

# Towards Causal Analysis of Empirical Software Engineering Data

## The Impact of Programming Languages on Coding Competitions

Carlo A. Furia<sup>1</sup> · Richard Torkar<sup>2,3</sup> · Robert Feldt<sup>2</sup>

<sup>1</sup> Software Institute, USI Università della Svizzera italiana, Switzerland

<sup>2</sup> Chalmers and the University of Gothenburg, Sweden

<sup>3</sup> Stellenbosch Institute for Advanced Study (STIAS), South Africa

First version: 2023-01-18. Current version: 2023-06-27.

### Abstract

There is abundant observational data in the software engineering domain, whereas running large-scale controlled experiments is often practically impossible. Thus, most empirical studies can only report statistical *correlations*—instead of potentially more insightful and robust *causal* relations.

To support analyzing purely observational data for causal relations, and to assess any differences between purely predictive and causal models of the same data, this paper discusses some novel techniques based on structural causal models (such as directed acyclic graphs of causal Bayesian networks). Using these techniques, one can rigorously express, and partially validate, causal hypotheses; and then use the causal information to guide the construction of a statistical model that captures genuine causal relations—such that correlation *does imply* causation.

We apply these ideas to analyzing public data about programmer performance in Code Jam, a large world-wide coding contest organized by Google every year. Specifically, we look at the impact of different programming languages on a participant’s performance in the contest. While the overall effect associated with programming languages is weak compared to other variables—regardless of whether we consider correlational or causal links—we found considerable differences between a purely associational and a causal analysis of the very same data.

The takeaway message is that even an imperfect causal analysis of observational data can help answer the salient research questions more precisely and more robustly than with just purely predictive techniques—where genuine causal effects may be confounded.

## 1 Introduction

It is commonplace that “correlation does not imply causation”: just because two variables appear to change together, it does not mean that one *makes* the other change. This fundamental limitation is especially damning for fields like empirical software engineering, where there is plenty of detailed observational data—for example by mining software repositories—but fully controlled experiments are challenging to design and to run. As a result, most (quantitative) empirical studies in software engineering apply statistical analysis to detect *associations* (such as correlations), which may suggest and be consistent with—but do not necessarily establish—causal effects.

On the bright side, if our only, or primary, goal is *predicting* an outcome from measured variables, we do not necessarily need to understand causal relations, and a more traditional approach to statistical analysis may be enough. Practical solutions for engineering can often be built on prediction alone: the spectacular success of machine learning—which often learns from and predicts observational data—indicates that lots of applications can go a long way with correlational data alone. In contrast, if we want to *understand* the specific contributions of some variables on some outcomes, and to *generalize* their effects in a different scenario, we must be able to distinguish between mere correlations and true causal effects. After all, an understanding grounded in causal relations has not only practical value, but should be a fundamental goal of any rigorous scientific discipline.

In this paper, we demonstrate on a case study (outlined in Section 1.1) how thinking about causality can make a difference—even when all we have is observational data—to reliably isolate the impact of specific factors in software engineering empirical data, and to develop a statistical analysis that precisely answers our research questions. In a nutshell, we will apply some basic tools of causal analysis (in particular, causal directed acyclic graphs) to model and assess possible causal relations between variables [42].<sup>1</sup> This, in turn, will help us choose a statistical model that reliably identifies the magnitude of causal effects. We will also demonstrate that *not* accounting for causality leads to choosing different statistical models for our case study, which make accurate predictions “in the small” but spuriously misrepresent the impact of the variables of interest.

The main lesson following this exercise will be that software engineering empirical research needs conceptual and technical tools to model and reason about causal relations. Failing to do so—sticking to purely associational interpretations of empirical data—can severely limit the robustness and generalizability of any empirical results, and would ultimately stifle the long-term standing of software engineering’s scientific progress.

## 1.1 Programming: Competitions and Languages

As a case study of applying causal modeling to empirical software engineering data, we consider programmers that take part in *programming competitions*. More precisely, we analyze data collected during seven past editions of the Code Jam coding competition—organized every year by Google. Section 3 describes how the Code Jam contest works, and what public data we analyzed.

As they grow in popularity and accessibility, programming competitions help promote programming and its applications in informal and fun settings. From the point of view of empirical software engineering, programming competitions provide a setting where one can easily collect data about the behavior and performance of devoted programmers. Of course, competitions usually take place under somewhat artificial settings with numerous constraints, which do not necessarily reflect software development as done in a professional environment. On the flip side, the same constraints also make them a partially controlled environment, where it’s easier to zero in on some specific aspects of the behavior of programmers without too many confounding factors. In particular, a programmer’s work in a programming competition usually has very clear goals and constraints, and there are objective criteria to determine to what extent a submission meets those goals and satisfies those constraints.

In this paper, we use Code Jam data to study, in a specific setting, a long-standing question in computer science: the impact of using different programming languages. This evergreen topic featured in countless empirical studies—including some recent hot-button work [48; 6; 14]—that usually analyzed observational data by purely associational means. As summarized in Section 2, different studies do not always agree on their findings; but they tend to suggest that the association to using different programming languages is generally small compared to other correlational factors. In this paper, we take yet another look at the same topic but through the lens of causal relations. Does this stance affect the conclusions that we can draw from an empirical analysis?

Code Jam accepts submissions in a broad variety of programming languages; in practice, as we discuss in Section 3, only a handful of programming languages are overwhelmingly chosen by participants to Code Jam. In order to focus our analysis on data that is extensive and consistent, we only consider *i*) submissions in the most *popular languages* (namely, C++, Java, and Python), and *ii*) *experienced* participants—that is, participants who successfully took part in several editions of the Code Jam contest. Thus, we investigate the following fundamental research question.

**RQ:** For experienced participants to the Code Jam contest, how does using different programming languages relate to their results in the contest?

## 1.2 Correlation vs. Causation

The phrasing of question RQ is deliberately vague about what “relate to” means—in other words, what kind of relations we are looking for. Two common interpretations are:

**correlational:** are some programming languages *associated with* a certain performance in Code Jam?

---

<sup>1</sup>As we further discuss in Section 2.1, the over-arching goal of modern causal analysis is to formally capture the intuitive notion of causality with mathematical/statistical constructs, so that one can rigorously reason about causal links in experimental or observational data.

**causal:** does using some programming languages *affect* the performance in Code Jam?

The first interpretation only looks for correlations among the programming language and other variables (in particular, the outcome in the competition). The second interpretation is stricter, in a way, as it demands evidence that the programming language tends to *cause* a change in the competition outcome.

In this paper, we first carry out a statistical analysis following the *correlational/associational interpretation*. This represents a type of data analysis that is typical in present-day empirical software engineering. The analysis, described in Section 4, indicates that: *i*) using Java is associated with worse-than-average results in the Code Jam competition; *ii*) using Python is associated with better-than-average results; and *iii*) using C++ has no consistent association with better or worse results.

Section 5, revisits the correlational analysis under a *causal* lens. This represents a type of data analysis that could add value if used more frequently in empirical software engineering. In particular, Section 5.1 explains in what sense “correlation does not imply causation”: the Code Jam contest is obviously not a randomized experiment, and hence we cannot distinguish a causal relation from a mere correlation based on its data alone. However, we can still build, under some reasonable assumptions, a different statistical model than the one used in the correlational analysis that is poised to measure not mere correlations but causal effects. Remarkably, the analysis of this “causally-consistent” model, presented in Section 5.5, leads to results that are nearly the opposite of the correlational analysis’s: *i*) using Python is associated with worse-than-average results; *ii*) using C++ is associated with better-than-average results; and *iii*) using Java has no consistent association with better or worse results.

This difference in results between associational and causal analysis is especially troubling given that the former clearly outperforms the latter in terms of purely *predictive* accuracy; yet, it fails to properly isolate causal effects.

As we further discuss in Section 6, an underlying issue is that, in our data, the *magnitude* of the impact of choosing different programming languages is anyway small compared to other factors (such as a programmer’s individual skills). Thus, the main contribution of our work lies not so much in picking the “right” programming language but rather in demonstrating—on a case study of empirical programming data—that thinking about causality can be crucial to choose the statistical models that most accurately answer our research questions, and to identify confounding factors that may affect the reliability of a study’s results.

### 1.3 Contributions

This paper makes the following contributions:

- Highlights the importance, for empirical software engineering, of modeling and reasoning about causal relations to fully understand, and properly analyze, observational data;
- Demonstrates using some basic tools of causal analysis on a case study, where the main findings differ from those that one would draw based on a purely correlational (associational) statistical analysis;
- Argues, based on the case study, that a causal analysis’s capability of identifying confounders (sources of non-causal, spurious correlations) provides a basis to more clearly identify the relative importance of different factors, and more easily disentangle them;
- For reproducibility, all data and analysis scripts are available online—together with additional results and detailed data visualization:

REPLICATION PACKAGE: <https://doi.org/10.5281/zenodo.7541480> [15].

### 1.4 Organization

The rest of the paper is organized as follows. Section 2 summarizes some fundamental work on causal modeling and analysis, recalls some (recent) empirical work on the analysis of programming languages on software development practices, and discusses the few other applications of causal models to software engineering data analysis to date. Section 3 presents the Code Jam programming contest, the available data from several of its editions, and how we selected a subset of this data for analysis. Section 4 follows a “canonical”, purely correlational, statistical analysis of the Code Jam data—primarily using Bayesian techniques. Section 5 peruses the statistical models developed in the previous section according to a model of causal relations among observed variables, which leads to revising some key study results. Section 6 explains how to reconcile the two conflicting outcomes of Section 4’s and Section 5’s analysis, and corroborates previous evidence that the

absolute impact of programming languages is often small compared to other factors. Section 7 discusses further threats to the validity of the case study’s findings; and Section 8 concludes with a high-level discussion of the paper’s ideas and results.

## 2 Related Work

This section reviews related work in the main areas that are relevant for this paper. Section 2.1 outlines fundamental concepts of modern causal inference—applicable, in principle, to every branch of science and engineering. Section 2.2 summarizes studies that have applied causal analysis techniques to empirical data in the software engineering domain. Section 2.3 reviews studies of the impact of using different programming languages, including on data emerging from programming contests of various kind—this paper’s area of application.

### 2.1 Causal Inference

Causal inference denotes a process, usually built on top of mathematical/statistical concepts, for identifying and analyzing which factors lead to specific outcomes [41; 40]. It has grown in popularity as a tool for data analysis, and can be used in conjunction with, or in place of, traditional statistical and data analysis methods [46]. In a nutshell, while plain statistical methods identify *associations*, causal analysis determines which factors *lead to* (cause) changes in others, and *quantifies* such causal effects.

The so-called *ladder of causality* illustrates the crucial differences between various types of analysis [44]. Classical, associative statistical analysis is the bottom level (“seeing”), addressing questions of the form “How does seeing  $X$  change my belief in  $Y$ ?”. The analysis of interventions (“doing”) is one step higher: “What would happen to  $Y$  if I do  $X$ ?”. Counterfactual analysis is the highest level (“imagining”): “What would have happened to  $Y$  if  $X$  had not occurred?”.

In principle, the benefits of causal inference in empirical research are clear: finding the causes of studied effects is a key goal of science. However, it can also have practical benefits, leading to more robust and accurate estimates. For instance, a re-analysis of medical data on hip fractures among elderly patients found that causal analysis could identify which factors contribute to the occurrence of hip fractures and to what extent, and also provide predictions that were comparable to traditional methods [7]. Another study [49] demonstrated that medical diagnosis based on causal inference was almost twice as effective (25th percentile vs. 48th percentile of human doctors) as those based on traditional, associational statistical methods.

One key ingredient of modern-day causal inference is the use of directed acyclic graphs (DAGs) to model the observed factors’ dependence *structure*. Nodes in a DAG denote factors, and edges denote which nodes (edge source/parent) cause changes in other nodes (edge target/child). Thus, the DAG of a causal model makes the causal dependence structure explicit. In addition to DAGs, a causal analysis framework normally offers means of *estimating* the value of effect nodes from the values of their cause nodes. In a so-called structural equation model, this is achieved via equations—typically linear, but more general forms have also been used—that relate each child node to its parents in the DAG. Another kind of causal model, the probabilistic causal model, uses instead a single probability distribution over all factors [41].

The causal structure (the DAG) underlying some empirical data may be known a priori or be deduced from domain theory. In some cases, one can apply so-called structural learning (also called causal discovery) to identify causal structures from data [57]. Structural learning methods may combine observational data with interventional data—that is, data measured after taking actions that cause changes in some of a DAG’s factors. A key benefit of structural causal models is that DAGs can be used both to guide which factors to intervene on (change), and to calculate the causal effects from the observations.

There are frameworks for causal inference that are not based on the modeling of causal structure with DAGs. A notable example is the potential outcomes framework [51], also referred to as the Neyman-Rubin causal model. This framework estimates the causal effect of an intervention by comparing the outcome that would have occurred under the intervention to the outcome that would have occurred in the absence of the intervention. Pearl has criticized this approach by pointing out concrete mistakes in the estimation of causal effects that may occur if failing to consider the causal structure [43]. However, the potential outcomes framework may still have advantages in some fields such as econometrics [27], where it fits better their typical assumptions and restrictions. In this paper, we focus on DAG-based causal inference, and leave the evaluation of other causal inference approaches to future work.

## 2.2 Causal Inference and Analysis in Software Engineering

While causal analysis is certainly not (yet) a common practice in software engineering research, there is a growing number of studies that tried to apply it to specific problems. In a recent survey [55], Siebert reviews a total of 31 studies that applied some kind of causal analysis to software engineering data. All these papers were published after 2010, and are considered “software engineering” because they focus on at least one of the activities of the SWEBOK (Software Engineering Body of Knowledge) [1]. The papers belong to four groups depending on their main activity: fault localization (17 papers), testing (4), performance analysis (5), and others (5).

The 17 fault localization papers build on, or at least cite, Baah et al.’s work [3]. Most of them model the causal effect that program elements (such as variables and statements) have on values and, ultimately, failures in an executing program; identifying these effects can help locate and debug actual faults. One paper in this group [31], targets a different unit of analysis, as it is concerned with fault localization in services composed of multiple, interacting programs.

Among the four testing papers, one builds on the authors’ fault localization results to identify elements to mutate [30]. Clark et al. [9] test simulation models to determine whether they exhibit the expected causal effects. Liu et al. [32] apply Bayesian causal modeling to improve the A/B testing used to evaluate changes in automotive software; they use the potential outcomes framework as their model of causality.

Most of the five papers on performance analysis focus on identifying components that determine a system’s observed performance. An exception is the paper by Scholz et al. [53], which focuses instead on the performance of fault prediction algorithms.

The remaining five papers surveyed by Siebert [55] are on miscellaneous topics that do not fit any of the other main categories: the impact of open-source license choice, newsletter strategies, the effect of gender on pull request acceptance, and productivity analysis.

Not all papers considered by Siebert apply a full set of causal modeling tools. Most of them deploy some specific statistical methods that can help detect causal effects, but do not use structural/graphical models (DAGs). The few papers that did consider graphical causal models built them by applying causal discovery algorithms—possibly complementing their output with expert, manual modeling.

A few additional recent papers applying causal analysis to software engineering data did not make it into Siebert’s survey. Heyn and Knauss [26] use causal relations and DAGs to clarify, communicate, and apply human knowledge to artificial intelligence system development. Fang et al. [12] use an econometric causal inference technique (based on the aforementioned potential outcomes framework) to study if tweeting about open-source development projects can affect their popularity. Finally, Dubslaff et al. [11] apply Halpern and Pearl’s concept of actual causality [22] to understand and analyze dependencies in configurable software systems.

## 2.3 Empirical Studies of Programming Languages

Empirical studies of programming languages fall into three main categories according to what kind of data they collect and analyze. *Controlled experiments* usually compare a small set of language features on a specific task; for example, static vs. dynamic type systems [23], or different concurrency primitives [50; 39]. Their precise, controlled experimental setups work well to investigate fine-grained questions; for the same reason, their results may have limited generalizability. *Surveys*, interviews, and other ways of methodically collecting individual observations are the tool of choice to investigate qualitative questions and to assess programmer perceptions and subjective preferences; for example, to understand trends behind language preferences and adoption [36]. *Repository* data, obtained by mining GitHub or other artifact repositories (such as the Code Jam contest data we analyze in this paper), offer plenty of information, which can support large-scale studies [54; 48; 38]; the flip-side is that the data may be noisy, or too heterogeneous to discern robust, general trends within it.

In fact, empirical studies of programming languages based on mining repositories can lead to clear, solid results when they target characteristics that are easy to measure reliably (such as the *conciseness* of programs written in different programming languages); in contrast, many confounding factors may arise that complicate discerning the actual, practical impact of programming languages on features lacking a universally accepted, easy-to-measure operationalization (such as *error proneness*). As an example, take Nanz and Furi’a’s analysis of the Rosetta Code repository [37], which compared solutions to programming tasks written in eight programming languages. Their comparison of conciseness found strong effect-size differences (Cohen’s  $d > 0.7$ ) for 14 out of 28 language pairs, corroborating other studies of conciseness [47]. In contrast, their comparison of failure proneness (simplistically measured as how many programs terminate with a runtime

failure) found no strong effect-size differences, and only five medium effect-size differences ( $0.3 \leq d < 0.5$ ), among the same 28 language pairs.

Fault proneness is an especially tricky measure to associate with the chosen programming language. Ray et al.’s study of “code quality” in GitHub projects [48] also failed to find *strong* associations between programming languages and presence of bugs, concluding that, while “some languages have a greater association with defects than other languages [...] the effect is small”; Berger et al.’s critical reanalysis of [48] suggested to further “reduce the number of languages with an association with defects” and found that “the practical effect size is exceedingly small” [6]. Our Bayesian reanalysis of the same data [14] also found that “the fault proneness of a language over another strongly depends on the conditions in which the languages are to be used”; namely, any “disproportionate differences [...] are project-specific rather than language-specific”. Therefore, it is not surprising that the present paper will also conclude (Section 6) that the choice of programming language has a modest effect on predicting the success of a person participating in the Code Jam competition, and it is largely secondary compared to other dominant factors—most important, the intrinsic skills and experience of each participant.

**Empirical studies of programming contests.** Programming contest empirical data have also been the subject of empirical studies—usually with a focus on using such contests to foster educational programs. Often, such studies are mainly qualitative, aiming at surveying the state of the art as broadly as possible [21; 29], and at suggesting ways to improve the impact of programming contests in education [61].

To our knowledge, Back and Westman’s master’s thesis [4]<sup>2</sup> is the only other empirical study of Google Code Jam data to date. It analyzes solutions to the contest’s problems written in C, C++, C#, Java, and Python, and compares them for attributes such as size, running time, and memory consumption (along the lines of [38]).

### 3 The Google Code Jam Data

Code Jam is a worldwide programming contest organized by Google every year since 2008. Each edition consists of several elimination rounds; all rounds—except possibly the final one—take place online.

Each round begins and ends at predefined times set by the organizers. As long as a round is open, registered participants can download the coding problems for that round and submit their solutions. Solutions can be written in a wide variety of programming languages, provided they are runnable on the platform Google sets up for the contest.

Participants can submit as many solution attempts as they want for each problem, in any order. A submission gets a number of points that depends on the number of tests it passes in the test suite that accompanies each problem. Participants are ranked based on the points of their best submission in each problem, using the submission time as a tie breaker. At the end of each round, the top participants in the ranking advance to the next round.

Figure 1 illustrates the progression of the 2010 edition of Code Jam, which included seven rounds. Initially, 8459 programmers entered the preliminary qualification round; 5553 were admitted to the next stage, and allowed to submit to one or more of the three parallel rounds 1A, 1B, and 1C; 1924 participants were ranked high enough to progress to round 2; 361 further progressed to round 3; and 24 took part in the world finals that determined the overall contest winners.

While some details of the contest (such as the number of rounds, how many advance to the next round, or some details of how points are counted) may change from year to year, the general contest structure has remained largely unchanged. Thus, we can merge the data about submissions to Code Jam in several yearly editions of the contest and analyze them consistently. Furthermore, we do not study the progression of participants within a contest, but treat each round as a separate ranker of the participants’ performance based exclusively on their submissions to that round.

#### 3.1 The Code Jam Dataset

Data about all submissions to Code Jam are publicly available on the contest’s website.<sup>3</sup> We analyzed data for the first seven yearly editions of the contest, 2008–2014, which we got from an unofficial database dump of the contest data.<sup>4</sup>

<sup>2</sup>This thesis was supervised by the first author of the present paper.

<sup>3</sup><https://codingcompetitions.withgoogle.com/codejam>

<sup>4</sup><https://www.go-hero.net/jam/>

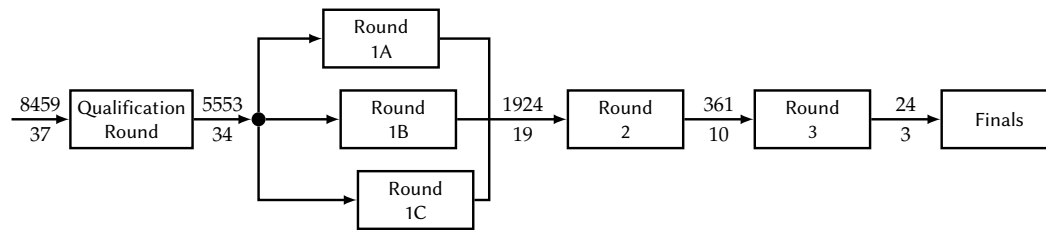


Figure 1: The structure of the 2010 edition of Google Code Jam. Each arrow indicates the number of participants entering each round of the contest (above the arrow) and the number of different programming languages used by these participants (below the arrow). Rounds 1A, 1B, and 1C are accessible to all participants who passed the qualification round.

A *datapoint* in this dataset reports information about a participant’s results in a single *round* of the Code Jam contest, consisting of values for following variables:

*challenge*: an identifier of the yearly edition of Code Jam and the round number within that edition

*nickname*: the unique identifier of the programmer who submitted in this challenge

*language*: the programming language used by these submissions

*size*: the submissions’ total size (in bytes)

*rank*: the final rank of the participant’s submissions within the challenge (1 is the top rank)

As we discussed above, each round includes one or more problems; for each problem, a participant’s submission with the most points is taken as their final submission. Since a participant’s rank depends, in general, on the points that all their (final) submissions collected in the round, *size* is simply the sum of the sizes of all their final submissions in that round, to give an idea of their overall programming effort. In the following, we sometimes refer to any such datapoints as “round/challenge participation”.

A clarification about terminology: a *high* rank denotes a rank that is close to the top rank (the first rank), whereas a *low* rank denotes a rank that is further away from the top rank. Thus, a high-rank submission is one with a small ordinal number; and a low-rank submission is one with a large ordinal number. Consistently with this standard usage, we will talk about *better-than-average* ranks to denote smaller-than-average rank ordinals, and *worse-than-average* ranks to denote larger-than-average rank ordinals.

### 3.2 Selecting Programming Languages

The Code Jam dataset includes submission in 75 different programming languages, including esoterica such as Whitespace<sup>5</sup> and GolfScript.<sup>6</sup> As we might expect, though, the popularity of languages in Code Jam is not uniform but highly skewed, with the usual suspects dominating. More precisely, the top 10 most used languages are: 1. C++ (50% of all datapoints); 2. Java (20%); 3. Python (13%); 4. C# (5%); 5. C (5%); 6. Ruby (2%); 7. PHP (1%); 8. Perl (1%); 9. Pascal (1%); 10. Haskell (1%).

We want to focus our analysis on a small number of languages that are widely used in Code Jam, so that there is abundant data to reliably estimate correlations and effects for those languages. Therefore, we only consider submissions in the top-3 most used languages (C++, Java, and Python), which account for an order of magnitude more submissions than all other languages combined. In particular, these languages are disproportionately used by the programmers that advance to the later rounds of the contest; for example, they are the three languages used in the 2010 final round (rightmost arrow in Figure 1).

### 3.3 Selecting Experienced Participants

Even after ignoring all but the three most used programming languages, the Code Jam dataset still includes 150 539 datapoints for the submissions by 49 808 unique participants. Besides being an impractically large dataset to analyze, it dilutes a smaller number of *experienced* participants within a far greater number of casual participants. As we can see more clearly in Figure 2, the overwhelming majority of participants took part in only one or two yearly editions of Code Jam, where they advanced at most to the second round.

<sup>5</sup>[https://en.wikipedia.org/wiki/Whitespace\\_\(programming\\_language\)](https://en.wikipedia.org/wiki/Whitespace_(programming_language))

<sup>6</sup>[https://en.wikipedia.org/wiki/Code\\_golf#Dedicated\\_golfing\\_languages](https://en.wikipedia.org/wiki/Code_golf#Dedicated_golfing_languages)

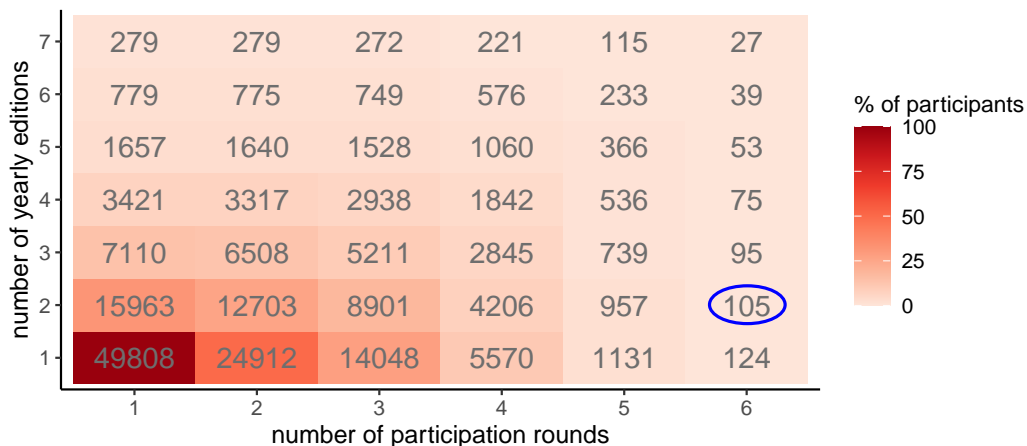


Figure 2: Number of participants to Code Jam who entered at least  $y$  yearly editions of Code Jam and at least  $x$  participation rounds in any one edition. The background color denotes the corresponding percentage of all 49 808 participants to Code Jam. An ellipse marks the subset of participants featuring in the main analysis.

To select experienced participants, we further filter the dataset and only retain data about participants who took part in at least two yearly editions of Code Jam and submitted, in any one edition, to at least six rounds. Since each edition of Code Jam in our dataset includes seven rounds or more, the latter requirement is met by who qualifies for the penultimate round at least once—clearly, an above-average result. This selection criterion leaves us with 1 896 datapoints corresponding to the submissions of 105 experienced participants to Code Jam.

We found that this data selection reasonably balances dataset size and experience of participants: it’s not so small as to be insignificant, nor so large as to be impractical to analyze and including too many casual participants. However, there is nothing special about it: other data selections could also be reasonable and feasible. To demonstrate this claim, Section 7.1 discusses the results of running the same analysis on significantly larger and smaller samples of the Code Jam dataset. While some detailed results may change, the overall answer to our research question mostly does not, and remains largely independent of the data selection criteria within a broad range.

## 4 Correlational Analysis of Programming Language Effects

This section presents a statistical analysis of the Code Jam dataset. We build four Bayesian regression models  $m_1, m_2, m_3, m_4$  of increasing complexity, and we compare them quantitatively according to their predictive capabilities. More precisely, we use generalized linear (regression) models: a broad, widely-used family of flexible statistical models [18]. The analysis follows the guidelines we discussed in previous work [14]; for brevity, we do not discuss here the application of the guidelines, but refer the interested readers to the replication package for details.

By following the recommended guidelines to build and evaluate regression models, this section’s analysis is arguably similar to any standard, up-to-date, rigorous statistical analysis of the same data made by an informed researcher. Even with some latitude to accommodate their favorite statistical techniques, the big picture that emerges should still be consistent with this section’s. In particular, Section 4.4 explains that using frequentist rather than Bayesian techniques would lead to similar overall conclusions.

### 4.1 Modeling

Figure 3 shows the four models, whose components we now describe and justify. These four models follow standard guidelines, and are serviceable to support the analysis of the paper; but we do not imply that they are the definitive or only models for the data. The replication package includes a few model variants, as well as the data that can be fitted on any other kind of statistical model to explore different research questions.



$$\begin{aligned}
 \text{rank}_i &\sim \text{NegativeBinomial}(\lambda_i, \phi) \\
 \log(\lambda_i) &= \underbrace{\alpha_{\text{language}[i]}}_{m_1} + \underbrace{\alpha_{\text{nickname}[i]} + \alpha_{\text{challenge}[i]}}_{m_2} + \underbrace{\beta \cdot \log(\text{size}_i)}_{m_3}
 \end{aligned}$$

(a) Likelihoods of models  $m_1$ ,  $m_2$ ,  $m_3$ , and  $m_4$ .

$$\begin{aligned}
 \alpha_\ell &\sim \text{Normal}(5, 3.5) && \text{for every language } \ell \\
 \alpha_n &\sim \text{Normal}(0, 0.5) && \text{for every nickname } n \\
 \alpha_c &\sim \text{Normal}(0, 0.5) && \text{for every challenge } c \\
 \beta &\sim \text{Normal}(0, 1) \\
 \phi &\sim \text{Gamma}(0.01, 0.01)
 \end{aligned}$$

(b) Priors for the coefficients  $\alpha$ ,  $\beta$ , and  $\phi$  of models  $m_1$ ,  $m_2$ ,  $m_3$ , and  $m_4$ .

Figure 3: Definitions of regression models  $m_1$ ,  $m_2$ ,  $m_3$ , and  $m_4$ . All these models have the same negative binomial likelihood with log-linear parameter  $\lambda$ ; each model  $m_k$  includes exactly  $k$  predictors.

**Variables.** First of all, we should choose the *variables* of our models. Since our goal is to analyze the impact of programming languages and other factors on the results in rounds of the Code Jam contest, we pick *rank* as *outcome* variable in all our models.

The other variables in the dataset can be used as *predictors* (also: *treatments*). Since we are especially interested in the effect of programming languages, we include variable *language* as a predictor in all our models. Then, we introduce progressively more complex models by adding the other variables as predictors: model  $m_1$  only uses *language*; model  $m_2$  adds *nickname*, since different participants are likely to have different results that reflect their skills; model  $m_3$  adds *challenge*, since each year/round of the competition is different, and arguably has a different intrinsic complexity; and model  $m_4$  adds *size*, which partially indicates the amount of work done by a participant in each round.

Plausibly, some of these four “independent” variables interact (for example, participants may have their own favorite programming language), and hence they are not really independent. To account for this, Section 4.2 compares the four models using a kind of regularization that weighs off predictive accuracy against complexity: any model with redundant variables will be penalized under this criterion. In a frequentist setting (Section 4.4), a variable selection process provides a similar safeguard against including strongly interacting variables. Later, Section 5 will revisit variable interactions and interpret them under a causal lens.

**Likelihood.** The *likelihood* is a probability distribution of the outcome given some parameters; in a regression model, the likelihood’s parameters are (generalized) linear functions of the predictors.

Since the outcome variable *rank* is a positive integer, the Poisson distribution family is the most appropriate choice since it has the highest information entropy for this kind of data [28]; in other words, it does not encode any assumption other than that *rank* must be integer and not negative [33]. Specifically, we pick the *negative binomial* distribution within the Poisson family, which is better suited for data that is overdispersed. This is the case of variable *rank* whose mean 929 is much smaller than its variance 3 962 575. Mean and variance coincide in a Poisson distribution, which would fail to accurately capture the distribution of *rank*; in contrast, a negative binomial distribution  $\text{NegativeBinomial}(\lambda, \phi)$  has two parameters  $\lambda$  and  $\phi$ , so that its mean  $\lambda$  can differ from its variance  $\lambda + \lambda^2/\phi$ .

**Parameters.** The *mean*  $\lambda$  of the negative binomial distribution in our models is a linear function of the predictors after applying a customary logarithmic *link function*, which ensures that the mean remains non-negative.

**Model  $m_1$ .** In model  $m_1$ ,  $\log(\lambda_i) = \alpha_{\text{language}[i]}$ : in each datapoint  $i$ , the logarithm of parameter  $\lambda$  equals a language-dependent parameter  $\alpha_{\text{language}[i]}$ . In other words, model  $m_1$  estimates a different mean for each programming language, and uses that to predict a challenge’s rank.

**Model  $m_2$ .** In model  $m_2$ ,  $\log(\lambda_i) = \alpha_{\text{language}[i]} + \alpha_{\text{nickname}[i]}$ , where  $\alpha_{\text{nickname}[i]}$  is another parameter, which depends on *nickname*. In other words, model  $m_2$  estimates the mean rank by combining mean estimates for each programming language and for each participant (*nickname*).

**Model  $m_3$ .** In model  $m_3$ ,  $\log(\lambda_i) = \alpha_{\text{language}[i]} + \alpha_{\text{nickname}[i]} + \alpha_{\text{challenge}[i]}$ , where  $\alpha_{\text{challenge}[i]}$  is another parameter, which depends on *challenge*. In other words, model  $m_3$  estimates the mean rank by also combining mean estimates for each challenge (a round in a specific yearly edition of the contest).

**Model  $m_4$ .** In model  $m_4$ ,  $\log(\lambda_i) = \alpha_{\text{language}[i]} + \alpha_{\text{nickname}[i]} + \alpha_{\text{challenge}[i]} + \beta \cdot \log(\text{size}_i)$ , where  $\beta$  is a new linear parameter, which multiplies the logarithm of *size*. In other words, model  $m_4$  estimates the mean rank by also estimating a contribution proportional to a submission’s size. We use the *logarithm* of *size* because this variable varies greatly in the dataset (from 457 to 202 965 bytes); taking the logarithm is a standard practice to smoothen a wide variability range and focus on its “orders of magnitude” changes.

Parameters  $\alpha_\ell$  for every language  $\ell$ ,  $\alpha_n$  for every nickname  $n$ , and  $\alpha_c$  for every challenge  $c$  are also called “intercepts”, since they correspond to constant terms in the linear model. For a similar reason, parameter  $\beta$  is also called “slope”.

**Priors.** The last components of a Bayesian model are *priors*, that is “initial” probability distributions on the parameters  $\alpha_\ell$  for every language  $\ell$ ,  $\alpha_n$  for every nickname  $n$ ,  $\beta$ ,  $\alpha_c$  for every challenge  $c$ , and  $\phi$ . The rest of this section explains and justifies our choice of priors; however, there is a good deal of latitude in how we precisely select them—as long as they pass the validation that we outline next.

We use standard weakly informative priors for all parameters: normal distributions for the  $\alpha$ s and  $\beta$ , and a gamma distribution for  $\phi$ . As you can see in Figure 3b, the priors of the language intercepts  $\alpha_\ell$  are not centered (their mean is 5); in contrast, the priors of  $\alpha_n$ ,  $\alpha_c$ , and  $\beta$  are centered (their mean is 0) and have smaller standard deviations than the prior of the language intercepts  $\alpha_\ell$ . This reflects the analysis’s focus on the effect of languages: every model includes the language intercepts, which are thus the primary determinant of the overall expected rank and are comparable between models; then, each other predictor may move the overall estimate higher or lower, starting from the language’s baseline. Since the rank is a positive integer, the mean rank must also be positive; hence, a prior with a positive mean is more likely to be consistent with the data. However, the language intercepts’ broad prior standard deviation makes the prior weak; this means that the data can freely sway the estimate of the actual impact of languages in either direction. As shown by the validation step, these priors make for an efficient fitting of the data; but there’s nothing “special” about them, since they are broad, and in fact different choices for the priors would lead to very similar overall estimates.

**Model validation.** Before we fit the models and use them for analyzing the data and answering our RQ, we should *validate* them to ensure that they are well-formed and their predictions can be trusted [14]. Here, we summarize the outcome of validation; all details are in the replication package.

First, we check that the priors are *plausible*, that is they are not unnecessarily constraining compared to the observed data. The prior predictive simulation plots confirm that this is the case for all four models.

Second, we check that the models are *workable*, that is they can be fitted without computational problems such as divergences. To this end, we inspect the fitting diagnostic metrics (divergent transitions,  $\hat{R}$  ratio of within-to-between chain variance, effective sample size, and trace plots) and confirm that they are within the expected ranges for all four models.

Third, we check that the models are *adequate*, that is they can generate data that resembles the analyzed empirical data. This is an important property to ensure that the model does not merely pass some diagnostic checks but can actually be trusted to capture (at least some) aspects of the data. Inspecting the posterior predictive checks plots for all four models confirms that they are all reasonably adequate.<sup>7</sup>

## 4.2 Model Comparison

Given that all four models pass the fundamental well-formedness checks, which model should we use for our analysis? Our current analysis goal is making good predictions—that is, estimating correlations between predictors and outcome. Thus, we can use an information-theoretic criterion to assess whether some model offers better out-of-sample predictions than some other models.

---

<sup>7</sup>As expected, the adequacy of the simplistic model  $m_1$  is visibly lower than the other, more complex models; however, they all are capable of reflecting at least the general shape and trend of the empirical data.

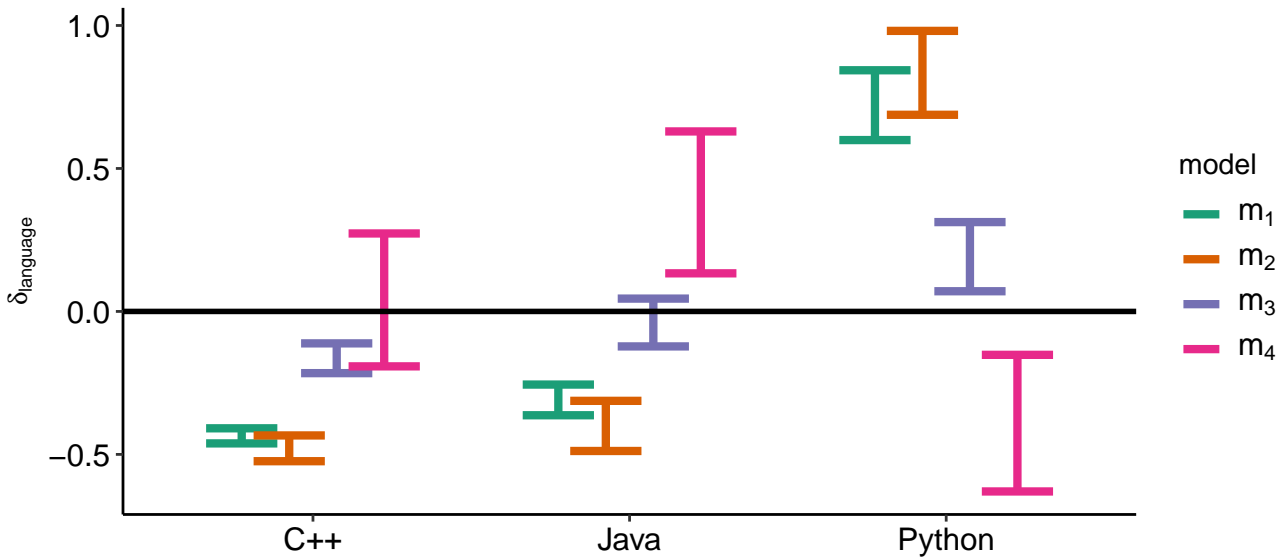


Figure 4: For each language  $\ell$ , in each model  $m_1, m_2, m_3, m_4$ , the 50% highest posterior density intervals of the distribution of the difference  $\delta_\ell$  between  $\ell$ 's intercept  $\alpha_\ell$  and the average intercept  $\bar{\alpha}_L$  for all languages in the model. See Section 4.3 for details on how these intervals are computed.

In particular, we use the widely used PSIS-LOO criterion [60], which estimates the *relative* adequacy of a number of competing models using a form of leave-one-out validation. Applying PSIS-LOO ranks the four models according to a score that estimates the relative predictive capabilities of one model compared to the others; each score difference has a standard error, which quantifies the uncertainty in the score difference.

| MODEL | RANKING | SCORE DIFFERENCE |              |
|-------|---------|------------------|--------------|
|       |         | ABSOLUTE         | STANDARDIZED |
| $m_4$ | 1       | –                | –            |
| $m_3$ | 2       | –387.6           | –14.6        |
| $m_2$ | 3       | –785.4           | –18.5        |
| $m_1$ | 4       | –85.8            | –2.1         |

Table 1: Comparison of four MODELS according to the PSIS-LOO validation criterion. The models are ranked from best to worst according to the criterion (the “best” model has rank 1). The table also reports, the relative DIFFERENCE in predictive capability score between each model and the one that precedes it in the ranking, both in ABSOLUTE value and STANDARDIZED (i.e., absolute value difference divided by standard error).

Table 1 shows the results of applying PSIS-LOO to the models  $m_1, m_2, m_3, m_4$ . Model  $m_4$  clearly outperforms all others in terms of predictive capabilities: the score of  $m_3$ —the next best model in the ranking—is a massive 14.6 standard errors lower. In turn, model  $m_3$  outperforms model  $m_2$ : their difference is still massive (18.5 standard errors). Model  $m_1$  is the worst, although it’s closer to  $m_2$  (2.1 standard errors).

The conclusion of this comparison is indisputable: if our goal is *predicting* with accuracy the statistical associations between the predictors (including *language*) and the outcome *rank*, we should definitely use model  $m_4$ , which offers far better out-of-sample predictive accuracy than the other, simpler models that incorporate less information.<sup>8</sup>

### 4.3 Analysis: Programming Language Associations

Fitting a model on the Code Jam data gives a *posterior distribution* of the parameters  $\alpha_\ell$  for each language  $\ell = \text{C++}, \text{Java}, \text{Python}$ —as well as distributions for the other parameters  $\alpha_n, \beta$ , and  $\alpha_c$ , but our RQ focuses on

<sup>8</sup>The replication package shows that  $m_4$  even outperforms considerably more sophisticated multilevel models that explicitly model variable interactions but do not include variable *size*.

| MODEL | BETTER    | WORSE  | NO ASSOCIATION |
|-------|-----------|--------|----------------|
| $m_1$ | C++, Java | Python |                |
| $m_2$ | C++, Java | Python |                |
| $m_3$ | C++       | Python | Java           |
| $m_4$ | Python    | Java   | C++            |

Table 2: For each MODEL, which languages are associated with BETTER and WORSE contest results, or have NO consistent ASSOCIATION, at the 50% probability level.

the analysis of languages.

To answer our RQ, we compute *contrasts* that quantify the difference between the absolute value of the  $\alpha_\ell$  for different languages. For each  $\ell$ , we derive the distribution<sup>9</sup> of  $\delta_\ell = \alpha_\ell - \bar{\alpha}_L$ , where  $\bar{\alpha}_L$  is the mean of the distribution of all coefficients  $\alpha_{C++}, \alpha_{Java}, \alpha_{Python}$  taken together. In other words,  $\delta_\ell$  is the difference between language  $\ell$ 's contribution to the rank and the average language contribution to the rank in the model.

Figure 4 plots the 50% highest-posterior density intervals of the distributions of  $\delta_\ell$  for each language  $\ell$  and in each model. If an interval  $I_\ell$  for language  $\ell$  is entirely below the zero line, it means that, with 50% probability, submissions written in language  $\ell$  are associated with better-than-average ranks (that is, smaller rank ordinals in Code Jam). If  $I_\ell$  is entirely above the zero line, it means that, with 50% probability, submissions written in language  $\ell$  are associated with worse-than-average ranks (that is, larger rank ordinals in Code Jam). If  $I_\ell$  includes the zero line, there is no consistent association with better- or worse-than-average results for language  $\ell$  at 50% probability.

Why do we focus on 50% probability intervals, instead of a higher, more common probability such as 95%? As we demonstrate in Section 6, the absolute magnitude of the impact of programming languages is modest compared to other variables in our data. Therefore, there is quite some uncertainty about the precise effects associated with languages, and we are unlikely to find consistent effects at high probability levels. The 50% probability intervals are still meaningful and give a “sense of where the parameters and predicted values will be” instead of aiming for “an unrealistic near-certainty” [17]. In the replication package, we also compute the 70%, 90%, and 95% probability intervals, showing that they still largely exhibit the same *tendencies* as the 50% intervals, but with larger uncertainty (for example, an interval that is completely above zero at 50% probability is *mostly* above zero at 95% probability—but also includes it). In other words, higher-probability intervals display the *same* tendency as the 50% intervals—just with more uncertainty about the precise interval spans. In summary, 50% probability intervals provide robust, valuable information, that clearly conveys the main qualitative findings of our analysis.

**Results: model  $m_4$ .** According to model  $m_4$  (which model comparison identified as the one delivering the most accurate predictions), at the 50% probability level:

- Python is associated with smaller-than-average rank ordinals, that is *better* contest results;
- Java is associated with larger-than-average rank ordinals, that is *worse* contest results;
- there is no consistent association for C++, as its 50% interval is nearly centered around zero.

Based on this, we can answer the paper’s main RQ under the correlational interpretation:

**AQ (correlation):** For experienced participants to the Code Jam contest, using Python is associated with better contest results, using Java is associated with worse contest results, and using C++ has no consistent association.

**Results: models  $m_1, m_2, m_3$ .** Further inspecting Figure 4 clearly indicates that the predictions of the four models  $m_1, m_2, m_3$ , and  $m_4$  are often inconsistent with each other. Table 2 summarizes the predictions of each model: in particular, models  $m_1, m_2$ , and  $m_3$ 's estimates are nearly the opposite of model  $m_4$ 's. Since we have shown that  $m_4$  outperforms the other three models in terms of predictive accuracy, it makes sense to base the correlational analysis on  $m_4$ .

<sup>9</sup>More precisely, we numerically estimate it on the posterior samples.

**Robustness.** When the qualitative results of a statistical analysis “flip” as we refine models (such as when going from  $m_3$  to  $m_4$ ), it may indicate that the associations we are measuring are weak. Section 6 will confirm this suspicion, detailing how the specific contribution of *language* is small compared to the other variables’. From a different angle, this lack of robustness in the statistical analysis also prompts us to think about possible *causal* relations, in order to go beyond mere predictive accuracy towards understanding the genuine causes of the observed associations.

#### 4.4 Frequentist Statistics

In related work [13; 14], we argued extensively about the advantages of Bayesian statistics over the more traditional frequentist statistics. This recommendation applies regardless of whether we are just interested in prediction or are looking for causal links.

Nevertheless, most of the observations and analyses we discuss in the present paper would remain valid when using frequentist statistics. To substantiate this claim, the replication package presents four statistical models  $f_1, f_2, f_3, f_4$  that are the frequentist counterparts to the four Bayesian models  $m_1, m_2, m_3, m_4$  discussed in the paper. Except for a few technical differences—such as how we encode contrasts and binary indicator variables, and using flat rather than weakly informative priors—each frequentist model  $f_k$  is a regression model using the same predictors and outcome variables, the same negative binomial likelihood, and the same log-linear parameterization as the corresponding Bayesian model  $m_k$ . Fitting  $f_k$  on the Code Jam dataset and analyzing it similarly to how we did for  $m_k$  also leads to very similar results in terms of which programming languages are associated with worse or better performance in the contest. The frequentist analysis also finds that the predictions based on model  $f_4$  (like  $m_4$ ) are nearly opposite of those based on model  $f_3$  (like  $m_3$ ).

The frequentist approach also agrees with our Bayesian approach regarding which model is “best” for predictions. To this end, we apply a variable selection process that starts from model  $f_1$  (with a single predictor, like model  $m_1$ ), adds<sup>10</sup> one predictor at a time, and checks whether it “significantly” improves model performance (measured by  $p$ -values or by information criteria such as AIC). This process continues until it selects model  $f_4$ , which is strictly better than all other simpler models according to this criterion.

In summary, we can go back to our four Bayesian models for the rest of the analysis, knowing that the high-level contributions and results of the paper—most important, the causal analysis results—are largely independent of whether one prefers to use Bayesian or frequentist regression models.

## 5 Causal Analysis of Programming Language Effects

This section discusses how to build an analysis of the Code Jam dataset that tries to identify true *causal* effects of the different programming languages; and compares its results to the purely associational analysis presented in the previous section.

### 5.1 From Correlation To Causation

In general, one cannot identify causal relations by analyzing data alone, as these relations depend also on how the data was *generated*.<sup>11</sup> This implies that we cannot just lift the results of the statistical analysis to infer causal effects without any assumptions on the data-generation process.

To illustrate this point, imagine that Code Jam were run as a fully controlled experiment designed to find the impact of using different programming languages. This means that all independent variables would be assigned randomly within the population of interest: random programmers (variable *nickname*) would be selected to join the competition; they would be assigned a random problem (variable *challenge*), which they would have to solve using a randomly chosen language (variable *language*) and producing a solution of a given size (variable *size*). In this scenario, the randomization would cancel out any indirect correlation between independent variables (in other words, the groups of programmers using different programming languages would be homogeneous in all characteristics except for the used language); thus, the statistical relation between *language* and *rank* (measured, for instance, by a regression model) would measure exclusively the direct causal effect of *language* on *rank*.

Obviously, this is not how Code Jam works. Participants are free to join or leave the contest, can use the programming language they prefer, and do not have any upfront constraints on the size of the program they

<sup>10</sup>A backward process has the same outcome: it suggests  $f_4$  as the “best” model.

<sup>11</sup>“In causal analysis we must incorporate some understanding of the process that produces the data, and then we get something that was not in the data to begin with.” [45, p. 85]

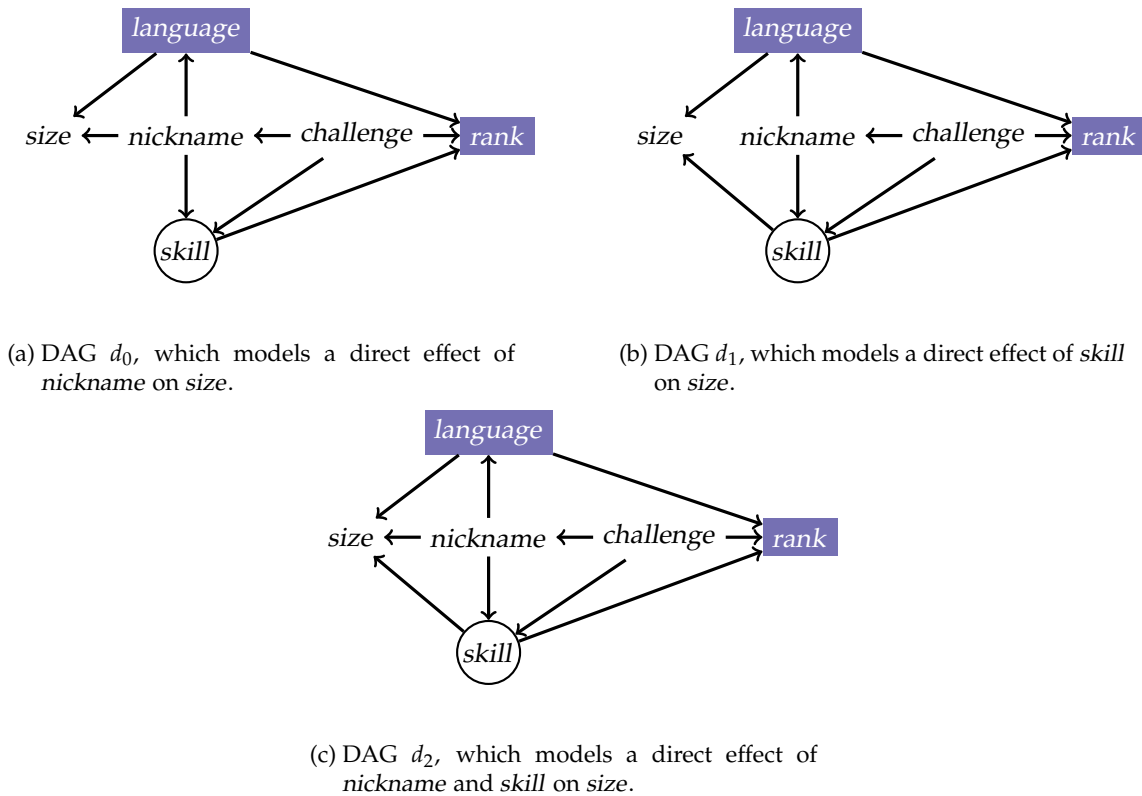


Figure 5: Three DAGs encoding possible causal relations among variables in the Code Jam dataset. Variables *rank* and *language* are colored because their relation is the primary target of the causal analysis. Variable *skill* is circled to denote that it is unobserved, that is not actually measured. All three DAGs have the same nodes (variables) but differ in the arrows (causal relations) that enter *size*.

submit. Therefore, the statistical relation between *language* and *rank* may actually measure an unknown mix of direct causal effect of *language* on *rank* and other confounding effects between independent variables. For example, imagine that all best programmers like to use C++; then, a statistical analysis would report a clear association between submissions in C++ and worse ranks. But this would mostly be an indirect effect of the best programmers' preference for one language over the others; if we asked them to only use Java, they would likely still get top ranks just because of their superior programming skills.

Given that the Code Jam data is observational, and hence such confounding effects are likely to take place, what can we do to isolate any causal effects? While we cannot infer them from the data alone, we can: *i*) design *qualitative causal models* that capture plausible causal relations among variables; *ii*) partially *validate* such models on the data, determining which models are more or less likely to be consistent with the data; *iii*) use a qualitative causal model that is consistent with the data as a guide to build a *statistical model* that properly factors out any confounding effects. If built following these guidelines, the statistical model's association between *language* and *rank* should measure the true causal effect between the two variables—at least, within the uncertainty given by the data, and assuming the qualitative causal model is accurate. The rest of this section goes through these steps for the Code Jam dataset.

## 5.2 Qualitative Causal Models: DAGs

Causal DAGs (Directed Acyclic Graphs) are intuitive graphical models to express qualitative causal relations between variables. Let's demonstrate how they work by designing qualitative causal models for the Code Jam data.

A DAG consists of nodes (also called vertices) and arrows (also called arcs, edges, or links) connecting them. In a causal DAG, *nodes* correspond to *variables*. For the Code Jam data, these are *language*, *rank*, *size*, and *nickname*. A DAG may also include *unobserved* variables, which are in a cause-and-effect relations with other variables, but are not directly available in the dataset. For the Code Jam data, we introduce unobserved

variable<sup>12</sup> *skill* as a placeholder for the programming skills of participants that are relevant for the contest. Figure 5 displays all nodes—observed and unobserved—for the Code Jam contest, marking unobserved variables with a circle, and highlighting the two variables *language* and *rank* whose causal relation our analysis focuses on.

A causal DAG’s *arrows* denote causal relations: an arrow  $A \rightarrow B$  from node  $A$  to node  $B$  means that there is a direct causal effect of  $A$  on  $B$  (in other words, changing  $A$  causes  $B$  to change as well). The model is qualitative in that an arrow’s causal relation may be weak or strong, and is in general probabilistic (but it should be detectable within the precision of the data measurements). For the Code Jam data, we design the DAGs in Figure 5 based on the following observations:

- Measuring the causal relation between *language* and *rank* is the main focus of our analysis; thus, we include an arrow  $language \rightarrow rank$ .<sup>13</sup>
- Different programmers prefer to use some languages over others; thus, we include an arrow  $nickname \rightarrow language$ .
- A program’s size depends also on the programming language it is written in;<sup>14</sup> thus, we include an arrow  $language \rightarrow size$ .
- Programmers differ in their coding skills, which obviously have an impact on their success in the contest; thus, we include arrows  $nickname \rightarrow skill$  and  $skill \rightarrow rank$ .
- Not all challenges in Code Jam have the same intrinsic difficulty; hence, the range of ranks that are assigned varies from challenge to challenge (for example, because fewer participants submit to harder challenges). Thus, we include an arrow  $challenge \rightarrow rank$ .<sup>15</sup>
- A programmer’s skills are usually problem-dependent: different programmers may be more familiar with certain kinds of problems; thus, we include an arrow  $challenge \rightarrow skill$ .
- Similarly, a programmer may selectively decide to submit solutions only for challenges that they find congenial; thus, we include an arrow  $challenge \rightarrow nickname$ .<sup>16</sup>
- Connecting nodes *size* and *rank* would be inconsistent with how the Code Jam contest works: the contest rules (Section 3) do not take a submission’s size into account to rank it (thus, no arrow from *size* to *rank*); and the rank is determined *after* submission, and hence changing a submission’s rank cannot affect the submission’s size (thus, no arrow from *rank* to *size*).

Finally, we may consider different plausible relations between *nickname* and *size*. It’s clear that programmers often have more concise or verbose programming styles, but we can express this relation in a DAG in different ways.

**DAG  $d_0$ .** If we consider a programmer’s conciseness a form of personal preference, we include an arrow  $nickname \rightarrow size$ .

**DAG  $d_1$ .** If we consider conciseness a direct result of a programmer’s skills, we include an arrow  $skill \rightarrow size$ .

**DAG  $d_2$ .** If we consider conciseness a combination of personal preference and skills, we include both arrows  $nickname \rightarrow size$  and  $skill \rightarrow size$ .

Figure 5 shows the resulting DAGs  $d_0$  (Figure 5a),  $d_1$  (Figure 5b), and  $d_2$  (Figure 5c).

<sup>12</sup>Unobserved variables are also called “latent” variables in frequentist terminology.

<sup>13</sup>The main analysis goal is quantitatively estimating the causal effect of *language* on *rank*, but this effect may not exist. Thus, the rest of the analysis builds on the tentative assumption that *language* may directly affect *rank*, and tries to remove sources of bias in the estimate of this causal effect as much as possible given the available information and any causal assumptions. In the end, the estimate itself may confirm a significant effect or indicate that the effect is negligible. Clearly, a causal link in the opposite direction (from *rank* to *language*) is impossible because a submission’s *rank* is determined *after* its *language* has been chosen.

<sup>14</sup>For example, Python programs are usually more concise than Java programs for the same task [38].

<sup>15</sup>Nevertheless, omitting the arrow  $challenge \rightarrow rank$  does not affect the conclusions of the ensuing causal analysis.

<sup>16</sup>This does not mean that a challenge can *change* a participant’s nickname; it means that the value of *challenge* can affect which participants (identified by their nicknames) *appear* in the data for that particular challenge.

### 5.3 Validating DAGs on Empirical Data

Which causal DAG among  $d_0, d_1, d_2$  should we adopt as qualitative causal model of the Code Jam data? We can perform a kind of *validation* where we test whether some of the DAG’s relations are consistent with the data. This validation is necessarily *partial*: as we discussed previously, we cannot derive causal relations from observational data in general (and exhaustively); this limitation also applies to a posteriori validation. However, even partial validation can be useful to choose among different plausible DAGs, and to corroborate the intuition that led us to construct the DAGs in the first place.

To validate a DAG, we first derive its *implied conditional independencies*: these are relations of statistical independence among DAG variables that can be derived from the DAG’s topology. A statistical independence relation is written as  $X \perp\!\!\!\perp Y \mid Z$  and means that variable  $X$  is independent of variable  $Y$  given variable  $Z$ ; intuitively, it captures the idea that there is no meaningful association between  $X$  and  $Y$  after we condition on (know the value of)  $Z$ .

To see an example of conditional independence implied by all DAGs  $d_0, d_1, d_2$ , consider the path  $language \leftarrow nickname \leftarrow challenge$ . Variable  $challenge$  affects  $language$  only indirectly through variable  $nickname$ ; if we condition on  $nickname$  (that is, we single it out explicitly), we should then observe no remaining meaningful correlation between  $language$  and  $challenge$ . Therefore, the three DAGs imply the conditional independence:

$$language \perp\!\!\!\perp challenge \mid nickname \quad (1)$$

Intuitively, this conditional independence is very plausible if we expect that participants tend to use the same programming language in all challenges they take part in;<sup>17</sup> then, after we fix  $nickname$ , we can already reliably predict  $language$ —regardless of  $challenge$ .

More precisely, (1) is the *only* conditional independence implied by DAGs  $d_1$  and  $d_2$ .<sup>18</sup> In contrast, DAG  $d_0$  includes subgraphs  $size \leftarrow nickname \leftarrow challenge$  and  $size \leftarrow nickname \rightarrow language \rightarrow rank$ , but no path  $nickname \rightarrow skill \rightarrow size$ , which imply two additional conditional independencies:

$$size \perp\!\!\!\perp challenge \mid nickname \quad (2)$$

$$rank \perp\!\!\!\perp size \mid nickname, language \quad (3)$$

How do we *test* a DAG’s implied conditional independencies? A simple way uses regression models to measure the association between allegedly independent variables. Given  $X \perp\!\!\!\perp Y \mid Z$ , we build a generalized linear model with  $X$  as outcome, and  $Y$  and  $Z$  as predictors;<sup>19</sup> we fit the model on the data, and check whether  $Y$ ’s fitted parameters are consistently different from zero. If they are, it means that there is a consistent association between  $X$  and  $Y$ , which invalidates the conditional independence; in this case, validation *fails*, as it seems the data exhibits correlations that should be negligible according to the causal DAG. Conversely, if  $Y$ ’s parameters are indistinguishable from zero, validation is *successful*, in that the data agrees with the causal DAG on this aspect. In practice, like every statistical analysis, we should not apply this kind of validation as a strictly binary (yes/no) check, but rather evaluate the parameters’ uncertainty in context and possibly in comparison with other model coefficients.

Figure 6 shows the definitions of models  $mid_1, mid_2$ , and  $mid_3$ , which test implied conditional independencies (1), (2), and (3) respectively. For brevity, we do not discuss choosing suitable priors for these models but only show the likelihoods. All models introduce a parameter  $\alpha$  that captures a population-level mean; this way, all other parameters are differences over  $\alpha$ . Model  $mid_1$  uses a negative binomial probability distribution, since outcome variable  $language$  is an integer identifier and its variance is one order of magnitude bigger than its mean. For similar reasons, model  $mid_3$  also uses a negative binomial distribution. Model  $mid_2$  uses instead a normal distribution, since its outcome variable varies on a continuous scale. Then, parameters  $\alpha_{challenge}, \alpha_{nickname}$ , and  $\alpha_{language}$  are index variables—one for each challenge, nickname, and language—used as intercepts; parameter  $\beta$  is a slope associated with the (logarithm of) a submission’s size.

Figure 7 plots the 50% probability intervals<sup>20</sup> of parameters  $\alpha_c$ , for each challenge  $c$ , in model  $mid_1$  (in brown color) and in model  $mid_2$  (in gray color). Only 2% of all 51 coefficients  $\alpha_c$  are consistently different from zero in  $mid_1$ ; in contrast, 71% of them are in  $mid_2$ . As for  $mid_3$ , coefficient  $\beta$ ’s 50% probability interval is  $(-1.02, -0.95)$ , which is quite clearly negative. This analysis suggests that:

<sup>17</sup>In our dataset, more than 80% of all participants used the same language in all their Code Jam submissions.

<sup>18</sup>More precisely, it is the only *testable* conditional independence: conditional independencies that involve unobserved variables (such as  $skill$ ) cannot be tested since they are not available in the data.

<sup>19</sup>Since  $X \perp\!\!\!\perp Y \mid Z$  is equivalent to  $Y \perp\!\!\!\perp X \mid Z$ , we could equivalently use  $Y$  as outcome, and  $X$  and  $Z$  as predictors.

<sup>20</sup>Analyzing higher probability intervals leads to similar overall conclusions.



$$\begin{aligned} \text{language}_i &\sim \text{NegativeBinomial}(\lambda_i, \phi) \\ \log(\lambda_i) &= \alpha + \alpha_{\text{challenge}[i]} + \alpha_{\text{nickname}[i]} \end{aligned}$$

$$\begin{aligned} \log(\text{size}_i) &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \alpha_{\text{challenge}[i]} + \alpha_{\text{nickname}[i]} \end{aligned}$$

(a) Likelihood of model  $mid_1$  for testing conditional independence (1).

(b) Likelihood of model  $mid_2$  for testing conditional independence (2).

$$\begin{aligned} \text{rank}_i &\sim \text{NegativeBinomial}(\lambda_i, \phi) \\ \log(\lambda_i) &= \alpha + \beta \cdot \log(\text{size}_i) + \alpha_{\text{nickname}[i]} + \alpha_{\text{language}[i]} \end{aligned}$$

(c) Likelihood of model  $mid_3$  for testing conditional independence (3).

Figure 6: Definitions of regression models  $mid_1$ ,  $mid_2$ , and  $mid_3$  for testing the implied conditional independencies of DAGs  $d_0$ ,  $d_1$ , and  $d_2$ .

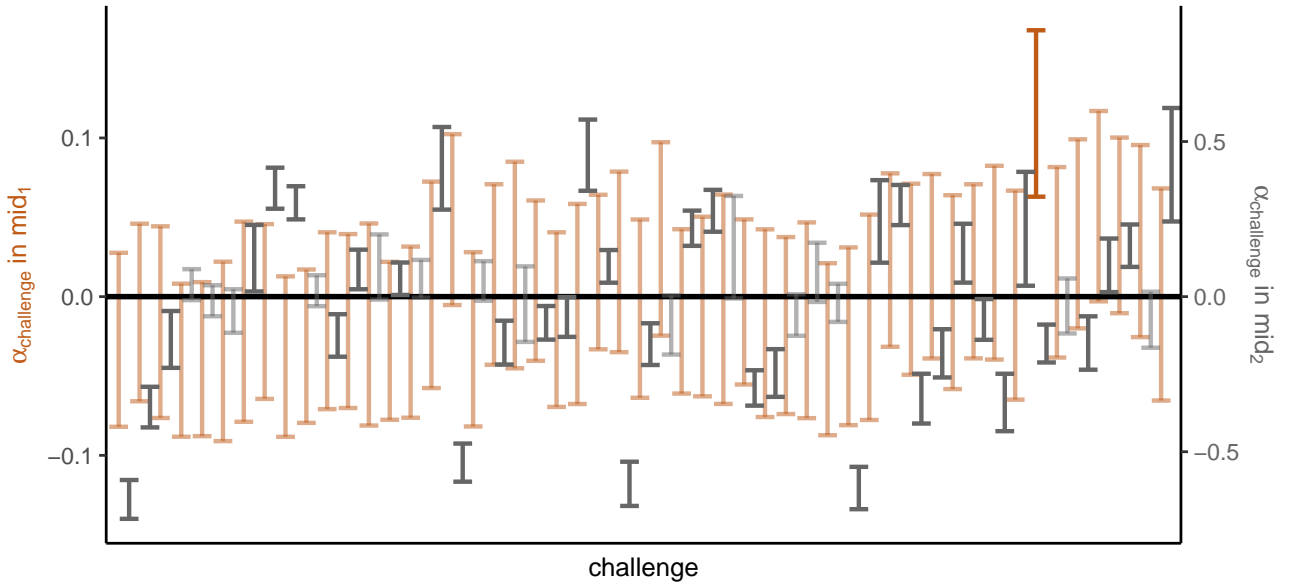


Figure 7: For each Code Jam challenge  $c$ , the 50% probability intervals of the estimates of parameters  $\alpha_c$  in models  $mid_1$  (brown, left scale) and  $mid_2$  (gray, right scale). Intervals that do not include zero are drawn in a darker color.

1. Variables *challenge* and *language* are negligibly associated in  $mid_1$ ; thus, the data confirms the conditional independence (1).
2. Variables *challenge* and *size* are consistently associated in  $mid_2$ ; thus, the data invalidates the conditional independence (2).
3. Variables *size* and *rank* are consistently associated in  $mid_3$ ; thus, the data invalidates the conditional independence (3).

We conclude that DAGs  $d_1$  and  $d_2$ , which imply the same conditional independence (1), pass validation, whereas DAG  $d_0$  fails it. Thus, we use DAGs  $d_1$  or  $d_2$  as the basis of our causal analysis of the effects of programming languages.

## 5.4 Choosing Predictors to Measure Causality

After validating it, a causal DAG can guide the choice of which predictors to include and which to exclude in a regression model in order to estimate the direct causal effects of one variable on another one. We'll demonstrate this on DAG  $d_2$  in Figure 5; using DAG  $d_1$  leads to the same conclusions.

### 5.4.1 Causal and Non-Causal Paths

A (simple) *path* on a DAG is a sequence of nodes connected by arrows (ignoring the arrow direction), where each node appears at most once. There are 14 paths in DAG  $d_2$  that go from *language* to *rank*:

- A.  $language \rightarrow rank$
- B.  $language \rightarrow size \leftarrow nickname \rightarrow skill \rightarrow rank$
- C.  $language \rightarrow size \leftarrow nickname \rightarrow skill \leftarrow challenge \rightarrow rank$
- D.  $language \rightarrow size \leftarrow nickname \leftarrow challenge \rightarrow rank$
- E.  $language \rightarrow size \leftarrow nickname \leftarrow challenge \rightarrow skill \rightarrow rank$
- F.  $language \rightarrow size \leftarrow skill \rightarrow rank$
- G.  $language \rightarrow size \leftarrow skill \leftarrow challenge \rightarrow rank$
- H.  $language \rightarrow size \leftarrow skill \leftarrow nickname \leftarrow challenge \rightarrow rank$
- I.  $language \leftarrow nickname \rightarrow size \leftarrow skill \rightarrow rank$
- J.  $language \leftarrow nickname \rightarrow size \leftarrow skill \leftarrow challenge \rightarrow rank$
- K.  $language \leftarrow nickname \rightarrow skill \rightarrow rank$
- L.  $language \leftarrow nickname \rightarrow skill \leftarrow challenge \rightarrow rank$
- M.  $language \leftarrow nickname \leftarrow challenge \rightarrow rank$
- N.  $language \leftarrow nickname \leftarrow challenge \rightarrow skill \rightarrow rank$

Since each pair of consecutive nodes in a path are correlated variables, a path from  $X$  to  $Y$  represents one contribution to the overall correlation that is observed between  $X$  and  $Y$ . Some paths correspond to *causal* correlations between  $X$  and  $Y$ ; others represent *spurious* correlations that do not correspond to any causal effects. For example,  $d_2$ 's path *A* is clearly causal; in contrast, path *M* is not, since the resulting correlation between *language* and *rank* is due to the effect of *challenge* on both *rank* and *language*: changing *challenge* simultaneously affects *rank* and *language*, and a spurious correlation emerges. In this scenario, *challenge* is called a *confounder*, as it biases the net causal effect of *language* on *rank*. More generally, every path from  $X$  to  $Y$  with an arrow entering  $X$  is a so-called *backdoor* path that biases the estimate of the true causal effect of  $X$  on  $Y$ . In  $d_2$ , paths *I* through *N* are backdoor paths.

### 5.4.2 Open and Closed Paths

Our goal is removing all backdoor paths, so that the correlation we observe between  $X$  and  $Y$  is *unbiased*—that is, it is exclusively due to causal effects. Obviously, we cannot do that by simply removing arrows from a DAG: the arrows are supposed to represent direct causal relations in the process that determined the data we are analyzing. Instead, we can select which variables (nodes) to include (“control for”) in a regression model that estimates the correlation between  $X$  and  $Y$ .

Intuitively, we select variables to use as predictors so as to *close* all backdoor paths. Closing a path means that the path's spurious correlation cancels out in the estimate of the regression model. Consider again path *M* in  $d_2$ ; if we include *nickname* among the regression variables, the path becomes closed, and it will not contribute any spurious correlation between *language* and *rank*. In fact, including a variable in a regression model means *conditioning* on the variable. Conditioning on *nickname* means that the model can specifically estimate the correlation of *nickname* on the other variables, which blocks the spurious correlation between *language* and *rank* otherwise induced by the path through *nickname*.

Unfortunately, adding variables as predictors may also backfire, as including some variables can *open* backdoor paths that would otherwise be closed. For example, take path *I* in  $d_2$ : as long as we do *not* include *size* among the regression variables, the path remains closed, and it will not contribute any spurious correlation between *language* and *rank*; but conditioning on *size* “opens the path” by introducing a correlation from *language* that goes all the way through the path to *rank*.

Here is an intuitive explanation of why conditioning on *size* may *bias* the estimate of the effect of *language* on *rank*. A submission's *size* is affected by its *language* (some languages are generally more concise [47;

38]) and by the participant’s *skill* (more skilled programmers presumably can produce more concise code). This implies that there are two main ways for a submission to be concise: either it’s written by a skilled programmer, or it’s written using a concise language. In other words, if we fix a submission’s *size*, it would appear as if *language* and *skill* correlate: thanks to their higher *skill*, the better programmers using verbose languages would still be able to produce submissions of *size* comparable to those of the worse programmers (lower *skill*) using concise languages. This correlation, however, is merely a result of self selection, and does not reflect any genuine causal relation between *skill* and *language*.<sup>21</sup> On top of this, *skill* and *language* both directly affect *rank*; therefore, the spurious correlation introduced by conditioning on *size* conflates these two effects—ultimately biasing our estimate of the true causal effect of *language* on *rank*.

Systematically, a path is *open* unless it includes a *collider*: a node  $Z$  where both arrows connecting it to the rest of the path enter it (as in  $X \cdots \rightarrow Z \leftarrow \cdots Y$ ). In the previous example of path  $I$ , *size* is a collider, which we should *exclude* from the regression variables lest backdoor path  $I$  becomes open.

### 5.4.3 Adjustment Sets

Finally, we can formulate a general recipe to select which variables to include as regression variables in a model to estimate the genuine causal effect of  $X$  on  $Y$  given a DAG capturing the causal relations among variables [8]. Consider all *backdoor* paths from  $X$  to  $Y$  in the DAG; for each open backdoor path, pick a variable  $Z$  within the path such that conditioning on  $Z$  would close the path; while doing so, make sure that you do not select any variable that is also a collider in some closed backdoor path. A set of variables that satisfy these constraints is called an *adjustment set*; using them as regression variables would close all open backdoor paths without opening any closed backdoor paths. Thus, a regression model that conditions on  $X$  as well as on all variables in an adjustment set provides an *unbiased* estimate of the correlation between  $X$  and  $Y$  that is solely due to the causal link between them.

The constraints on which variables to include or exclude from an adjustment set may be unsatisfiable—if conditioning on a collider is needed to close another backdoor path. Conversely, multiple valid adjustment sets may exist for the same DAG.

Adding predictors to a regression model is a common practice to address so-called “omitted variable bias”, which occurs when a causal effect’s estimate is biased because we did not correct for possible confounders. As we are demonstrating, causal DAGs are useful to rigorously find confounders and filter them out from the causal estimate. The flip side is that there may also be a risk of “included variable bias”: adding the wrong predictor to a regression may *introduce* a confounding effect and spoil the estimate of a causal relation. DAGs are also useful to detect and address this kind of confounding that occurs when we include variables that should not be included.

### 5.4.4 Unbiased Estimation for Code Jam Data

Analyzing the paths from *language* to *rank* in Figure 5’s DAG  $d_2$  indicates two adjustment sets<sup>22</sup> to accurately estimate the causal effect of *language* on *rank*:  $A_1 = \{\textit{nickname}\}$  and  $A_2 = \{\textit{nickname}, \textit{challenge}\}$ .  $A_1$  is the *minimal* adjustment set, as conditioning on *nickname* is required to close any backdoors.  $A_2$  is an alternative adjustment set, which indicates that conditioning on *challenge* is neither needed nor harmful to filter out spurious association. In contrast, *size* does not appear in either adjustment sets, which means that we should *not* condition on it.

This analysis indicates that models  $m_2$  and  $m_3$  in Figure 3 are suitable to reliably estimate the causal effect of *language* on *rank*. In contrast, we should not use models  $m_1$  or  $m_4$  for this: model  $m_1$  uses only *language* as predictor (and none of the adjustment variables), whereas model  $m_4$  also uses *size* as predictor (which adds a confounding effect). This conclusion holds quite robustly independent of the details of the causal DAG that we assume; in particular, DAGs  $d_1$  and even  $d_0$  have the same adjustment sets as  $d_2$ .

While including *challenge* as predictor is neutral as far as removing confounding bias is concerned, it may help improve the *precision* of the estimate of the effect of *language* on outcome *rank*. As a rule of thumb, adding predictors “close” to the outcome helps reduce variance, since it may filter out the effect of other unknown dependencies or non-linearities among variables [24]. Since model  $m_3$  also outperforms model  $m_2$  in terms of predictive accuracy (see Table 1), we will use the former as the basis of our causal analysis.

<sup>21</sup>Here is an analogous example of selection bias that is perhaps more intuitive to understand [25; 58]. An actor’s *success* is mainly determined by their acting *talent* and by their good *looks* ( $\textit{talent} \rightarrow \textit{success} \leftarrow \textit{looks}$ ). Thus, among successful actors, the more talented ones will be worse looking, and the less talented ones will be better looking. This inverse relation between *talent* and *looks* is spurious though: an actor’s talent and good looks are presumably independent, but conditioning on *success* measures a correlation that is simply a figment of the way in which the data is analyzed.

<sup>22</sup>There exist several other adjustment sets, but they all include *skill*, which is unobserved, and hence cannot be used as predictor.

## 5.5 Analysis: Programming Language Effects

The qualitative causal analysis based on DAGs indicates that we can use model  $m_2$  or  $m_3$ , but should prefer  $m_3$ , to estimate the causal effect of *language* on *rank*. Equipped with this knowledge, we can revisit the analysis of Section 4.3 by inspecting model  $m_3$ 's 50% highest-posterior density intervals of the distribution of  $\delta_\ell$  (the difference between language  $\ell$ 's contribution and the average language contribution) for each language  $\ell$ . Figure 4 displays these intervals.

**Results: model  $m_3$ .** According to model  $m_3$  (which causal analysis identified as the one capturing most accurately the direct causal effects of languages on Code Jam results), at the 50% probability level:

- C++ is associated with smaller-than-average rank ordinals, that is *better* contest results;
- Python is associated with larger-than-average rank ordinals, that is *worse* contest results;
- there is no consistent association for Java, although its 50% interval mostly covers better-than-average ranks.

Based on this, we can answer the paper's main RQ under the causal interpretation:

**AQ (causal):** For experienced participants to the Code Jam contest, using C++ leads to better contest results, using Python leads to worse contest results, and using Java has no consistent effect (but leans towards better results).

The answer to the paper's RQ has changed completely after shifting our focus on causal effects, as opposed to purely correlational associations!

## 5.6 Correlation vs. Causation

The difference between the correlational and causal answers to the question of the relation between programming languages and Code Jam contest results is striking, as they are nearly each other's mirror image.

The contrast is amplified by other features of the analysis. First, there is a remarkable agreement between the two "causally consistent" models  $m_2$  and  $m_3$ , whose predictions are essentially the opposite of model  $m_4$ . Second, model  $m_4$  obliterates the other models in terms of predictive capabilities (see Section 4.2); thus, there is no reason to use models  $m_3$  or  $m_2$  according to purely predictive considerations. Third, even though the simple model  $m_1$  happens to agree with models  $m_2$  and  $m_3$ 's predictions despite not controlling for confounders, it is dead last in Section 4.2's model comparison; thus, once again, model  $m_4$  is indisputably the one to prefer from a purely predictive point of view.

This confirms the intuition that the Code Jam contest is not run as a randomized controlled experiment. Therefore, not all associations that we observe in the data reflect actual causal effects. If we want to estimate the latter, as opposed to the former, we have to carefully select which information to include and which to exclude from our statistical model.

## 6 The Modest Impact of Languages

How can we explain the striking difference between the correlational and causal analyses of the same Code Jam data? In other words, how is it possible that adding a single predictor to model  $m_3$  leads to nearly opposite predictions about the relative impact of different programming languages?

In order to shed some light on the matter, we break down model  $m_4$  to compare the relative effect of each predictor variable it uses. We consider  $m_4$  because it is the one that delivers the most reliable predictions according to the model comparison of Section 4.2. Furthermore, it includes all available data, so that we can compare *all* variables to each other.

The following discussion hinges on Figure 8, which plots the estimates and 50% probability intervals of the effects of variables *nickname*, *challenge*, *language*, and *size* in model  $m_4$ . As detailed in Section 6.1, the effects are centered so that we can compare their absolute values on a consistent scale. Then, Section 6.2 and Section 6.3 analyze the plot in detail to appreciate the different relative impact of each variable on a submission's *rank*.

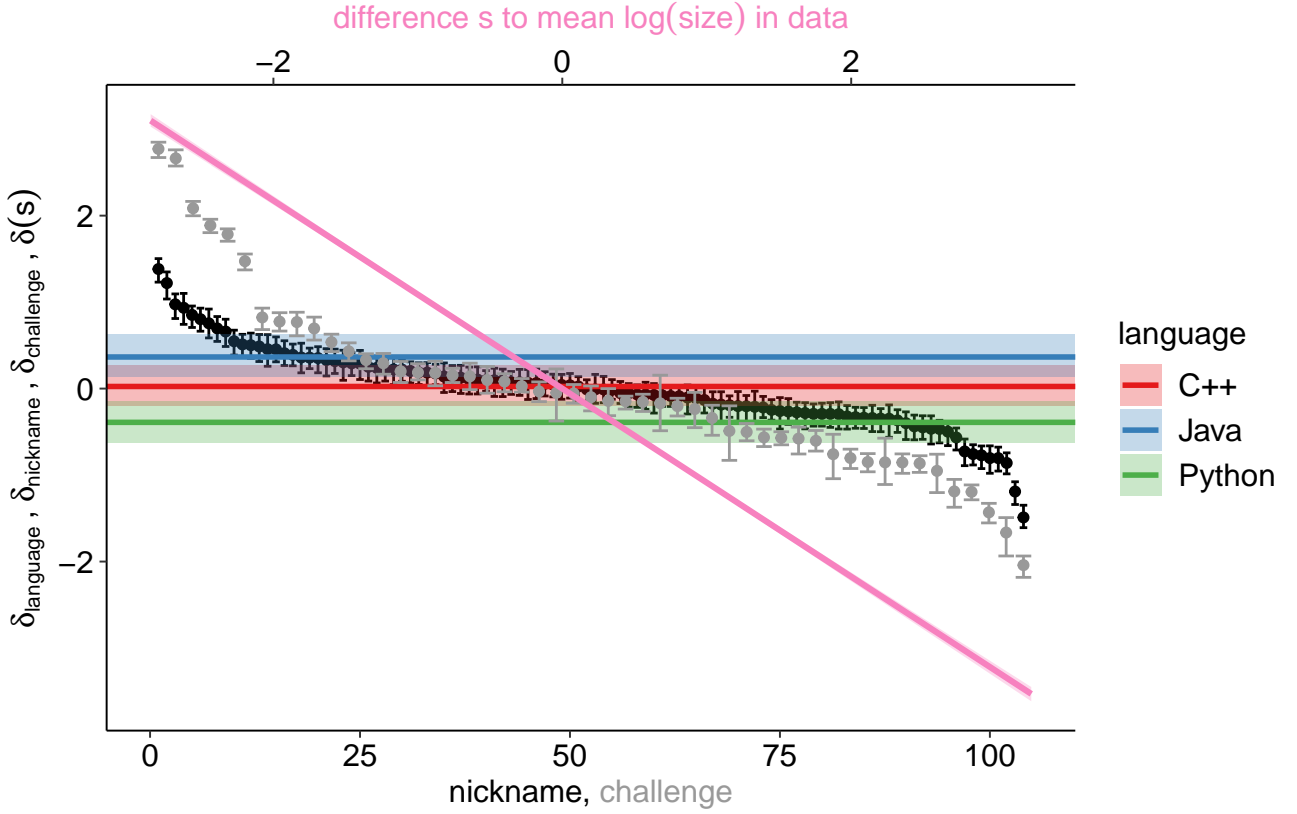


Figure 8: Estimates and 50% probability intervals of the centered effects:  $\delta_\ell$ , for every language  $\ell$ ;  $\delta_n$ , for every nickname  $n$ ;  $\delta_c$ , for every challenge  $c$ ; and  $\delta(s)$  for every difference  $s$  to mean  $\log(\text{size})$ .

### 6.1 Centered Effects in Model $m_4$

As shown in Figure 3a, model  $m_4$ 's main term is the sum of four components, each capturing the contribution of a different predictor among *language*, *nickname*, *challenge*, and *size*. Just like we did for *language* in the main analysis of the paper, we introduce a derived variable  $\delta_x$  for each nickname, challenge, and size; in a nutshell,  $\delta_x$  is a *centered* version of  $\alpha_x$ , which measures the differential contribution of  $x$  relative to the average contribution of other  $x$ s in the same group.

$$\begin{aligned} \delta_\ell &= \alpha_\ell - \frac{\mathbb{E}[\sum_{x \in L} \alpha_x]}{|L|} && \text{for every language } \ell \in L \\ \delta_n &= \alpha_n - \frac{\mathbb{E}[\sum_{x \in N} \alpha_x]}{|N|} && \text{for every nickname } n \in N \\ \delta_c &= \alpha_c - \frac{\mathbb{E}[\sum_{x \in C} \alpha_x]}{|C|} && \text{for every challenge } c \in C \\ \delta(s) &= \beta \cdot s && \text{for any value } s \text{ of } \log(\text{size}) \end{aligned}$$

Figure 9: Definitions of centered random variables  $\delta_\ell$ ,  $\delta_n$ ,  $\delta_c$ , and  $\delta(s)$ .  $\mathbb{E}[x]$  denotes the expected value (mean) of  $x$ ;  $L$ ,  $N$ , and  $C$  are the sets of all languages, nicknames, and challenges, respectively.

More precisely, consider the definitions in Figure 9. Each  $\delta_\ell$ ,  $\delta_n$ ,  $\delta_c$ , and  $\delta(s)$  is a random variable derived from one of  $m_4$ 's parameters. For example, for any language  $\ell$  from the set  $L = \{\text{C++}, \text{Java}, \text{Python}\}$  of all languages,  $\delta_\ell$  is the difference between  $\alpha_\ell$  and the mean (expected value) of all language-specific intercepts  $\alpha_{\text{C++}}$ ,  $\alpha_{\text{Java}}$ , and  $\alpha_{\text{Python}}$ . Since  $\alpha_\ell$  is a random variable, whereas the mean is a constant,  $\delta_\ell$  is itself a (shifted) random variable, with its own distribution that we can summarize with the usual statistics (mean, probability intervals, and so on). In Bayesian analysis,  $\delta_\ell$ 's distribution is approximated by a collection of *posterior samples*, on which we measure the statistics.

Similarly,  $\delta_n$  is the centered version of  $\alpha_n$  for any nickname  $n$  among all nicknames  $N$  in the Code Jam dataset; and  $\delta_c$  is the centered version of  $\alpha_c$  for any challenge  $c$  among all challenges  $C$  in the dataset. The definition of  $\delta(s)$  is analogous but for the slope  $\beta$  in model  $m_4$ , which multiplies the logarithm of continuous variable *size*. Consistently with the other centered variables, we want  $\delta(s)$  to measure *difference* over the “average” contribution associated with *size*; to this end,  $s$  ranges over differences  $x - \overline{\log(\text{size})}$ , where  $\overline{\log(\text{size})}$  is the average (mean) logarithmic size of a submission in the Code Jam dataset. For example,  $s = 0$  denotes a submission whose logarithmic size is the same as the average submission size.

With these definitions in place, we can plot the estimates, as well as the 50% probability intervals of  $\delta_\ell$ ,  $\delta_n$ ,  $\delta_c$ , and  $\delta(s)$  on the same figure since they all range over the same scale of “differences on top of the overall population mean”. This also means that we can meaningfully *compare* the relative contributions of each term; according to the model’s predictions, those that are larger have a larger overall effect on the *rank* of a submission, and hence are the main determinants of the overall results. The resulting plot is shown in Figure 8.

## 6.2 Comparison of Centered Effects

First, look at the contribution of *language*, which Figure 8 pictures by three colored lines with a shaded band, one per language  $\ell$ , that represent the estimate of  $\delta_\ell$  and the 50% probability interval of that estimate. As you can see, *language*’s effects span a narrow range, roughly centered around zero, and visibly overlap. This means that the overall impact, on the predicted results, of using one or the other programming language is modest and hardly definitive. The range between the upper endpoint of Java’s 50% probability interval and the lower endpoint of Python’s is just 1.26; this translates to  $\exp(1.26) \simeq 3.5$  on *rank*’s outcome scale,<sup>23</sup> which means that using one language over another is usually associated with only a few position’s differences in rank.

Now, look at the contribution of *nickname*, which Figure 8 pictures as black dots and intervals (for mean and 50% probability interval of the estimate); for readability, all values  $n$  of *nickname* are sorted by decreasing values of  $\delta_n$  and given a progressive numerical identifier (on the bottom horizontal axis). As you can see, *nickname*’s effects span a wider range than the language ones: the difference between the upper endpoint of the leftmost interval and the lower endpoint of the rightmost interval is 3.1, or 22.4 on the outcome scale. This value is one order of magnitude larger than the corresponding value for *language*; thus, the overall difference in contributions between different participants is more conspicuous and consequential. In other words, if we fix a participant and change the language<sup>24</sup> they use from Python to Java, we still can only expect a modest change in rank; but if we let the best and worst participant use even the least effective language, their results in the contest will remain substantially different.

We now repeat the analysis for *challenge*, whose contribution Figure 8 pictures as gray dots and intervals (for mean and 50% probability interval of the estimate); as for *nickname*, all values  $c$  of *challenge* are sorted by decreasing values of  $\delta_c$  and given a progressive numerical identifier (on the bottom horizontal axis). The intervals associated with the different *challenges* span a wider range yet: the difference between the upper endpoint of the leftmost interval and the lower endpoint of the rightmost interval is 5.0, or 153.2 on the outcome scale.

Finally, the pink line in Figure 8 diagrams the function  $y = \delta(x)$ , where  $x$  measures a difference between a submission’s logarithmic *size* and the mean logarithmic size of submissions to Code Jam. Around the line is a shaded ribbon, corresponding to the 50% probability interval of the estimate of  $\beta$ ; the ribbon is barely visible because it is narrow, as there is little uncertainty in the estimate. The horizontal range on the top horizontal axis spans the actual range of differences to mean logarithmic size that we observed in the Code Jam dataset. Within this range, *size*’s effects on *rank* are the biggest yet: the difference between the vertical coordinates of the top-left and bottom-right ends of the pink line is 6.6, or 752.7 on the outcome scale. In part, the large effect attributed to *size* is a result of measuring the *logarithm* of a submission’s size. A submission that is larger than another submission by  $s$  units on this scale is actually  $\exp(s)$  larger in bytes; thus, the effects of *size* emerge only when we observe substantial differences in actual size.

The relatively large effect of *size* also demonstrates a spurious (i.e., non-causal) correlation: clearly, a submission’s size cannot be a direct cause of the submission’s rank, since Code Jam rules do not take size into account. Similarly, if we take a poorly-ranked submission and artificially increase its size (for example, by adding dead code), its rank won’t improve. Therefore, *size* is an effective predictor (probably because it

<sup>23</sup>Taking the exponential inverts the logarithmic link function of model  $m_4$ ’s linear relation.

<sup>24</sup>This is a statistical prediction of the model; in reality, not all participants may be fluent in different languages or comfortable switching.

| INDEX | YEARS | ROUNDS | DATAPOINTS | PARTICIPANTS |
|-------|-------|--------|------------|--------------|
| 1     | 7     | 6      | 721        | 27           |
| 2     | 6     | 6      | 960        | 39           |
| 3     | 5     | 6      | 1230       | 53           |
| 4     | 4     | 6      | 1564       | 75           |
| 5     | 3     | 6      | 1807       | 95           |
| 6     | 2     | 6      | 1896       | 105          |
| 7     | 1     | 6      | 2010       | 124          |
| 8     | 7     | 5      | 2762       | 115          |
| 9     | 7     | 4      | 4889       | 221          |
| 10    | 6     | 5      | 5106       | 233          |
| 11    | 7     | 3      | 5707       | 272          |
| 12    | 7     | 1      | 5776       | 279          |
| 13    | 7     | 2      | 5776       | 279          |
| 14    | 5     | 5      | 7359       | 366          |
| 15    | 4     | 5      | 9667       | 536          |
| 16    | 6     | 4      | 11242      | 576          |
| 17    | 3     | 5      | 11814      | 739          |
| 18    | 2     | 5      | 13478      | 957          |
| 19    | 6     | 3      | 13652      | 749          |
| 20    | 6     | 2      | 13884      | 775          |
| 21    | 6     | 1      | 13908      | 779          |

Table 3: Different way of selecting the Code Jam data: each row identifies a dataset with DATAPOINTS about submissions made by a number of PARTICIPANTS. These are the participants who entered at least YEARS yearly edition of the Code Jam contest, at least ROUNDS consecutive rounds in any one year, and who used the C++, Java, or Python programming languages. The rows are sorted by number of datapoints and numbered with an ordinal INDEX of size. The paper’s main analyses target the dataset with INDEX 6.

serves as a proxy for the expected effort required to meet a submission’s requirements), but does not immediately suggest any practical strategy to improve success at Code Jam contests.

### 6.3 Language vs. Other Effects

Remember that model  $m_4$  makes better predictions than the other models. Thus, the above analysis suggests that knowing a submission’s *language* sways only modestly the prediction of the submission’s *rank*. In contrast, knowing who wrote the submission (*nickname*), how large the submission is (*size*), and which specific *challenge* the submission is for, all affect predictions more strongly.

In fact, this is not a feature of model  $m_4$  specifically: a plot similar to the one in Figure 8 but for  $m_3$ <sup>25</sup> would look very similar—except for the absence of *size*, which is not used in  $m_3$ —and confirm that the language-specific effect is modest compared to the other predictors’ effects.

One takeaway message is simply that programming language effects—whatever they are—are *modest* compared to others. Thus, they may be tricky to detect reliably, as they are easily confounded by other dominant factors. As we discuss in Section 2, this observation is consistent with several other empirical studies investigating the impact of programming languages. Another way of looking at this is that, since the effects of programming languages are easily confounded, it is very hard for an empirical study to collect all the necessary data in an accurate enough way; in other words, studies are easily underpowered [19].

How do these findings relate to the main focus of the paper—demonstrating how causal analysis techniques can help improve how we answer software engineering questions empirically? In a way, this section’s analysis helps validate the causal analysis, in that it confirms that the impact of programming languages is limited in our (as in other studies’) data, and hence not taking causality into account can easily distort their actual effects. Conversely, explicitly considering and modeling causal effects was instrumental in obtaining a precise understanding of the role of programming languages and their relations to other factors.

## 7 Robustness and Threats to Validity

Before discussing (in Section 7.2) any threats to the validity of our results, we outline (in Section 7.1) some additional analysis that we performed to understand the robustness and generalizability of our main results.

<sup>25</sup>This plot is available in the replication package.

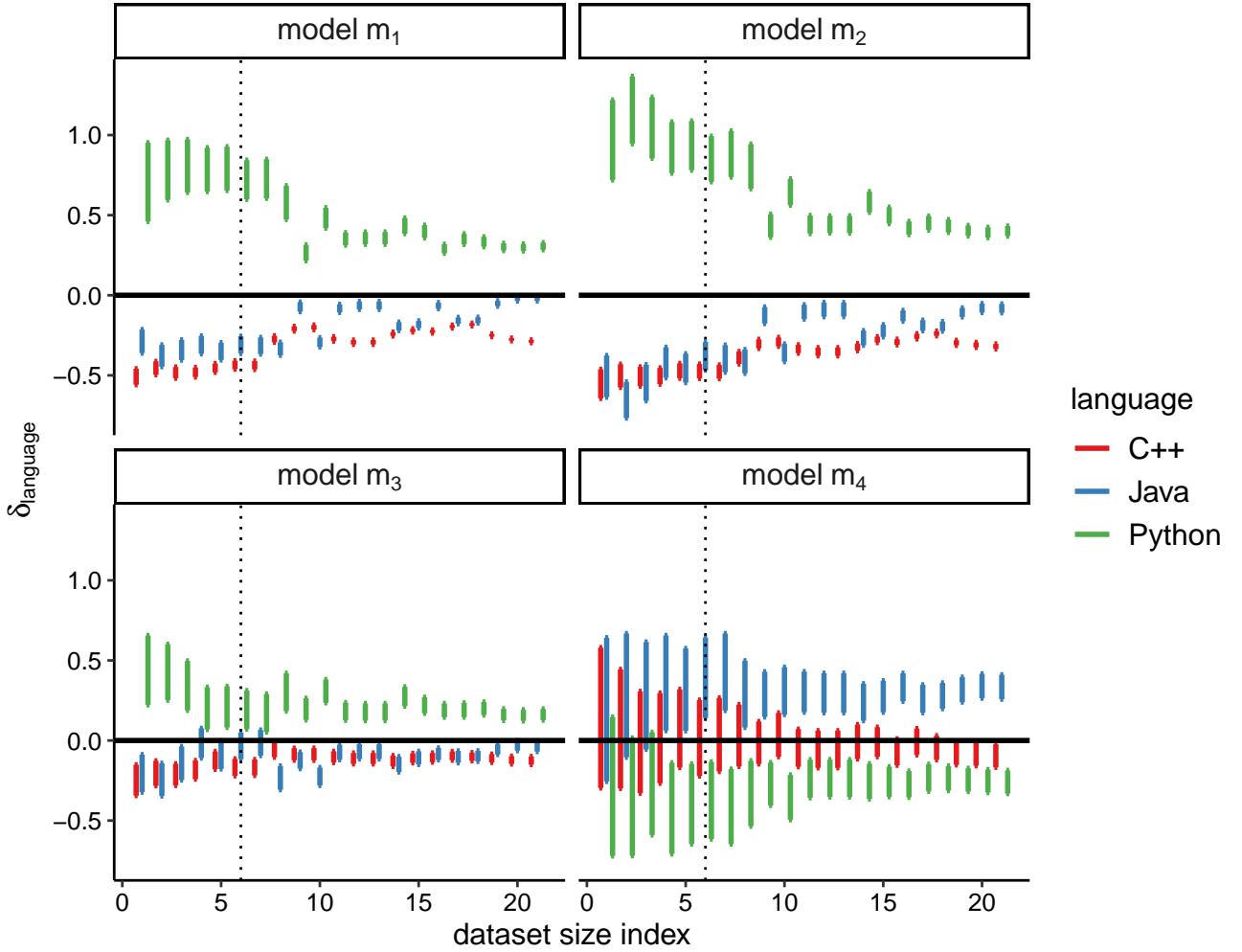


Figure 10: For each language  $\ell$ , in each model  $m_1, m_2, m_3, m_4$ , the 50% highest posterior density intervals of the distribution of  $\delta_\ell$  as we select larger datasets from all Code Jam data. The “dataset size index” corresponds to column INDEX in Table 3: the larger, the more participants and datapoints are included in the dataset. The dotted lines mark the dataset used in the main paper analyses.

## 7.1 Analysis Robustness

In Section 3.3 we explained how we selected a subset of all available Code Jam data to compile a dataset that is of reasonable size (neither unrealistically small nor impractically large) and captures the performance of *experienced* participants (as opposed to any casual participants). To this end, we only considered data of programmers who: *i*) took part in at least two yearly editions of Code Jam; *ii*) in any one edition, entered at least six rounds. In addition, we only considered submissions using the most popular programming languages C++, Java, and Python. Do the main results of our analysis change if we select a subset of data differently?

To answer this question, we consider different subsets of the Code Jam data built as follows. A pair  $Y, R$  identifies the subset of the Code Jam data including all submissions using languages C++, Java, and Python by programmers who: *i*) took part in at least  $Y$  yearly editions of Code Jam; *ii*) in any one edition, entered at least  $R$  rounds. Table 3 lists 21 datasets, each corresponding to a pair  $Y, R = \text{YEARS}, \text{ROUNDS}$ —where  $1 \leq Y \leq 7$  (since our data spans 7 editions of Code Jam) and  $1 \leq R \leq 6$  (since no edition has more than 6 rounds). More precisely, Table 3 enumerates the first 21 datasets in increasing order of size (column DATAPOINTS). We use column INDEX to identify each dataset according to its size; the paper’s main analysis thus corresponds to the dataset with index 6.

We fitted our four models on each of these 21 datasets. Figure 10 pictures the 50% probability intervals of the language-specific  $\delta_\ell$ . While the intervals change in size and position relative to zero as we include more submissions, the high-level picture remains generally consistent. In particular, in the “causal” models  $m_1, m_2, m_3$ , C++ and Java are usually associated with negative  $\delta_\ell$  (better-than-average ranks), whereas Python



| ANALYTICS SMELL                     | DISCUSSED IN | ADDRESSED IN |
|-------------------------------------|--------------|--------------|
| Using deprecated data               | 7.2.1        | 7.1          |
| Not interesting                     | 7.2.1        | 2.3          |
| Not exploring simplicity            | 7.2.2        | 4            |
| Assumptions in statistical analysis | 7.2.2        | 4            |
| $p < 0.05$ and all that!            | 7.2.2        | 4            |
| Not exploring stability             | 7.2.3        | 7.1          |
| Inadequate reporting                | 7.2.4        | [15]         |
| No data visualization               | 7.2.4        | 4, 5, 6, 7.1 |
| Not using related work              | 7.2.4        | 2            |

Table 4: A summary of the analytics smells [35] relevant for the paper. For each SMELL, the subsection where it is DISCUSSED, and the section(s) or reference where it is ADDRESSED.

is usually associated with positive  $\delta_\ell$  (worse-than-average ranks); the situation is opposite in model  $m_4$ , where Python is usually associated with negative  $\delta_\ell$ , Java is usually associated with positive  $\delta_\ell$ , and C++ usually has no consistent associations.

There are a few exceptions to these trends: with the smallest datasets, there is more uncertainty (i.e., the intervals are wider) because fewer data is used to fit the models; with the largest datasets, some associations shift because more heterogeneous participants are included. However, by and large, the contrast in predictions between models  $m_1, m_2, m_3$  and, on the other hand, model  $m_4$  remains.

Similarly to Section 6’s analysis, this section’s robustness analysis further validates the causal analysis by confirming that its results are not merely a product of how the data was selected. While we provided a justification for selecting the original dataset—based on the intuitive notion of “experienced” participant—ours was the first detailed analysis of Code Jam data, and hence validating our data selection process was important. In contrast, other studies, targeting more widely used and accepted curated datasets, may forgo such an exploration of the analysis robustness without losing much in terms of generalizability.

## 7.2 Threats to Validity and Analytics Smells

We scrutinize several aspects of the paper’s analysis, and connect them both to the traditional “threats to validity” categories (internal, external, construct, and conclusion) and to Menzies and Shepperd’s “Bad smells” in analytics [35]. Table 4 gives an overview of the analytics smells that are covered in this section.

### 7.2.1 Data

The operationalizations in the Code Jam data are generally straightforward, as they consist of standard attributes (participant’s nickname, size, and so on) that should be unproblematic to obtain. In particular, variable *rank* is computed automatically by the Code Jam organizers according to predefined rules.

Determining a submission’s programming language may be tricky in general, but should not be an issue for Code Jam data, since ranking a submission requires to run it on the contest’s predefined environment, and hence it must use one of the available language compilers. Besides, we only used the three most widely used programming languages, which helps avoid odd corner cases.

We mostly used the data as is, except for aggregating multiple submissions by participant in a round, and for taking the logarithm of the size. The former is consistent with how the rank is determined by *all* participant submissions; the latter is customary when a quantity spans several order of magnitude.

The above observations indicate how threats to *construct validity* are mitigated.

We could not use *all* available Code Jam data both because it is impractically too much, and because it is too heterogeneous in terms of participants’ skills. Section 3 explains how we aimed for a reasonable selection that should be representative of “experienced” participants; but we also performed additional analysis on larger and smaller data selections to investigate robustness of the main results (Section 7.1). While this selection might somewhat restrict the validity of our results to a certain category of participants, it certainly helps address bad smell “using deprecated data”, as well as threats to *internal validity*. The latter, in particular, has to do with whether the paper’s statistical analysis is capable of identifying cause-effect relations by addressing biases and confounding effects; this is the crux of the paper’s contributions.

The advantages brought by using different programming languages is a popular question for both researchers and practitioners (see Section 2.3); thus, the bad smell “not interesting” is not a risk for our research.

### 7.2.2 Modeling

Statistical modeling in the paper (Section 4) is based on widely used generalized regression models, starting from a minimal model  $m_1$  and extending it with other predictors that are available in the data. We then argued—based on statistical model comparison techniques, as well as on causal analysis considerations—about the advantages and disadvantages of each model over the others. This process helps avoid the bad smell “not exploring simplicity.”

Conversely, no analysis can be truly exhaustive and consider all possible models—nor can it claim that the models that it considered are *definitive*. Thus, it is always possible that more refined, complex, or extended models may reveal more nuanced trends in the data. As we briefly mentioned earlier in the paper, the replication package includes two more complex variants of model  $m_3$ , which however do not lead to qualitatively different conclusions about the impact of languages in Code Jam. As we have seen in Section 5, the causal models we consider in the paper are also somewhat robust, in that adding or removing some assumed causal links does not affect the model’s implications.

As we argued in previous work [13; 59], using Bayesian modeling techniques offers distinct advantages over the frequentist approaches that were dominant in the past. One of them is that there exist flexible *validation* techniques [14]; these techniques buttress the choice of statistical models by helping detect flawed models that are unsuitable to capture the analyzed data, as well as other problems related to unwarranted statistical “assumptions” (another bad smell [35]). For brevity, we did not discuss in detail the validation process in the paper, but the replication package documents how we followed it scrupulously, so as to mitigate any threats to *conclusion validity*.

Using Bayesian data analysis techniques also simplifies moving away from dichotomous (binary) views of “statistical significance”, whose skewing effects and questionable theoretical justifications have been repeatedly criticized [10; 56; 20; 62; 16; 34; 2; 5] (also as the related bad smell “ $p < 0.05$  and all that!”).

### 7.2.3 Analysis

We observed how the main object of our analysis—the effect of programming languages—is elusive as any programming-specific effects tend to be small in comparison to other factors. This necessarily limits any strong claims of *generalizability*—which is what *external validity* is mostly concerned with.

We took some measures to mitigate these issues as best as possible. First, we mainly focused on 50% probability intervals instead of the stricter 95% that are customarily used. This is a way of avoiding ending up with a greatly *underpowered study* (another bad smell of [35]). Second, we explored how the results change as we change the subset of Code Jam data that we analyze (Section 7.1), which is also a way of exploring robustness/stability (avoiding bad smell “not exploring stability”).

Finally, the replication package includes further analyses of sensitivity and using more complex multi-level models [18], which we do not present here for brevity. The sensitivity analysis confirms the modest relative effect of the programming language. The analysis results of using more complex models corroborate those obtained by using the paper’s plain regression models—which provides a further form of validation.

### 7.2.4 Other Issues

Providing a detailed replication package adheres to the best practices, and avoids the bad smell “inadequate reporting”. The replication package uses data visualization extensively, and so does the paper (addressing bad smell “no data visualization”).

As we discuss in Section 2 there is very little work on causal analysis for software engineering data; thus, this paper avoids the bad smell “not using related work”.

## 8 Discussion and Conclusions

As a concluding discussion, let’s review some of the high-level lessons that emerged from this paper’s case study.

**Small means small.** Regardless of the specific causal relations, our analysis confirms the observation, made by several other empirical studies (see Section 2), that the effects of using different programming languages tend to be small—especially relative to those of other variables. When effects are small, the robustness of any findings is necessarily lessened. However, causal analysis can still help to isolate general interactions whose overall impact may become more prominent in other scenarios; in other words, thinking causally can still improve external validity.

**When associations are enough.** If your exclusive or primary goal is making accurate predictions, a purely correlational model may be sufficient; in fact, it may even outperform a causally consistent model. For example, if you just want to predict the winner of the next Google Code Jam contest, you’re probably better off basing your predictions on model  $m_4$  (or an even more sophisticated model). Nevertheless, even in such situations, thinking about causality can still be useful as a sanity check and as a safeguard against misinterpreting or overgeneralizing your analysis results.

**Bayesian models remain more flexible.** As we repeatedly observed in the paper, the high-level results our analysis—in particular, the contrast between causal and non-causal statistical models—are largely independent of whether we deploy Bayesian or frequentist statistics. More generally, a lot of the causal modeling techniques we discussed in the paper can be applied to frequentist models as well. Nevertheless, we maintain that Bayesian statistics are preferable [13], as they are flexible and natural to interpret. For example, performing Section 6’s analysis on a frequentist model would be cumbersome (but still possible), whereas it was straightforward on the full posterior probability distribution that Bayesian models provide.

**Correlation vs. causation, again.** Thinking about causal relations is not the endgame of empirical data analysis. First, we already noted that there are scenarios in which predictive accuracy takes precedence. Second, causal and predictive models need not disagree—in fact, they are often consistent [52]. Third, the hypothesis underlying a causal model may need their own validation by other means, and may change as the data generating process changes or is better understood. It remains that, if we do not take causal relations into account, we may miss (important) parts of the picture. Conversely, thinking about causality prods us into looking at the data from a fresh perspective, rigorously thinking about confounding factors, and ultimately interpreting any potential findings in a more sound way.

## References

- [1] Alain Abran, James W Moore, Pierre Bourque, Robert Dupuis, and L Tripp. Software engineering body of knowledge. *IEEE Computer Society, Angela Burgess*, page 25, 2004.
- [2] Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. *Nature*, 567:305–307, 2019.
- [3] George K Baah, Andy Podgurski, and Mary Jean Harrold. Causal inference for statistical fault localization. In *Proceedings of the 19th international symposium on Software testing and analysis*, pages 73–84, 2010.
- [4] Alexandra Back and Emma Westman. Comparing programming languages in Google Code Jam. Master’s thesis, Chalmers University of Technology, 2017. <https://publications.lib.chalmers.se/records/fulltext/250672/250672.pdf>.
- [5] Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoiijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D.

- Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. Redefine statistical significance. *Human Nature Behavior*, 2:6–10, 2018.
- [6] Emery D. Berger, Celeste Hollenbeck, Petr Maj, Olga Vitek, and Jan Vitek. On the impact of programming languages on code quality: A reproduction study. *ACM Transactions on Programming Languages and Systems*, 41(4):21:1–21:24, 2019.
- [7] Pascal Caillet, Sarah Klemm, Michel Ducher, Alexandre Aussem, and Anne-Marie Schott. Hip fracture in the elderly: a re-analysis of the epidos study with causal bayesian networks. *PLoS One*, 10(3):e0120125, 2015.
- [8] Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 2022.
- [9] Andrew G Clark, Michael Foster, Benedikt Prifling, Neil Walkinshaw, Robert M Hierons, Volker Schmidt, and Robert D Turner. Testing causality in scientific modelling software. *arXiv preprint arXiv:2209.00357*, 2022.
- [10] Jacob Cohen. The earth is round ( $p < .05$ ). *American Psychologist*, 49(12):997–1003, 1994.
- [11] Clemens Dubslaff, Kallistos Weis, Christel Baier, and Sven Apel. Causality in configurable software systems. In *Proceedings of the 44th International Conference on Software Engineering, ICSE*. ACM, 2022.
- [12] Hongbo Fang, Hemank Lamba, James Herbsleb, and Bogdan Vasilescu. “This is damn slick!” Estimating the impact of tweets on open source project popularity and new contributors. In *Proceedings of the 44th International Conference on Software Engineering, ICSE*. ACM, 2022.
- [13] Carlo A. Furia, Robert Feldt, and Richard Torkar. Bayesian data analysis in empirical software engineering research. *IEEE Transactions on Software Engineering*, 47(9):1786–1810, September 2021. Publication date: August 2019.
- [14] Carlo A. Furia, Richard Torkar, and Robert Feldt. Applying Bayesian analysis guidelines to empirical software engineering data: The case of programming languages and code quality. *ACM Transactions on Software Engineering and Methodology*, 31(3):40:1–40:38, 2022.
- [15] Carlo A. Furia, Richard Torkar, and Robert Feldt. Replication package. <https://doi.org/10.5281/zenodo.7541480>, January 2023.
- [16] Andrew Gelman. The problems with  $p$ -values are not just with  $p$ -values. *The American Statistician*, 70, 2016. Online discussion: [http://www.stat.columbia.edu/~gelman/research/published/asa\\_pvalues.pdf](http://www.stat.columbia.edu/~gelman/research/published/asa_pvalues.pdf).
- [17] Andrew Gelman. Why I prefer 50% rather than 95% intervals. <https://statmodeling.stat.columbia.edu/2016/11/05/why-i-prefer-50-to-95-intervals/>, November 2016. From the blog *Statistical Modeling, Causal Inference, and Social Science*.
- [18] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Analytical Methods for Social Research. Cambridge University Press, Cambridge, UK, 2020.
- [19] Andrew Gelman and David Weakliem. Of beauty, sex and power. *American Scientist*, 97:310–316, 2009.
- [20] Steven N. Goodman. Toward evidence-based medical statistics. 1: The  $p$  value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999.
- [21] Lasse Hakulinen. Survey on informatics competitions: Developing tasks. In *Olympiads in Informatics*, volume 5, pages 12–25. IOI, 2011.
- [22] Joseph Halpern. A modification of the halpern-pearl definition of causality. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [23] Stefan Hanenberg. An experiment about static and dynamic type systems: Doubts about the positive impact of static type systems on development time. In *Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications, OOPSLA ’10*, pages 22–35, New York, NY, USA, 2010. ACM.

- [24] Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):579–599, 2022. Also <https://arxiv.org/abs/1907.02435>.
- [25] Miguel A Hernán and Susana Monge. Selection bias due to conditioning on a collider. *BMJ*, 381:1135, 2023.
- [26] Hans-Martin Heyn and Eric Knauss. Structural causal models as boundary objects in ai system development. In *1st International Conference on AI Engineering-Software Engineering for AI*, 2022.
- [27] Guido W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.
- [28] Edwin T. Jaynes. *Probability theory: The logic of science*. Cambridge University Press, Cambridge, 2003.
- [29] Hieke Keuning, Johan Jeuring, and Bastiaan Heeren. A systematic literature review of automated feedback generation for programming exercises. *ACM Trans. Comput. Educ.*, 19(1):3:1–3:43, 2019.
- [30] Seongmin Lee, Dave Binkley, Robert Feldt, Nicolas Gold, and Shin Yoo. Causal program dependence analysis. *arXiv preprint arXiv:2104.09107*, 2021.
- [31] Xing Li, Yinbo Yu, Kai Bu, Yan Chen, Jianfeng Yang, and Ruijie Quan. Thinking inside the box: Differential fault localization for sdn control plane. In *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pages 353–359. IEEE, 2019.
- [32] Yuchu Liu, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, and Jonn Lantz. Bayesian causal inference in automotive software engineering and online evaluation. *arXiv preprint arXiv:2207.00222*, 2022.
- [33] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, Florida, USA, 2 edition, 2020.
- [34] Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon statistical significance. *The American Statistician*, 73(S1):235–245, 2019.
- [35] Tim Menzies and Martin Shepperd. "Bad smells" in software analytics papers. *Information and Software Technology*, 112:35–47, 2019.
- [36] Leo A. Meyerovich and Ariel S. Rabkin. Empirical analysis of programming language adoption. In *Proceedings of the 2013 ACM SIGPLAN International Conference on Object Oriented Programming Systems Languages & Applications, OOPSLA '13*, pages 1–18, New York, NY, USA, 2013. ACM.
- [37] Sebastian Nanz and Carlo A. Furia. A comparative study of programming languages in Rosetta Code. <http://arxiv.org/abs/1409.0252>, September 2014.
- [38] Sebastian Nanz and Carlo A. Furia. A comparative study of programming languages in Rosetta Code. In Antonia Bertolino, Gerardo Canfora, and Sebastian Elbaum, editors, *Proceedings of the 37th International Conference on Software Engineering (ICSE)*, pages 778–788. ACM, May 2015.
- [39] Sebastian Nanz, Faraz Torshizi, Michela Pedroni, and Bertrand Meyer. Design of an empirical study for comparing the usability of concurrent programming languages. In *Proceedings of the 2011 International Symposium on Empirical Software Engineering and Measurement, ESEM '11*, pages 325–334, Washington, DC, USA, 2011. IEEE Computer Society.
- [40] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.
- [41] Judea Pearl. *Causality*. Cambridge University Press, 2009.
- [42] Judea Pearl. *Causality: Models, reasoning and inference*. Cambridge University Press, 2nd edition, 2009.
- [43] Judea Pearl. The mathematics of causal inference. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*, page 5, New York, NY, USA, 2011. Association for Computing Machinery.

- [44] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.
- [45] Judea Pearl and Dana Mackenzie. *The Book of Why*. Penguin Random House, 2018.
- [46] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- [47] Lutz Prechelt. An empirical comparison of seven programming languages. *IEEE Computer*, 33(10):23–29, October 2000.
- [48] Baishakhi Ray, Daryl Posnett, Vladimir Filkov, and Premkumar Devanbu. A large scale study of programming languages and code quality in Github. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, FSE 2014*, pages 155–165, New York, NY, USA, 2014. Association for Computing Machinery.
- [49] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.
- [50] Christopher J. Rossbach, Owen S. Hofmann, and Emmett Witchel. Is transactional programming actually easier? In *Proceedings of the 15th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP ’10*, pages 47–56, New York, NY, USA, 2010. ACM.
- [51] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- [52] Maximilian Scholz and Paul-Christian Bürkner. Prediction can be safely used as a proxy for explanation in causally consistent Bayesian generalized linear models. <https://arxiv.org/abs/2210.06927>, October 2022.
- [53] Maximilian Scholz and Richard Torkar. An empirical study of linespots: A novel past-fault algorithm. *Software Testing, Verification and Reliability*, 31(8):e1787, 2021.
- [54] Hyunmin Seo, Caitlin Sadowski, Sebastian G. Elbaum, Edward Aftandilian, and Robert W. Bowdidge. Programmers’ build errors: a case study (at Google). In *36th International Conference on Software Engineering, ICSE ’14, Hyderabad, India - May 31 - June 07, 2014*, pages 724–734. ACM, 2014.
- [55] Julian Siebert. Applications of statistical causal inference in software engineering. *arXiv preprint arXiv:2211.11482*, 2022.
- [56] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology. *Psychological Science*, 22(11):1359–1366, 2011.
- [57] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen, 2016.
- [58] Jonathan Sterne. We should be cautious about associations of patient characteristics with COVID-19 outcomes that are identified in hospitalised patients. Health Data Research UK – <https://www.hdruk.ac.uk/news/we-should-be-cautious-about-associations-of-patient-characteristics-with-covid-19-outcomes-that-are-identified-in-hospitalised-patients>, 2020.
- [59] Richard Torkar, Carlo A. Furia, Robert Feldt, Francisco Gomes de Oliveira Neto, Lucas Gren, Per Lenberg, and Neil A. Ernst. A method to assess and argue for practical significance in software engineering. *IEEE Transactions on Software Engineering*, 48(6):2053–2065, June 2022. Online since January 2021.
- [60] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.
- [61] Tom Verhoeff. The role of competitions in education. In *Future World: Educating for the 21st Century*. IOI, 12 1997.
- [62] Ronald L. Wasserstein and Nicole A. Lazar. The ASA statement on  $p$ -values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. <https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf>.