used this approach on 13 papers in our pilot study (20% of the available data). We identified two flaws. (1) Raters struggled to differentiate between levels of our four-point scale. Inter-rater disagreements were common and hard to resolve. (2) Some PPP Theory elements were missed because the paper sections targeted in our analysis were too specific.

3.2.2 Final Analysis Approach. First, we clarified definitions to make categories easier to differentiate. Second, we characterized papers holistically rather than considering individual sections. Lastly, given the relatively rare use of PPP Theory relationships in the pilot papers, we reduced the scope of the measurement to simply reporting whether process/product relationships were considered at all, or actively controlled for in the papers. Figure 2 provides an overview of the final analysis approach we used to assess each paper.

Metrics: We assessed the use of PPP Theory with two metrics:

(1) *Construct Focus*: Which PPP Theory construct(s) did the paper focus on?

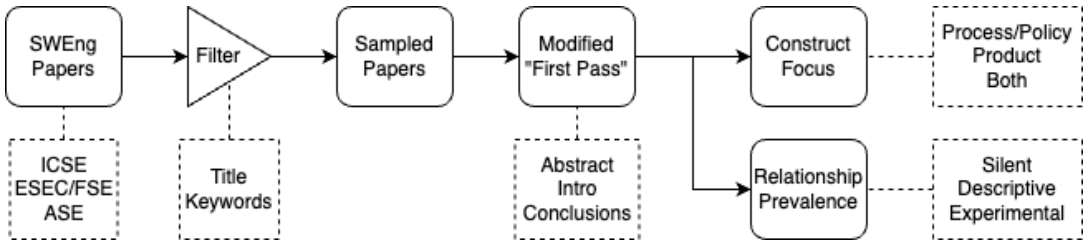

Figure 2: Our streamlined paper analysis approach. We begin by searching for full-length papers in leading software engineering conferences, filtering based on titles and keywords, resulting in empirical studies. We randomly sample half of these studies. This is followed by a "First Pass" analysis akin to Keshav [37] to comprehend paper content. We identify the paper's PPP Theory construct(s) focus and assess the presence of PPP Theory relationships.

(2) *Relationship Prevalence:* Did the paper identify relationships between PPP Theory constructs?

These metrics allow us to categorize what a study is *about* and whether it *considers* PPP Theory.

For the construct focus metric (item 1 above), we categorize each paper into one of the following three types:

(1) *Process/Policy:* The paper observes or measures process/policy. For example, He *et al.* measures the library migration process in the Java ecosystem [38].
(2) *Product:* The paper observes or measures a product. For example, Shen *et al.* study root causes and symptoms of deep learning compiler bugs [39].
(3) *Both:* The paper considers both. For example, Di Grazia *et al.* measure the adoption and use of Python type annotations (process) *and* the resulting statically-detectable type errors in Python projects (product) [40].

For the relationship prevalence metric (item 2 above), we categorize each paper into one of the following three types:

(1) *Silent:* The paper makes no mention of PPP Theory relationships. For example, Shen *et al.* [39] report the characteristics of deep learning compiler bugs but do not explicitly describe how software engineering processes or policies can cause these bugs or should be influenced by bugs.
(2) *Descriptive:* The paper *mentions* a relationship between PPP Theory constructs. For example, He *et al.* [38] measure the library migration process and describe the importance of this process with respect to Java software products, but do not directly measure this relationship.
(3) *Experimental:* The paper *measures* a relationship between PPP Theory constructs. For example, Di Grazia *et al.* [40] measures the relationship between using type annotations and the resulting number of type errors.

Also, we categorized each paper based on the ownership of the empirical data employed in the study: *Public* (papers involving publicly accessible data), *Private* (data not accessible/proprietary), and *Both* (papers incorporating both private and public data).

Analysis Process: Our raters followed a modified version of Keshav's "First Pass" to quickly assess the PPP-Theoretic content of selected papers [37]. Raters proceeded as follows:

- Read title, abstract, introduction, and research questions.
- Read section headings.
- Read the findings — this includes highlighted key results, the discussion, and the Conclusion section.

After performing this "First Pass," raters categorized the construct focus and relationship prevalence metrics for the paper according to the descriptions mentioned above.

Each paper was evaluated by two raters. Inter-rater agreement was measured using Cohen's Kappa score [41], and all disagreements were settled through discussion. Prior to settling disagreements, our process produced a Kappa score of 0.86 for the construct classifications, 0.67 for rating the PPP relationships in each work, and 0.88 for the Data accessibility.

4 RESULTS

In this section, we identify our results from assessing 33 empirical software engineering papers. Figure 3 summarizes our findings.

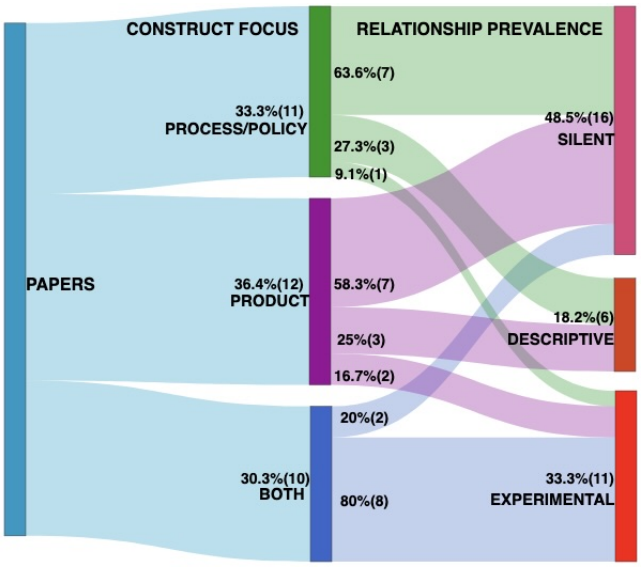

Figure 3: Distribution of analyzed papers based on Construct Focus (process/policy, products, both) and Relationship Prevalence (silent, descriptive, experimental). 30% of papers consider multiple constructs and 49% of papers are silent about relationships between them.

4.1 Construct Focus

The first stage of Figure 3 divides papers by their Construct focus. Papers typically consider products (37%) or policy/processes (33%). The former observes the actions of software engineers and organizations, and the latter measures information about software artifacts. **The smallest set considered both policy/process and products** (30%).

4.2 Relationship Prevalence

The second stage of Figure 3 divides papers by their relationship prevalence. **Work experimenting with or measuring PPP Theory relationships was rare**: 33% or 11 of the papers. About half of the papers (16 papers or 49%) did not discuss relationships from the PPP Theory, and the remaining 6 papers (18%) were descriptive. As expected, considering the construct focus of a paper, the studies focused on a single construct tend to be silent or descriptive, and studies that consider both constructs tend to relate them.

4.3 Other observations

We considered the ownership of the data used in these works. Most papers used public data (64%), some private (24%), rarely both (12%). We observed one trend: of the four studies that used both data types, three experimentally showed a PPP relationship. Perhaps studies with diverse data provide more insight into PPP relationships.

5 IMPLICATIONS FOR RESEARCH

We suggest four implications for the research community.

(1) Incorporate PPP Theory into study designs: A surprising fraction of papers (49%) did not consider these PPP relationships. We do not wish to criticize these works; there is value in characterizing processes and in characterizing products, whether or not relationships are demonstrated between these constructs. However, we wonder if the Empirical Software Engineering research community would benefit from a greater focus on the PPP-Theoretic basis for their measurements. This was the original vision of empirical software engineering introduced in the 1980s and 1990s [10, 42, 43]. This could provide a meaningful way to address concerns about the interpretability and generalizability of our community's empirical research [9, 13, 15, 16]. As a step towards this, the SIGSOFT Empirical Standards [44] could be extended to provide guidance about epistemology (*What is software engineering knowledge?* [45]), not just about methodology (*How to obtain knowledge?*). Some thoughtfulness about the PPP Theory could help authors analyze the Threats to the Validity of their work, without resorting to vague statements about generalizability in "other contexts".

(2) Study the feedback relationship: Among the papers that did consider a relationship between policy/process and product, we note that there was a bias toward measuring the "forward" direction of Figure 1. In our sample it is unusual for researchers to characterize and measure the role of feedback in the engineering process. Although many works have observed the opportunity for failures to inform future engineering approaches [28, 29], this appears to still be a gap in the literature.

(3) Study more feed-forward relationships: Although the "forward" direction of relationships was more commonly examined, there are classes of constructs whose relationships were not examined in our sample. Papers in this category considered topics like software organizational structure, software evolution, and software maintenance. We did not see any papers on topics such as the effect of regulations (*e.g.,* GDPR), cybersecurity policies (*e.g.,* NIST 8397), or industry standards (*e.g.,* MISRA). Greater industry collaboration might facilitate the study of such relationships. Empirical software engineering research often examines open-source software — those engineers lack the liability that motivates organizations to promote such policies and processes.

(4) Disentangle Policy and Process: Lastly, the PPP Theory predicts separate roles of policy and process. Policy interfaces with organizational goals, while process interfaces with the engineered product. These constructs are generally entangled in the empirical software engineering literature, so in our model, we combined them in Figure 2. Considering them as separate constructs may help the community develop a richer theory of software engineering.

6 THREATS TO VALIDITY

Construct: We rely on constructs and relationships defined by the PPP Theory. Our specific operationalizations were derived from literature (§2), but distinguishing these can be difficult because they are often entangled. We addressed this concern through inter-rater agreement, achieving reasonable Kappa scores of 0.86 for PPP constructs and 0.67 for relationships between constructs.

Internal: We make no claims of cause and effect.

External: We sampled half of the empirical works from ICSE, ESEC/FSE, and ASE 2021-2022. A longer time span or alternative venues might affect our results. Based on our understanding of the recent research literature, we do not think time is a crucial variable. These three venues are large general venues, so considering other venues seems unlikely to substantially shift our result.

7 CONCLUSION

In this paper, we have investigated the degree to which current empirical software engineering works consider the relationship between policies, processes, and software products (PPP relationship). We have reviewed 33 published works. Our results show that: (1) Most empirical software engineering works are focussed on single constructs of the PPP theory model, and (2) 18% of the reviewed works provided a description for the PPP relationships, while 49% of the works were silent on this PPP relationship.

Consequent to our results, we have made 4 suggestions to the software engineering research community on the significance of adopting the PPP Theory in future empirical studies. Our recommendations center on the need to further study the PPP theory constructs, and their forward and feedback relations.

DATA AVAILABILITY

Data is available on Zenodo: https://doi.org/10.5281/zenodo.8277429.

ACKNOWLEDGMENTS

We thank A. Kazerouni and the reviewers for their feedback. We acknowledge financial support from NSF awards IIS-1813935, SaTC-2135156, and POSE-2229703, as well as Cisco and Rolls Royce.

REFERENCES

[1] V. R. Basili, "The Role of Empirical Study in Software Engineering," in *2006 30th Annual IEEE/NASA Software Engineering Workshop*. Columbia, MD, USA: IEEE, Apr. 2006, pp. 3–6.

[2] S. Xu, "Empirical research methods for software engineering: Keynote address," in *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, Jun. 2017, pp. 1–1.

[3] "Quality management systems – requirements," International Organization for Standardization, September 2015, iSO Standard 9001. [Online]. Available: https://www.iso.org/standard/62085.html

[4] MISRA, *MISRA C 2012: Guidelines for the Use of the C Language in Critical Systems: March 2013*. Motor Industry Software Research Association, 2013.

[5] "Ieee standard for software quality assurance processes," IEEE Standards Association, June 2014, iEEE Std 730-2014. [Online]. Available: https://ieeexplore.ieee.org/document/6838485

[6] I. Sommerville, *Software engineering*, 9th ed. Boston: Pearson, 2011, oCLC: ocn462909026.

[7] K. Petersen and C. Wohlin, "Context in industrial software engineering research," in *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, 2009, pp. 401–404.

[8] C. Hobbs, "Software development standards," in *Embedded Software Development for Safety-Critical Systems*, 2nd ed., C. Hobbs, Ed. Elsevier, 2016, ch. 3, pp. 65–91.

[9] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–734, Aug. 2002.

[10] V. R. Basili, F. Shull, and F. Lanubile, "Building Knowledge through Families of Experiments," *IEEE Transactions on Software Engineering*, vol. 25, no. 04, pp. 456–473, Jul. 1999, publisher: IEEE Computer Society. [Online]. Available: https://www.computer.org/csdl/journal/ts/1999/04/e0456/13rRUx0xPVD

[11] B. Dit, E. Moritz, M. Linares-Vasquez, and D. Poshyvanyk, "Supporting and Accelerating Reproducible Research in Software Maintenance Using TraceLab Component Library," in *2013 IEEE International Conference on Software Maintenance*. Eindhoven, Netherlands: IEEE, Sep. 2013, pp. 330–339. [Online]. Available: http://ieeexplore.ieee.org/document/6676904/

[12] M. Nagappan, T. Zimmermann, and C. Bird, "Diversity in software engineering research," in *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*. Saint Petersburg Russia: ACM, Aug. 2013, pp. 466–476. [Online]. Available: https://dl.acm.org/doi/10.1145/2491411.2491415

[13] L. Madeyski and B. Kitchenham, "Would wider adoption of reproducible research be beneficial for empirical software engineering research?" *Journal of Intelligent & Fuzzy Systems*, vol. 32, no. 2, pp. 1509–1521, Jan. 2017.

[14] E. Murphy-Hill, G. C. Murphy, and W. G. Griswold, "Understanding context: creating a lasting impact in experimental software engineering research," in *FSE/SDP workshop on Future of SWEng research*, ser. FoSER '10. New York, NY, USA: Association for Computing Machinery, Nov. 2010, pp. 255–258. [Online]. Available: https://dl.acm.org/doi/10.1145/1882362.1882415

[15] J. Siegmund, N. Siegmund, and S. Apel, "Views on Internal and External Validity in Empirical Software Engineering," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, vol. 1, May 2015, pp. 9–19, iSSN: 1558-1225.

[16] J. M. González-Barahona and G. Robles, "On the reproducibility of empirical software engineering studies based on data retrieved from development repositories," *Empirical Software Engineering*, vol. 17, no. 1-2, pp. 75–89, Feb. 2012. [Online]. Available: http://link.springer.com/10.1007/s10664-011-9181-9

[17] H. Colebatch, "Introduction to the Handbook on Policy, Process and Governing," in *Handbook on Policy, Process and Governing*. Edward Elgar Publishing, 2018, pp. 1–13. [Online]. Available: https://www.elgaronline.com/view/edcoll/9781784714864/9781784714864.00005.xml

[18] W. A. Cram, J. G. Proudfoot, and J. D'arcy, "Organizational information security policies: a review and research framework," *European Journal of Information Systems*, vol. 26, pp. 605–641, 2017.

[19] S. J. Ball, "What is policy? 21 years later: reflections on the possibilities of policy research," *Discourse: Studies in the Cultural Politics of Education*, vol. 36, no. 3, pp. 306–313, May 2015.

[20] "NIST SP 800-12: Chapter 5 - Computer Security Policy." [Online]. Available: https://csrc.nist.rip/publications/nistpubs/800-12/800-12-html/chapter5.html

[21] R. Wies, "Using a Classification of Management Policies for Policy Specification and Policy Transformation," in *Integrated Network Management IV*, A. S. Sethi, Y. Raynaud, and F. Faure-Vincent, Eds. Boston, MA: Springer US, 1995, pp. 44–56. [Online]. Available: http://link.springer.com/10.1007/978-0-387-34890-2_4

[22] Ö. Kafali, J. Jones, M. Petruso, L. Williams, and M. P. Singh, "How Good Is a Security Policy against Real Breaches? A HIPAA Case Study," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, May 2017, pp. 530–540.

[23] M. Sloman, "Policy driven management for distributed systems," *Journal of Network and Systems Management*, vol. 2, no. 4, pp. 333–360, Dec. 1994.

[24] P. Naldurg and R. H. Campbell, "Modeling insecurity: policy engineering for survivability," in *ACM workshop on Survivable and self-regenerative systems*, ser. SSRS '03. New York, NY, USA: Association for Computing Machinery, Oct. 2003, pp. 91–98. [Online]. Available: https://doi.org/10.1145/1036921.1036931

[25] N. Dulay, E. Lupu, M. Sloman, and N. Damianou, *A policy deployment model for the Ponder language*, ser. IEEE/IFIP International Symposium on Integrated Network Management (IM'2001). Seattle: IEEE Press., Feb. 2001, pages: 543.

[26] R. S. Pressman, *Software engineering: a practitioner's approach*. McGraw-Hill Education, 2014.

[27] ISO/IEC 12207, *Systems and software engineering–Software life cycle processes*, International Standard Organization Std., 2017. [Online]. Available: https://www.iso.org/standard/63711.html

[28] D. Anandayuvaraj and J. C. Davis, "Reflecting on recurring failures in iot development," in *37th IEEE/ACM International Conference on Automated Software Engineering*, 2022, pp. 1–5.

[29] P. C. Amusuo, A. Sharma, S. R. Rao, A. Vincent, and J. C. Davis, "Reflections on software failure analysis," in *European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 1615–1620. [Online]. Available: https://dl.acm.org/doi/10.1145/3540250.3560879

[30] C. Okafor, T. R. Schorlemmer, S. Torres-Arias, and J. C. Davis, "Sok: Analysis of software supply chain security by establishing secure design properties," in *Proceedings of the 2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses*, 2022, pp. 15–24.

[31] M. Souppaya, K. Scarfone, and D. Dodson, *Secure Software Development Framework (SSDF) version 1.1 recommendations for mitigating the risk of software vulnerabilities*. Gaithersburg, MD: National Institute of Standards and Technology (U.S.), Feb 2022, no. NIST SP 800-218. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218.pdf

[32] Z. Newman, J. S. Meyers, and S. Torres-Arias, "Sigstore: software signing for everybody," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2353–2367.

[33] W. Humphrey, "Using a defined and measured personal software process," *IEEE Software*, vol. 13, no. 3, p. 77–88, May 1996.

[34] C. Wohlin, M. Höst, and A. Wesslén, "Can the personal software process be used for empirical studies," in *Proceedings ICSE Workshop on Empirical Studies of Software Development and Evolution*, 1999.

[35] C. Wohlin and A. Wesslen, "Understanding software defect detection in the personal software process," in *Proceedings Ninth International Symposium on Software Reliability Engineering (Cat. No.98TB100257)*, Nov 1998, p. 49–58.

[36] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. [Online]. Available: http://link.springer.com/10.1007/978-3-642-29044-2

[37] S. Keshav, "How to read a paper," *SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 3, p. 83–84, jul 2007. [Online]. Available: https://doi.org/10.1145/1273445.1273458

[38] H. He, R. He, H. Gu, and M. Zhou, "A large-scale empirical study on Java library migrations: prevalence, trends, and rationales," in *European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, Aug. 2021, pp. 478–490. [Online]. Available: https://dl.acm.org/doi/10.1145/3468264.3468571

[39] Q. Shen, H. Ma, J. Chen, Y. Tian, S.-C. Cheung, and X. Chen, "A comprehensive study of deep learning compiler bugs," in *ESEC/FSE*, ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, Aug. 2021, pp. 968–980. [Online]. Available: https://dl.acm.org/doi/10.1145/3468264.3468591

[40] L. Di Grazia and M. Pradel, "The evolution of type annotations in python: an empirical study," in *ESEC/FSE*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 209–220.

[41] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[42] S. S. Brilliant, J. C. Knight, and N. G. Leveson, "Analysis of faults in an n-version software experiment," *IEEE Transactions on software engineering*, vol. 16, no. 2, pp. 238–247, 1990.

[43] A. J. Perlis, F. Sayward, and M. Shaw, *Software metrics: an analysis and evaluation*. Mit Press, 1981, vol. 5.

[44] ACM SIGSOFT, "Acm sigsoft empirical standards," https://github.com/acmsigsoft/EmpiricalStandards, 2021, accessed: May 3, 2023.

[45] P. Bourque, R. E. Fairley, and I. C. Society, *Guide to the Software Engineering Body of Knowledge (SWEBOK(R)): Version 4.0*, 4th ed. Washington, DC, USA: IEEE Computer Society Press, 2023.