Towards Improving the External Validity of Software Engineering Experiments with Transportability Methods

Julian Frattini
Richard Torkar*
Robert Feldt†
{firstname}.{lastname}@chalmers.se
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden

Carlo A. Furia
furiac@usi.ch
USI Università della Svizzera italiana
Lugano, Switzerland

Abstract

Controlled experiments are a core research method in software engineering (SE) for validating causal claims. However, recruiting a sample of participants that represents the intended target population is often difficult or expensive, which limits the external validity of experimental results. At the same time, SE researchers often have access to much larger amounts of observational than experimental data (e.g., from repositories, issue trackers, logs, surveys and industrial processes). *Transportability methods* combine these data from experimental and observational studies to "transport" results from the experimental sample to a broader, more representative sample of the target population. Although the ability to combine observational and experimental data in a principled way could substantially benefit empirical SE research, transportability methods have—to our knowledge—not been adopted in SE. In this vision, we aim to help make that adoption possible. To that end, we introduce transportability methods and their prerequisites, and demonstrate their potential through a simulation. We then outline several SE research scenarios in which these methods could apply, e.g., how to effectively use students as substitutes for developers. Finally, we outline a road map and practical guidelines to support SE researchers in applying them. Adopting transportability methods in SE research can strengthen the external validity of controlled experiments and help the field produce results that are both more reliable and more useful in practice.

CCS Concepts

• **General and reference** → **Experimentation**; **Reliability**; • **Computing methodologies** → *Scientific visualization.*

Keywords

Controlled Experiment, Transportability, External Validity

*Also with The Stellenbosch Institute for Advanced Study.
†Also with Mid Sweden University.

EASE 2026, Glasgow, United Kingdom

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Julian Frattini, Richard Torkar, Robert Feldt, and Carlo A. Furia. 2026. Towards Improving the External Validity of Software Engineering Experiments with Transportability Methods. In *Proceedings of The 30th International Conference on Evaluation and Assessment in Software Engineering (EASE 2026).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Controlled experiments are an essential research method in software engineering (SE)—as in any empirical discipline—for validating claims about causal relationships between variables [30]. The random assignment of study subjects to a treatment or control group eliminates the influence of confounding factors on the relationship of interest [19]. Therefore, the observed effect can be attributed to the treatment rather than to confounding effects.

However, this internal validity often comes at the expense of external validity. Experiments are conducted in a contrived setting [24] where a representative sample of subjects must be drawn from a target population [19, 23]. Achieving a broad and representative sample is particularly challenging when an experiment involves human subjects: the intended target population of SE professionals is difficult to reach and expensive to recruit [1]. Consequently, controlled experiments in SE often settle for small samples and participants with a limited experience (e.g., students), thus jeopardizing statistical power and external validity [6]. This hinders transfer of scientific results into practice [21].

Other empirical disciplines face the same challenges. For example, medical researchers aim to predict how well a treatment response observed in a sample will hold in the target population, i.e., all potential recipients of that treatment [16]. To address this, the field of statistical causal inference has developed a formal framework for *transportability* of statistical relations across populations [15]. Within this broader line of work, "[e]stimation methods to generalize trial findings to a target population of interest" [3] emerged, which we will refer to as *transportability methods* from here on out. These methods combine experimental results with typically much larger observational data on relevant covariates, allowing to transport results from limited experiments to a target population without collecting more experimental data [3].

Despite their potential, such methods have—to our knowledge—not been adopted in SE research to date. With this vision paper, our goal is to pave the way for the adoption of transportability methods in SE research by contributing: (1) a high-level description of transportability methods and their necessary preconditions (Section 2),

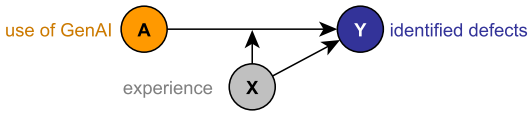

Figure 1: DAG visualizing causal assumptions of the illustrative, running example

(2) a simulation demonstrating the methods' usefulness (Section 3), (3) a list of valuable cases for application in SE research (Section 4), and (4) a road map for enabling adoption in SE research (Section 5).

Data Availability Statement

All figures, scripts, and documentation can be found in our replication package [10].

2 Transportability Methods

Section 2.1 introduces an illustrative example contextualizing the subsequent, methodological descriptions. Then, Section 2.2 lists relevant preconditions that need to be met for the actual transportability methods described in Section 2.3 to work.

2.1 Illustrative Example

At a high level, a controlled experiment estimates the causal effect of a treatment A on an outcome Y (i.e., $A \rightarrow Y$). As a running example, we will consider the effect of using (i.e., the treatment $A = 1$) or not using (i.e., the control $A = 0$) generative AI (GenAI) on the number Y of successfully identified defects during code reviews [27]. The quantity of interest to estimate from the experiment is the *average treatment effect* (ATE) τ, i.e., the average difference in detected defects when using GenAI instead of not using it.

The level of experience X of a subject is an example of a covariate that may affect the outcome Y directly ($X \rightarrow Y$), but may also *moderate* the ATE [20]: Subjects with less experience may benefit more from using GenAI during code reviews than subjects with more experience. In the absence of experience, suggestions from GenAI may be a decent help, while the same suggestions may be trivial for an experienced reviewer. This makes X a *treatment effect modifier*. Figure 1 visualizes these relations as a directed acyclic graph (DAG), commonly used in Pearl's framework for causal inference [15].

Controlled experiments aim to approximate the ATE τ in the target population, but can realistically only measure the *trial* ATE τ_1 in the experimental sample. The ATE of interest τ may differ from the measurable trial ATE τ_1. In our illustrative example, one reason for this difference may stem from the challenge of recruiting subjects. In particular, *trial eligibility* S (i.e., the likelihood of a subject from the target population to be included in the experimental sample) is often affected by a treatment effect modifier such as X: We can assume that subjects with more experience X are in more senior positions because it increases the likelihood of getting promoted [29]. This makes them less accessible and more expensive to recruit as subjects to the experiment. In contrast, subjects with less experience X may be more available to participate in the experiment. For this reason, SE experiments commonly use university students to represent the target population of software

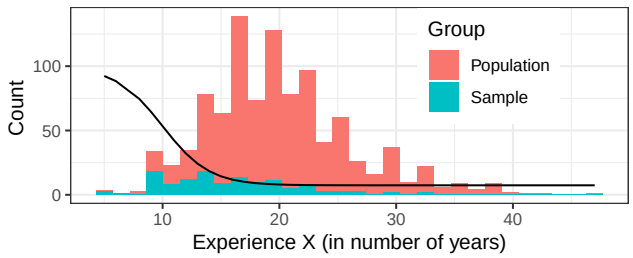

Figure 2: Distributions of the covariate X in the target population and experimental sample due to covariate shift

engineers [2]. Figure 2 visualizes this challenge. The black line represents subjects' trial eligibility (in percentage) which decreases for higher values of X. The distribution of the covariate X in the experimental sample (teal bars) therefore ends up different from the distribution in the target population (red bars)—a phenomenon known as *covariate shift* [25]. When X is both a treatment effect modifier and experiences a covariate shift, the measurable τ_1 can differ from τ. In our example, the experiment likely involves more easy-to-recruit but inexperienced subjects for which the measured effect is particularly strong. As a consequence, the experiment will overestimate $\tau_1 > \tau$ and suggest that using GenAI is much more effective than it would be in the target population.

2.2 Preconditions

Table 1 lists the preconditions that must hold for the transportability methods (presented in Section 2.3) to work, as elicited by Colnet et al. [3]. Preconditions A1 through A4 are fundamental requirements for a valid controlled experiment. Preconditions A5, A6, and A7 are specific to transportability methods. Thus, we discuss what they mean and what happens if they are violated in the following.

A5 requires that at least one covariate X affects trial eligibility S, i.e., the black line in Figure 2 is not just a horizontal line. If A5 does not hold, then the distribution of X would be the same in the experimental sample as in the target population. In such a case, the experimental sample perfectly represents the target population (i.e., there is no covariate shift), the trial ATE would perfectly generalize ($\tau = \tau_1$), and there would be no need for transporting. As we argued in the illustrative example in Section 2.1, and as we will further elaborate in Section 4, in most practical cases there would be at least some covariate shift, i.e., A5 normally holds.

A6 requires that for every stratum of $x \in X$ the ATE in the target population is the same as the trial ATE, i.e., the *conditional* ATE is the same even if the marginal ATE may not be. In other words, if we stratify by the covariate X, the trial ATE generalizes to the target population. This implies that if A6 holds, X acts as the only treatment effect modifier. A situation where A6 would not hold is if there are other *unobserved* covariates that moderates the treatment effect. In this case, applying transportabilty methods to only X may fail to correct for all of the covariate shift.

Finally, A7 requires that every subject from the target population has non-zero probability of being included in the experimental sample, i.e., the black line in Figure 2 is always above 0%. In general, the distribution of X in the experimental sample will differ from in the target population (see A5). A7 only requires that the two

Table 1: Assumptions to apply transportability methods to controlled experiments. An asterisk * marks those that are specific to transportabilty techniques.

ID	Name	Definition	Explanation
A1	Consistency	$Y = AY(1) + (1-A)Y(0)$	The observed outcome is the potential outcome given the assigned treatment, i.e., we have a connection between treatment and outcome.
A2	Randomization	$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid S = 1, X$	The treatment is independent of all the potential outcomes and covariates as in a controlled experiment.
A3	Ignorability on trial participation	$\{Y(0), Y(1)\} \perp\!\!\!\perp S \mid X$	The outcome Y is unaffected by trial participation S when controlling all relevant covariates X.
A4	Mean exchangeability	$\mathbb{E}[Y(a) \mid X = x, S = 1] = \mathbb{E}[Y(a) \mid X = x]$	Instead of requiring that every individual behaves identically in and out of the trial, we only assume that *on average* the treatment effect is the same between groups with the same observed characteristics.
A5*	Sample ignorability for treatment effects	$Y(1) - Y(0) \perp\!\!\!\perp S \mid X$	The effect of the treatment is independent of trial eligibility when knowing all covariates X.
A6*	Transportability of the conditional ATE	$\forall x \in X: \tau_1(x) = \tau(x)$	In every stratum $x \in X$, the ATE in the experimental sample τ_1 is equal to the ATE in the target population τ.
A7*	Positivity of trial participation	$\exists c: \mathbb{P}(S = 1 \mid X) \geq c$	Every subject in the target population must have at least some chance (i.e., non-zero probability) of being included in the experiment.

distributions have the same support. If A7 does not hold—i.e., some stratum of X has a 0% probability of being sampled—no statistical method could recover the ATE from the unobserved stratum. Transportability is, hence, constrained to the range of X covered in the experimental sample.

2.3 Formulae and Estimation Methods

One approach to approximate the ATE τ is to model the treatment effect modification as an interaction effect in a regression formula:

$$Y \sim \mathcal{N}(\alpha + \tau \cdot A \cdot X, \epsilon) \tag{1}$$

This formula regresses the outcome Y (here assumed to be normally distributed $\mathcal{N}$ with variance ϵ) on a linear combination of an intercept α (the baseline value for Y) and the treatment A, which has an effect of τ on the outcome but is moderated by X. For simplicity of the demonstration, we ignore all marginal effects of A and X on Y.

However, this approach of estimating τ only works if the interaction between the continuous X and A is linear. This would require that every increase in the covariate X causes the same proportional increase in the treatment effect moderation. However, not every effect behaves this way, particularly when considering human factors [13]. In the example where the covariate X is a continuous measure of experience in number of years, it is possible that an increase of experience from 0 to 1 year has a greater effect than from 20 to 21 years. To handle these more complex interactions, a more general approach is needed.

Enter transportability methods. Under the conditions in Section 2.2, a transportability method can recover the actual ATE τ of the target population from (1) the trial ATE τ_1 and (2) the distribution the the covariates X in the target population, but without requiring further data about A or Y.

Colnet et al. discuss two classes of identification formulae [3]:

(1) **Reweighting**: $\tau = \mathbb{E}\left[\frac{n}{m \times \alpha(X)} \tau_1(X) \mid S = 1\right]$
(2) **Regression**: $\tau = \mathbb{E}\left[\mu_{A=1,S=1}(X) - \mu_{A=0,S=1}(X)\right] = \mathbb{E}\left[\tau_1(X)\right]$

Based on these formulae, they elaborate several estimation methods for transportability. For brevity, we will only present one from each class and refer the interested reader to Colnet et al. [3].

Transport with reweighting: The **inverse probability of sampling weighting** (IPSW) is an estimator of τ based on reweighting. IPSW weighs each data point in the controlled experiment based on trial eligibility:

$$\hat{\tau}_{\text{IPSW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{n}{m} \frac{Y_i}{\hat{\alpha}_{n,m}(X_i)} \left(\frac{A_i}{e_1(X_i)} - \frac{1 - A_i}{1 - e_1(X_i)} \right) \tag{2}$$

Here, n is the size of the experimental sample, m the size of the larger target population, $\hat{\alpha}_{n,m}(X_i)$ represents the trial eligibility S, and $e_1(x)$ the propensity score [3] (i.e., the likelihood of being assigned to a treatment, which is fixed at 50% in most experiments with only one treatment and one control level). Trial eligibility $\hat{\alpha}_{n,m}(X_i)$ can be estimated via logistic regression based on the distribution of X in the experimental sample and in the target population. Values of X that occur often in the sample and in the target population have a high trial eligibility, values of X that occur rarely in the sample but more often in the target population have a low trial eligibility. Based on this estimated trial eligibility, the data points from the controlled experiment are re-weighted. Data points from subjects with high trial eligibility contribute less to the ATE than from subjects with low trial eligibility. In the illustrative example, this would mean that the results obtained from one participating senior engineer (high experience X, and therefore, low trial eligibility) are weighted more strongly in estimating the ATE than the results obtained from several participating master students (low experience X, and therefore, high trial eligibility). This weighting of results by the inverse probability of sampling counteracts the effect of the covariate on the trial eligibility.

Transport with regression: The **plug-in g-formula** is an estimator of τ based on regression. This estimator approximates τ by fitting two separate linear models.

$$\hat{\tau}_G = \frac{1}{m} \sum_{i=n+1}^{n+m} \left(\hat{\mu}_{1,1,n}(X_i) - \hat{\mu}_{0,1,n}(X_i) \right) \tag{3}$$

The two linear models predict the outcome Y based on X, one for the control group ($\hat{\mu}_{0,1}$) and one for the treatment group ($\hat{\mu}_{1,1}$). These regressions $Y \sim X$ for the two levels of A directly model the treatment effect moderation of X. The plug-in g-formula estimation then applies the covariate value X_i of all m observational data points to both linear models, averages the results, and calculates the ATE as the difference between the two averages.

3 Simulation

To demonstrate how transportability methods work in practice, we perform a computer simulation [24]. We simulate a target population and draw a sample from it that represents participants of an experimental study. We then simulate this experiment with known causal effects among variables. Finally, we estimate the ATE using four methods: mean difference, linear regression with an interaction term, and the two presented transportability methods. We compare the four methods in their ability to recover the simulated causal effect from the data.

3.1 Dataset construction

We simulate the illustrative example described in Section 2.1. The main factor A has two levels: control ($A = 0$, i.e., not using AI) and treatment ($A = 1$, i.e., using AI). We use a normally distributed measure representing *defect detection performance* instead of the number of identified defects for the outcome $Y \in \mathbb{R}$. Using this normally distributed outcome simplifies interpretation by avoiding link functions required for count data [14], though the transportability methods work just as well for such data. Finally, the covariate $X \in \mathbb{R}^+$ represents experience measured in number of years and follows a negative-binomial (NB) distribution (as in Figure 2).

We created a data set by first simulating a target population of 1000 subjects with a random distribution of the covariate $X \sim \text{NB}(10, 3)$. The scale parameter $\mu = 10$ and dispersion parameter $\gamma = 3$ are arbitrary but produce realistic values between 0 and about 50 years of experience with a peak around $X = 20$, as seen in Figure 2. Next, we simulated the trial eligibility $S \sim \text{Bernoulli}(p)$, where the likelihood of being included in the experimental sample decreases with X (as shown as the black line in Figure 2). Subjects where $S = 1$ are included in the controlled experiment, the remaining subjects where $S = 0$ remain in the observational group. This split the data set into roughly $n = 175$ experimental subjects and $m = 825$ observational subjects, though the exact numbers vary due to the random distribution. Finally, we randomly divided the experimental subjects into control ($A = 0$) and treatment ($A = 1$) groups and simulated the outcome Y which is affected by the treatment A but moderated by the covariate X. For the ATE, we chose an arbitrary value of $\tau = 16.7$. For the treatment effect moderation, our simulation decreased the ATE with higher values of X in a non-linear way, which models diminishing returns of increasing experience. We did not simulate a marginal effect of $X \to Y$, i.e., the outcome Y did not change for different values of X directly, only through the treatment effect moderation.

3.2 Estimation Setup

In the evaluation, we compare four methods to estimate the ATE:

(1) **Mean difference** (baseline) between the outcome Y in the control and treatment group
(2) **Linear regression model** with an interaction effect representing the treatment effect moderation (Equation (1))
(3) **IPSW estimator** from the reweighting-class (Equation (2))
(4) **Plug-in g-formula** from the regression-class (Equation (3))

We run the simulation described above 50 times. For each simulated dataset, we record the ATE estimated by each of the four methods and then plot the distribution of these estimates.

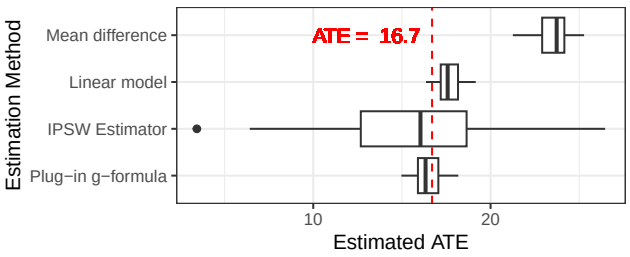

Figure 3: Results from the simulation compared against the simulated effect (red line).

3.3 Results

Figure 3 shows the results of the simulation. The box plots represent the estimated results of each of the four methods over 50 iterations. The red dashed line shows the simulated ATE ($\tau = 16.7$) that these methods attempted to recover.

The naïve mean difference vastly overestimates the simulated ATE. Since the experimental sample predominantly contained subjects with lower experience X and the ATE of the main factor A is moderated to be stronger for lower values of X, the naïve estimation assumes the ATE to be much stronger than it truly is ($\tau_1 \gg \tau$).

The linear model including an interaction effect performs significantly better, but still overestimates the simulated ATE. This is because it models the interaction to be linear, while the treatment effect moderation is actually non-linear.

The 50%-quantiles of estimations of both transportability methods include the simulated ATE thanks to the covariate distribution X in the target population. However, the IPSW estimator shows substantially greater uncertainty around its mean estimate. As Colnet et al. explain, this estimator can be highly unstable, particularly when the trial-eligibility weights become extreme [3]. The plug-in g-formula performs better on both accounts: it is more accurate and more robust. Using the information about the covariate distribution from the target population, it is able to correct the treatment effect moderation and recover the simulated ATE.

4 Motivating Examples

Beyond the illustrative example used in Sections 2 and 3, we identify three classes of challenges in empirical SE research for which transportability methods may be worth considering.

4.1 Experiment Participant Experience

A long-running debate in SE research asks whether (undergraduate) students can serve as valid substitutes for SE professionals in controlled experiments [4, 17]. Students are easier to recruit, but they may lack the skills or domain knowledge of professional practitioners [5]. This question has fueled an extensive public discussion, with prominent empirical SE researchers arguing both sides [2, 7, 8]. Yet the debate has relied mostly on hypotheses, assumptions, and anecdotal evidence rather than direct empirical tests. Transportability methods offer a constructive way forward for understanding and addressing the issue of representative subjects in SE experiments.

4.2 System Properties

Not only human participants but also the artifacts used in experiments may fail to represent the target population. Researchers often study software systems built in student projects [11], specifications mocked for the experiment [9], or artificial bugs injected into software [12]. Industry-grade artifacts may be unavailable, unsuitable for time-constrained experiments, or missing properties that the study requires (e.g., ground-truth traceability links [11]). Even when experiments use industry-grade artifacts, they are often restricted to open-source systems because those are accessible [18].

Smaller, simpler, hand-crafted, or open-source artifacts are often more practical, but they may not represent the target population of software systems, specifications, or other artifacts. This creates covariate shift in characteristics such as size, complexity, documentation quality, which affects how well results generalize. Framed as a transportability problem, the objects' representativeness becomes a tangible property and limitations to external validity clear.

4.3 Task Complexity

In addition to human and artifact subjects, the experimental tasks themselves are often not fully representative of real-world practice [23]. Researchers often limit the scope of a task to minimize the required time commitment of participants, e.g., code reviews without extensive familiarization with the source code [27]. This sacrifices representativeness of a task, raising the question whether effects observed during the experimental task still hold in reality.

5 Road Map and Guidelines

Even in medical research, where transportability methods originated, their application is still limited [3]. We see the opportunity to enable this useful class of methods for SE research by focusing effort on the following steps.

5.1 Understanding Treatment Effect Modifiers

Firstly, SE research should develop a clear understanding of which covariates act as moderators on ATEs of interest, as these limit external validity of experimental results. Identifying them would support a more rigorous and systematic analysis of the threats to external validity, rather than resorting to common practice [31].

We anticipate that several covariates will be specific to certain SE tasks, while others apply to a broader scope. For example, *experience*, *domain knowledge*, or *skill* probably moderate many causal effects of interest [28], as they are likely to influence almost any SE activity. In contrast, a covariate like *programming language proficiency* will affect some SE tasks (e.g., source code development and code reviews) [27] more than others (e.g., requirements elicitation).

Although identifying all moderators is difficult, causal models (Figure 1) make these assumptions explicit. Rather than aiming for "perfect" knowledge, researchers should use these models for sensitivity analyses that quantify how unobserved moderators could bias the transported ATE [14]. This shifts the focus from exhaustive completeness to the statistical robustness of the external-validity claim. These analyses can also help researchers prioritize the factors that matter most when designing an experiment: They should collect data on key moderating covariates and seek a representative sample that spans the full range of each covariate.

5.2 Operationalizing Covariates

Once relevant covariates are identified, SE research must develop appropriate and agreed-upon operationalizations. Since many of the moderating covariates are likely to be latent variables and context factors, their operationalization is critical [22]. For example, *experience* is often operationalized via the *number of years working as a software engineer*, which may not adequately reflect the underlying concept: If one software engineer has worked for twice as long as another, there is no guarantee that they are also "twice as experienced." A proper operationalization underpins the construct validity of these covariates. Without it, the previously introduced transportability methods lose effectiveness. Therefore, thoroughly assessing the construct validity of operationalizations of covariates moderating an ATE [26] will pave the way towards adjusting for them using transportability methods.

5.3 Collecting Observational Data

With relevant covariates identified and operationalized, the SE research community can steer its efforts towards collecting observational data on these covariates in the target population. While surveying the total target population remains unrealistic, observational studies collecting covariate distributions are likely to involve larger samples of the target distribution compared to interventional studies (e.g., controlled experiments or action research studies), given that they are less obtrusive [24]. For example, if *experience* is identified as a relevant, ATE-moderating covariate for several SE tasks, surveys collecting the distribution of developer experience in different countries and companies can be conducted to approximate the distribution of that covariate in the general target population.

5.4 Transporting Results

With observational data sets approximating the distribution of relevant covariates, SE researchers can transport the results of controlled experiments from an experimental to an observational sample, where the latter is more representative of the target population. Thanks to the previously presented methods, covariate shift in controlled experiments can be partially addressed when the assumptions hold. For example, controlled experiments can be conducted primarily with students (i.e., subjects with lower *experience*) as long as there are still a few subjects representing the other end of the spectrum of the covariate (i.e., subjects with higher *experience*) to meet A7. This also implies the advice that—given an existing sample of students—effort is better spent on recruiting a few senior software engineers instead of a lot more students. Ultimately, when an experiment meets the assumptions in Section 2.2 and observational data on ATE-moderating covariates is available, transportability methods can improve the external validity of results without requiring additional experimental data.

5.5 Presenting Results

Finally, these methods allow contextualizing obtained results in two regards. First, the presence of a treatment effect modifier allows complementing the ATE with the results about the actual moderation. While the ATE represents the *average* effect aggregated over the full range of the covariate X, a stratified view into how the effect changes along X provides more detailed insights. In the

illustrative example, this would allow the conclusion that the use of GenAI is beneficial for inexperienced subjects but irrelevant for experienced ones. Second, assessing the degree to which precondition A7—the positivity of trial participation—is met allows confining the external validity of the achieved results. If it was impossible to recruit subjects or infeasible to sample objects that cover the full spectrum of a treatment effect modifying covariate, the obtained range should be reported to confine the scope of generalizability. In the illustrative example, the data point with the larges value for X (33 in Figure 2) defines the upper end of transportability.

6 Conclusion

Transportability methods have the potential to improve the external validity of results from controlled experiments and increase their practical relevance. If covariates moderate the ATE of a phenomenon of interest while simultaneously affecting trial eligibility, external validity is under threat. However, if observational data about those covariates from a larger sample is available, results can be transported to this larger sample using transportability methods. Their application could help address several long-standing issues with experimentation in SE. Still, the path to adopt transportability methods in SE requires addressing several challenges in order to meet all preconditions. Targeting this goal will encourage SE researchers to explore and understand relevant covariates, collect data about them, and actively reason about the representativeness of their experimental subjects, objects, and tasks. In future work, we aim to demonstrate the application to real cases of SE research.

References

[1] Sebastian Baltes and Paul Ralph. 2022. Sampling in software engineering research: A critical review and guidelines. *Empirical Software Engineering* 27, 4 (2022), 94. doi:10.1007/s10664-021-10072-8

[2] Jeffrey Carver, Letizia Jaccheri, Sandro Morasca, and Forrest Shull. 2004. Issues in using students in empirical studies in software engineering education. In *Proceedings. 5th international workshop on enterprise networking and computing in healthcare industry (IEEE Cat. No. 03EX717)*. IEEE, 239–249. doi:10.1109/METRIC.2003.1232471

[3] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. 2024. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science* 39, 1 (2024), 165–191. doi:10.1214/23-STS889

[4] Bill Curtis. 1986. By the way, did anyone study any real programmers?. In *Papers presented at the first workshop on empirical studies of programmers on Empirical studies of programmers*. 256–262. doi:10.5555/21842.28899

[5] Oscar Dieste, Natalia Juristo, and Mauro Danilo Martínez. 2013. Software industry experiments: A systematic literature review. In *2013 1st International Workshop on Conducting Empirical Studies in Industry (CESI)*. IEEE, 2–8. doi:10.1109/CESI.2013.6618462

[6] Tore Dybå, Vigdis By Kampenes, and Dag IK Sjøberg. 2006. A systematic review of statistical power in software engineering experiments. *Information and Software Technology* 48, 8 (2006), 745–755. doi:10.1016/j.infsof.2005.08.009

[7] Davide Falessi, Natalia Juristo, Claes Wohlin, Burak Turhan, Jürgen Münch, Andreas Jedlitschka, and Markku Oivo. 2018. Empirical software engineering experts on the use of students and professionals in experiments. *Empirical Software Engineering* 23, 1 (2018), 452–489. doi:10.1007/s10664-017-9523-3

[8] Robert Feldt, Thomas Zimmermann, Gunnar R Bergersen, Davide Falessi, Andreas Jedlitschka, Natalia Juristo, Jürgen Münch, Markku Oivo, Per Runeson, Martin Shepperd, et al. 2018. Four commentaries on the use of students and professionals in empirical software engineering experiments. *Empirical Software Engineering* 23, 6 (2018), 3801–3820.

[9] Julian Frattini, Davide Fucci, Richard Torkar, Lloyd Montgomery, Michael Unterkalmsteiner, Jannik Fischbach, and Daniel Mendez. 2025. Applying bayesian data analysis for causal inference about requirements quality: a controlled experiment. *Empirical Software Engineering* 30, 1 (2025), 29. doi:10.1007/s10664-024-10582-1

[10] Julian Frattini, Richard Torkar, Robert Feldt, and Carlo Furia. 2026. Replication Package. https://doi.org/10.5281/zenodo.19451793. Last accessed 2026-04-07.

[11] Tobias Hey, Jan Keim, and Sophie Corallo. 2024. Requirements classification for traceability link recovery. In *2024 IEEE 32nd International Requirements Engineering Conference (RE)*. IEEE, 155–167. doi:10.1109/RE59067.2024.00024

[12] René Just, Darioush Jalali, Laura Inozemtseva, Michael D. Ernst, Reid Holmes, and Gordon Fraser. 2014. Are mutants a valid substitute for real faults in software testing?. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, (FSE-22), Hong Kong, China, November 16 - 22, 2014*, Shing-Chi Cheung, Alessandro Orso, and Margaret-Anne D. Storey (Eds.). ACM, 654–665. doi:10.1145/2635868.2635929

[13] Johnson Ching-Hong Li. 2018. Curvilinear moderation—a more complete examination of moderation effects in behavioral sciences. *Frontiers in Applied Mathematics and Statistics* 4 (2018), 7. doi:10.3389/fams.2018.00007

[14] Richard McElreath. 2018. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC. doi:10.1201/9781315372495

[15] Judea Pearl and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25. 247–254.

[16] Peter M Rothwell. 2005. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet* 365, 9453 (2005), 82–93. doi:10.1016/S0140-6736(04)17670-8

[17] Iflaah Salman, Ayse Tosun Misirli, and Natalia Juristo. 2015. Are students representatives of professionals in software engineering experiments?. In *2015 IEEE/ACM 37th IEEE international conference on software engineering*, Vol. 1. IEEE, 666–676. doi:10.1109/ICSE.2015.82

[18] Yorick Sens, Henriette Knopp, Sven Peldszus, and Thorsten Berger. 2025. A Large-Scale Study of Model Integration in ML-Enabled Software Systems. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. 1165–1177. doi:10.1109/ICSE55347.2025.00185

[19] William R Shadish, Thomas D Cook, and Donald T Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company. doi:10.1086/345281

[20] Janet Siegmund and Jana Schumann. 2015. Confounding parameters on program comprehension: a literature survey. *Empirical Software Engineering* 20, 4 (2015), 1159–1192. doi:10.1007/s10664-014-9318-8

[21] Dag IK Sjøberg, Bente Anda, Erik Arisholm, Tore Dybå, Magne Jørgensen, Amela Karahasanović, and Marek Vokáč. 2003. Challenges and recommendations when increasing the realism of controlled software engineering experiments. In *Empirical Methods and Studies in Software Engineering: Experiences from ESERNET*. Springer, 24–38. doi:10.1007/978-3-540-45143-3_3

[22] Dag IK Sjøberg and Gunnar Rye Bergersen. 2022. Construct validity in software engineering. *IEEE Transactions on Software Engineering* 49, 3 (2022), 1374–1396. doi:10.1109/TSE.2022.3176725

[23] Dag IK Sjøberg, Jo Erskine Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanovic, N-K Liborg, and Anette C Rekdal. 2005. A survey of controlled experiments in software engineering. *IEEE transactions on software engineering* 31, 9 (2005), 733–753. doi:10.1109/TSE.2005.97

[24] Klaas-Jan Stol and Brian Fitzgerald. 2018. The ABC of software engineering research. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 27, 3 (2018), 1–51. doi:10.1145/3241743

[25] Masashi Sugiyama and Motoaki Kawanabe. 2012. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press. doi:10.7551/mitpress/9780262017091.001.0001

[26] Caroline B Terwee, Cecilia AC Prinsen, Alessandro Chiarotto, Marjan J Westerman, Donald L Patrick, Jordi Alonso, Lex M Bouter, Henrica CW De Vet, and Lidwine B Mokkink. 2018. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of life research* 27, 5 (2018), 1159–1170. doi:10.1007/s11136-018-1829-0

[27] Rosalia Tufano, Alberto Martin-Lopez, Ahmad Tayeb, Ozren Dabic, Sonia Haiduc, and Gabriele Bavota. 2025. Deep Learning-based Code Reviews: A Paradigm Shift or a Double-Edged Sword?. In *47th IEEE/ACM International Conference on Software Engineering, ICSE 2025, Ottawa, ON, Canada, April 26 - May 6, 2025*. IEEE, 1640–1652. doi:10.1109/ICSE55347.2025.00060

[28] Stefan Wagner and Marvin Wyrich. 2021. Code comprehension confounders: A study of intelligence and personality. *IEEE Transactions on Software Engineering* 48, 12 (2021), 4789–4801. doi:10.1109/TSE.2021.3127131

[29] Michael Waldman. 1984. Worker allocation, hierarchies and the wage distribution. *The Review of Economic Studies* 51, 1 (1984), 95–109. doi:10.2307/2297707

[30] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, Anders Wesslén, et al. 2012. *Experimentation in software engineering*. Vol. 236. Springer. doi:10.1007/978-3-662-69306-3

[31] Marvin Wyrich and Sven Apel. 2024. Evidence Tetris in the Pixelated World of Validity Threats. In *Proceedings of the 1st IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering*. 13–16. doi:10.1145/3643664.3648203

Received 23 January 2026; accepted 2 April 2026