The Journal of Experimental Education

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

# The Probability of a Robust Inference for External Validity: A Probabilistic Generalizability Index for Randomized Experiments

Tenglong Li, Mingming Chen, Yi Lin & Zixi Chen

**To cite this article:** Tenglong Li, Mingming Chen, Yi Lin & Zixi Chen (08 Dec 2025): The Probability of a Robust Inference for External Validity: A Probabilistic Generalizability Index for Randomized Experiments, The Journal of Experimental Education, DOI: 10.1080/00220973.2025.2599809

**To link to this article:** https://doi.org/10.1080/00220973.2025.2599809

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2025.2599809

Routledge
Taylor & Francis Group

MEASUREMENT, STATISTICS, AND RESEARCH DESIGN



# The Probability of a Robust Inference for External Validity: A Probabilistic Generalizability Index for Randomized Experiments

Tenglong Li[a], Mingming Chen[a], Yi Lin[a], and Zixi Chen[b]

[a]Xi'an Jiaotong-Liverpool University, Suzhou, P.R. China; [b]New York University Shanghai, Shanghai, P.R. China

**ABSTRACT**

External validity is often questionable in empirical research, especially in randomized experiments due to the tradeoff between internal validity and external validity. To quantify the robustness of external validity, one should first conceptualize the part of the target population that cannot be represented by the observed sample (i.e., the unobserved part) and thus the unobserved sample which is thought of as a random sample drawn from the unobserved part of the target population. In this article, we define the probability of a robust inference for external validity, i.e., the PEV, as the probability of rejecting the null hypothesis again based on both the observed and unobserved samples, given a significant result based on the observed sample. Drawing on the relationship between the unobserved sample and the PEV, we propose a six-step procedure for evaluating external validity and illustrate this procedure with an empirical example. We show that the PEV can be interpreted as the statistical power and our analysis of external validity can be translated as power analysis across all plausible forms of the unobserved sample.

**KEYWORDS**

external validity; randomized experiment; robustness indices; causal inference; hypothesis testing; power analysis

## Introduction

Randomized experiments have long been the benchmark for making causal inferences (Cook et al., 2010; Fisher, 1937; Imbens & Rubin, 2015; Morgan & Rubin, 2012; Rubin, 2007, 2008; Shadish et al., 2002, 2011; Thomas, 2016; What Works Clearinghouse, 2014). However, a randomized experiment usually suffers from a nonrepresentative sample or a convenient sample because it could be impractical to randomly sample a human subject and then randomly assign that person to a treatment (Cook, 2002, 2003, 2007; Cronbach, 1975; Imai et al., 2008; O'Muircheartaigh & Hedges, 2014; Olsen et al., 2013; Reichardt & Gollob, 1999; Schneider et al., 2007; Stuart et al., 2001; Tipton, 2014). One prominent example is Borman et al. (2008)., which examined the causal effect of Open Court Reading (OCR) curriculum (National Reading Panel, 2000) by randomly assigning classrooms in each sampled school and each grade to the treatment (OCR) group and the control (non-OCR) group. Their sampled schools were randomly drawn from schools which volunteered for this study, and they might not represent classrooms in non-volunteer schools, potentially limiting the external validity of their inference if the non-volunteer schools responded to OCR differently from the volunteer schools. (Frank et al., 2013). Cook (2002, 2003, 2007) has formalized such risk as the potential tradeoff between internal

**CONTACT** Tenglong Li Tenglong.Li@xjtlu.edu.cn Academy of Pharmacy, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, Jiangsu, 215123, P.R. China

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00220973.2025.2599809.

validity and external validity, i.e., one increase internal validity at the cost of reducing external validity.

The loss of external validity could limit replicability (Avellar et al., 2017; Garcia & Wantchekon, 2010; Hedges, 2013). Considering the difficulty and cost of implementing a randomized experiment, one naturally would (and should) expect the payoff worth the investment, which suggests a significant randomized experiment should be able to generalize widely (Cook et al., 2010; Orr, 2015; Stuart & Rhodes, 2017; Tipton & Peck, 2017). To respond, assessing the evidence regarding external validity has been recommended as an indispensable part of causal research and literature on providing such assessment tools has emerged (e.g., O'Muircheartaigh & Hedges, 2014; Stuart et al., 2001; Tipton, 2014).

Bearing the same goal of indexing the external validity of randomized experiment in mind, this paper approaches it differently. In particular, this article addresses this question by considering the relationship between hypothesis testing and an unobserved sample which would largely mitigate the concern of limited external validity. To do so, the target population is considered to be consisted of the observed part from which the observed sample was taken, and the unobserved part from which no data were sampled but to which one would like to generalize. For example, Borman et al. (2008) conducted a study and reported a significant positive effect of Open Court Reading curriculum on students' reading achievement, based on a sample of classrooms from six schools that volunteered to participate in their randomized trial. To generalize the inference of Borman et al. (2008) to all American schools, one needs to conceptualize the unobserved sample of classrooms which would be potentially randomly drawn from schools which did not volunteer for their randomized trial. Naturally, one would wonder whether Borman et al. could have found a significant positive effect of OCR again had such unobserved sample became available to them. Drawing on this thought, the probability of a robust inference for external validity (henceforth it is abbreviated as the PEV) is proposed as a probabilistic measure of external validity, which is novel for the current literature (Frank et al., 2013; Frank & Min, 2007). Furthermore, we investigate how the unobserved sample is related to the result of hypothesis testing that is founded on both the observed and the unobserved samples, and we establish the relationship between the PEV and the unobserved sample statistics that are sufficient for null hypothesis statistical testing (NHST). This relationship will enable researchers to quantify external validity as the PEV value while varying the unobserved sample across thought experiments, leading to a systemized evaluation of external validity with regard to the unobserved part of a target population.

This paper is organized as follows: The research setting will be firstly outlined and then the unobserved and ideal samples are defined for the purpose of assessing external validity. Drawing on these definitions, the PEV can be formally defined and the relationship between the PEV and the unobserved sample is discussed, for a simple group-mean-difference estimator of average treatment effect. In the example section, the PEV analysis is formalized as a six-step procedure and Borman et al. (2008) is used to illustrate how this analytical procedure is employed to evaluate external validity of a study.

## Research setting and definitions

### *Research setting*

Throughout this paper, we assume a causal inference is made based on a randomized experiment with a non-representative sample from the target population, and in this randomized experiment participants are randomly assigned to either the treatment group or the control group, i.e., there are only two groups in this randomized experiment. Clearly, external validity should be thoroughly evaluated for the above inference. We further assume the above inference is made based on null hypothesis significance testing (NHST), and a significant result (i.e., estimated average

treatment effect) is reported with a debatable external validity. In addition, we assume the average treatment effect is estimated by computing the mean difference between the treatment and control groups of the experiment.[1]

### Definitions

**Definition 1:** The unobserved part of the target population refers to the part of the target population from which the observed sample could not be drawn. The observed part of the target population refers to the part of the target population from which the observed sample was drawn.

**Example:** The unobserved part of the target population of Borman et al. (2008) would be the collection of all classrooms in the non-volunteer schools. The observed part of the target population of Borman et al. (2008) would be the collection of all classrooms in the volunteer schools.

**Definition 2:** The unobserved sample refers to an imaginary random sample which is drawn from the unobserved part of the target population. The ideal sample refers to the combination of the observed sample and the unobserved sample.

**Example:** Figure 1 is created to illustrate the above concepts, based on Borman et al. (2008). In Figure 1, the target population of Borman et al. is partitioned into the observed part (i.e., the collection of all classrooms in the volunteer schools) which is represented by the left half and the unobserved part (i.e., the collection of all classrooms in the non-volunteer schools) which is represented by the right half, separated by the middle vertical dashed line. The observed sample of Borman et al. (2008) is a sample of classrooms drawn from the volunteer schools, and in Figure 1 it is represented by the collection of $O_i$ and $C_j$, where $O_i$ refers to an individual classroom that was randomly assigned to the observed Open Court Reading (OCR) group and $C_j$ refers to an individual classroom that was randomly assigned to the observed control group. The conceptualization of unobserved sample is represented by the two arrows which start from the observed sample (the rectangle on the left with blue shaded circles) from the volunteer schools and end at the unobserved sample (the rectangle on the right with unshaded circles) from the non-volunteer schools. Correspondingly, the unobserved sample of Borman et al. (2008) is an imaginary sample of classrooms which were randomly drawn from the non-volunteer schools and subsequently randomly assigned to the OCR group (denoted by $O_k$) or the control group (denoted by $C_l$). This means the unobserved sample (i.e., the collection of $O_k$ and $C_l$) also has a treatment (the OCR) group and a control group, by following the same sampling and random assignment procedure described in Borman et al. (2008). Lastly, the ideal sample for Borman et al. (2008) consists of all the classrooms in both the observed sample and the unobserved sample, i.e., it is the collection of $O_i, O_k, C_j, C_l$ in the Figure 1.

**Definition 3:** The following observed sample statistics are defined for computing the PEV: (i) $n_t^{ob}$ which denotes the number of the treated subjects in the observed sample (i.e., the number of classrooms that were randomly assigned to the OCR group in the observed sample); (ii) $n_c^{ob}$ which denotes the number of the control subjects in the observed sample (i.e., the number of classrooms that were randomly assigned to the control group in the observed sample); (iii) $\hat{\sigma}_t^2$ and $\hat{\sigma}_c^2$ which denote the variances of the observed outcomes under the treatment and the control, respectively; (iv) $\hat{\delta}^{ob}$ which denotes the estimated average treatment effect based on the observed sample. Moreover, the following parameters are defined based on the unobserved sample: (v) $\hat{\delta}^{un}$ which denotes the estimated average treatment effect based on the unobserved sample; (vi) $\pi_R$ which denotes the relative size of the observed sample in the ideal sample, i.e., $\pi_R = \frac{n^{ob}}{n^{ob}+n^{un}}$ where $n^{ob}$ and $n^{un}$ are the sample sizes of the observed and unobserved samples respectively. It's noteworthy that $\pi_R$ should also represent the relative size of the observed part in

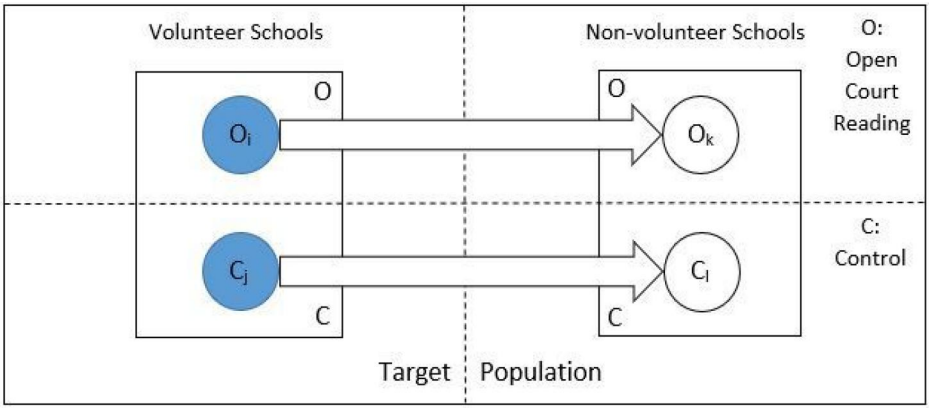


**Figure 1.** The observed and unobserved parts of the target population for Borman et al. (2008).

the target population. For example, $\pi_R$ would be the proportion of the volunteer schools (and arguably schools which are similar to the volunteer schools) in the population of all U.S. schools for Borman et al. (2008).

Frank and Min (2007) noted that one should think about the ideal sample when his inference is based on a randomized experiment whose observed sample is non-representative of the target population, as the bias associated with such inference is induced by the gap between the observed sample and the ideal sample one would use for inference, which by definition is the unobserved sample. The spirit of their approach is to reconsider the inference as if the ideal sample were available. Following this spirit, this article intends to achieve two goals: First, this article seeks to define the PEV as a probabilistic robustness index that would allow users to quantify the probability of a robust significant inference had the ideal sample became available, in the context of null hypothesis significance testing (NHST). Second, this article seeks to uncover the relationship between the PEV and the unobserved sample (specifically $\hat{\delta}^{un}$ and $\pi_R$), which would allow researchers to study the general pattern of how the PEV varies conditional on a plausible range of values of $\hat{\delta}^{un}$ and $\pi_R$ that characterize the unobserved sample.

## The probability of a robust inference for external validity

The PEV is defined under the theoretical framework of null hypothesis significance testing (NHST) which is commonly employed to determine if a treatment effect indeed exists. Specifically, the null hypothesis $H_0 : \delta = 0$ is tested against the alternative hypothesis $H_a : \delta \neq 0$.[2] The PEV becomes meaningful after the null hypothesis is rejected for the observed sample, i.e., the observed test statistic surpasses the critical value/decision threshold. Given the unobserved sample is expected to be different from the observed sample, the PEV is motivated by the question "would the null hypothesis be rejected again if the unobserved sample was available and included for NHST", i.e., the PEV aims to evaluate external validity by considering the unobserved sample and its relationship with NHST.

Formally, the probability of a robust inference for external validity (PEV) is defined as follows for a significant observed treatment effect estimate $\hat{\delta}^{ob}$, based on the ideal sample:

$$\mathbf{PEV} = \boldsymbol{P}(\hat{\boldsymbol{\delta}}^{\boldsymbol{id}} \ \textbf{is significant} \,|\, \hat{\boldsymbol{\delta}}^{\boldsymbol{ob}} \ \textbf{is significant}) \tag{1}$$

The PEV quantifies the likelihood of rejecting the null hypothesis $H_0$ again based on the ideal sample, conditional on the fact that the same null hypothesis has been already rejected based on the observed sample, assuming the unobserved sample is procured as specified in the definition

2. In addition, $\hat{\delta}^{id}$ (estimated average treatment effect based on the ideal sample) and $\hat{\delta}^{ob}$ should have the same sign, to ensure the conclusions drawn from rejecting $H_0$ are consistent for both the ideal and observed samples (Frank et al., 2013). For a point estimate of average treatment effect, NHST is likely built on either normal or student's t-distribution, and the PEV has the following relationship with the T-ratio $T = \frac{\hat{\delta}^{id}}{se(\hat{\delta}^{id})}$:

If $\hat{\delta}^{ob}$ is significantly positive:

$$\boldsymbol{\Phi}^{-1}(\boldsymbol{PEV}) = \boldsymbol{T} - \boldsymbol{C} \tag{2}$$

If $\hat{\delta}^{ob}$ is significantly negative:

$$\boldsymbol{\Phi}^{-1}(\boldsymbol{PEV}) = \boldsymbol{C} - \boldsymbol{T} \tag{3}$$

where C is the critical value and $\Phi^{-1}$ is the inverse of the standard normal CDF. From (2) and (3), it is obvious that the PEV is the statistical power of testing the null hypothesis $H_0 : \delta = 0$ versus the alternative hypothesis $H_a : \delta \neq 0$ based on the ideal sample, as one can replace the PEV with the statistical power for the above two equations to derive its formal definition. It's noteworthy that the Eqs. (2) and (3) are only approximately true for studies with small sample sizes, and C typically corresponds to the specified level of significance $\alpha$, i.e., C could be written as $Z_{1-\alpha/2}$ for a significantly positive $\hat{\delta}^{ob}$ or $Z_{\alpha/2}$ for a significantly negative $\hat{\delta}^{ob}$ (Li & Frank, 2022). For example, for a significantly positive $\hat{\delta}^{ob}$ and $\alpha = 0.05$, C typically would be 1.96. It's also possible that C is pragmatically chosen as a fixed value based on transaction cost and/or policy implications, rather than purely based on level of significance (Frank et al., 2013). For an instance, one may prefer to use a fixed effect size (like 1) as the decision threshold of whether Open Court Reading (OCR) group is significantly better than the control group, in the context of Borman et al. (2008), given the substantial cost of replacing existing reading curriculums with the OCR at many schools. In this case, one can transform the decision threshold into the critical value C and still be able to evaluate external validity with the PEV.

## The relationship between the PEV and the unobserved sample

The PEV essentially depends on the unobserved sample, specifically on the parameter $\pi_R$ which defines the relative size of the observed part in the target population (also the relative size of the observed sample in the ideal sample) and the parameter $\hat{\delta}^{un}$ which defines the estimated average treatment effect based on the unobserved sample. The relationship between the PEV and the two parameters is derived as follows:

**Theorem 1:** The probit link of the PEV is a function of $\pi_R$ and $\hat{\delta}^{un}$, given the observed sample statistics $\hat{\sigma}_t^2, \hat{\sigma}_c^2, n_t^{ob}, n_c^{ob}, \hat{\delta}^{ob}$ and the critical value C for deciding $\hat{\delta}^{ob}$ is statistical significant. Specifically, if $\hat{\delta}^{ob}$ is significantly positive, this function is expressed as:

$$\boldsymbol{\Phi}^{-1}(\boldsymbol{PEV}) = \frac{1}{\sqrt{\frac{\hat{\boldsymbol{\sigma}}_t^2}{\boldsymbol{n}_t^{ob}} + \frac{\hat{\boldsymbol{\sigma}}_c^2}{\boldsymbol{n}_c^{ob}}}} \left[\boldsymbol{\pi}_R^{0.5}\left(\hat{\boldsymbol{\delta}}^{ob} - \hat{\boldsymbol{\delta}}^{un}\right) + \boldsymbol{\pi}_R^{-0.5}\hat{\boldsymbol{\delta}}^{un}\right] - \boldsymbol{C} \tag{4}$$

If $\hat{\delta}^{ob}$ is significantly negative, this function is expressed as:

$$\boldsymbol{\Phi}^{-1}(\boldsymbol{PEV}) = \boldsymbol{C} - \frac{1}{\sqrt{\frac{\hat{\boldsymbol{\sigma}}_t^2}{\boldsymbol{n}_t^{ob}} + \frac{\hat{\boldsymbol{\sigma}}_c^2}{\boldsymbol{n}_c^{ob}}}} \left[\boldsymbol{\pi}_R^{0.5}\left(\hat{\boldsymbol{\delta}}^{ob} - \hat{\boldsymbol{\delta}}^{un}\right) + \boldsymbol{\pi}_R^{-0.5}\hat{\boldsymbol{\delta}}^{un}\right] \tag{5}$$

(Proof in Appendix).

Theorem 1 is derived from the relationship between the PEV and the T-ratio $\frac{\hat{\delta}^{id}}{se(\hat{\delta}^{id})}$ based on the ideal sample (i.e., the Eqs. (2) and (3)). Conditional on the normality assumption for testing $H_0 : \delta = 0$ versus $H_a : \delta \neq 0$ for the ideal sample, the Eqs. (4) and (5) can be directly derived from the definitions of the PEV (i.e, the Eq. (1)). The unobserved part is also assumed to have approximately the same variance for the outcome as the observed part in the target population, so that $\hat{\sigma}_t^2$ and $\hat{\sigma}_c^2$ can be treated as unbiased estimates for the variances of the outcome under the treatment and the control conditions. To determine the value of the PEV, one needs to first conceptualize the values of $\pi_R$ and $\hat{\delta}^{un}$ for the unobserved sample, that is, one must conceptualize the size of the unobserved sample as well as estimated average treatment effect based on the unobserved sample in a well-defined thought experiment. For the example of Open Court Reading (OCR) curriculum, $\pi_R$ characterizes the proportion of schools that can be represented by the schools sampled by Borman et al. in the United States, and equivalently it also should be the ratio between the size of the observed sample size and the size of the ideal sample conceptualized based on Borman et al. Moreover, $\hat{\delta}^{un}$ characterizes the estimated average treatment effect of OCR based on the unobserved sample, i.e., one needs to conceptualize a thought experiment where schools in the unobserved sample are randomly assigned to the OCR and control groups and the mean reading scores of the two groups are compared to obtain the estimate. Theorem 1 also has a Bayesian interpretation where the prior distribution is defined by $\pi_R$ and $\hat{\delta}^{un}$ (i.e., the unobserved sample) and the likelihood is defined by the observed sample (Frank & Min, 2007; Hoff, 2009; Li, 2018).

The relationship between the PEV and the unobserved sample can be extended to regression models, where one not only considers the outcomes under both treatment conditions in both the observed and unobserved samples but also the covariates in both the observed and unobserved samples. Therefore, this relationship would be much more complicated than the one stated by Theorem 1, given one needs to derive the ideal sample variances and covariances based on the sample means, variances as well as covariances in both the observed and unobserved samples for all variables involved, in order to compute the treatment effect based on regression. We refer readers with interests in regression models to the Appendix for technical details.

## Example: the effect of Open Court Reading curriculum on reading achievement

### *Overview*

The Open Court Reading (OCR) program is a curriculum that is rooted in research-based practices and has been in the market for a long time and widely adopted by many districts and schools. Although OCR is potentially a beneficial program because it responds to recommendations from research that focused on developing early reading skills, its effect had initially not been assessed and confirmed by a randomized experiment. Seeing this, Borman et al. (2008) designed a multisite cluster randomized experiment and randomly drew 6 schools from the schools had contacted and shown their interest to SRA/McGraw Hill, the publisher of the OCR curriculum. Those 6 schools came from six different states (Florida, Georgia, Idaho, Indiana, North Carolina and Texas) and were geographically, ethnically and socioeconomically representative of schools in the US. Subsequently, Borman et al. defined a block as a single grade of one sampled school and within each block classrooms were randomly assigned to the OCR group or the control group (business as usual). The final sample of Borman et al. (2008) included five schools (the Georgia school dropped out) and 49 classrooms of which 27 classrooms were assigned to the OCR group. Controlling for the pretest scores and block membership, Borman et al. (2008) estimated the effect of OCR as 7.95 (on reading composite scores) which was statistically significant and went on to conclude that "the outcomes from these analyses provided not only evidence of the

promising 1-year effects of OCR on students' reading outcomes but also suggest that these effects may be replicated across varying contexts with rather consistent and positive results".

Ideally, the findings of Borman et al. (2008) imply that, the estimated effect of OCR would be around 7.95 if one were to conduct a large-scale completely randomized experiment elsewhere or collected a nationally representative sample. At the very least, one should be able to conclude a significantly positive effect of OCR if the experiment were replicated. However, such claims are not necessarily warranted as the external validity of Borman et al. (2008) is debatable, given all sampled schools of Borman et al. (2008) were drawn from the volunteer schools. Specifically, Frank et al. (2013) note that the effect of the OCR program might not have been as large as the one reported by Borman et al. (2008) had their study been conducted in non-volunteer schools, possibly because volunteer schools might have anticipated the OCR program was highly effective for them given their student composition. If this were the case, the effect of the OCR curriculum would be overestimated and the inference drawn by Borman et al. (2008) may not be valid for all targeted schools.

Following the above argument, the external validity of Borman et al. (2008) is evaluated next by quantifying its robustness as the PEV proposed in this paper. Particularly, the analytical procedure includes six steps: (i) get the required sample statistics, (ii) choose the critical value C, (iii) obtain the relationship between the PEV and the focal parameters, i.e., $\pi_R$ and $\hat{\delta}^{un}$, (iv) specify a joint distribution about $\pi_R$ and $\hat{\delta}^{un}$, (v) calculate the expected value and confidence interval for the PEV, (vi) evaluate the robustness based on the expected value of the PEV.

### *Quantifying the robustness of the inference of Borman et al. (2008)*

i. Get the required sample statistics: The following sample statistics are needed for computing the PEV and can be extracted from Borman et al. (2008): $\hat{\delta}^{ob} = 7.95, \hat{\sigma}_t^2 = 45, \hat{\sigma}_c^2 = 45, n_t^{ob} = 27, n_c^{ob} = 22$. (Frank et al., 2013).

ii. Choose critical value C: Borman et al. (2008) reported that the Open Court Reading (OCR) curriculum had a significantly positive effect on reading composite scores (i.e., $\hat{\delta}^{ob} = 7.95$). Therefore, we decided to set $C = 1.96$ which corresponds to $\alpha = 0.05$ for testing the null hypothesis $H_0 : \delta = 0$, where $\delta$ represents the true average treatment effect of OCR on reading achievement.

iii. Obtain the relationship between the PEV and the focal parameters: By plugging the observed statistics collected in the first step and the value of C in the second step into the Eq. (4), the PEV is a probit function of the proportion of American schools that can be represented by the observed sample of Borman et al. (i.e., $\pi_R$) and the estimated average treatment effect of OCR on reading composite scores based on the unobserved sample (i.e., $\hat{\delta}^{un}$) as follows:

$$\mathbf{\Phi}^{-1}(\boldsymbol{PEV}) = \frac{1}{\sqrt{3.71}}\left[\boldsymbol{\pi_R^{0.5}}\left(7.95 - \hat{\boldsymbol{\delta}}^{\boldsymbol{un}}\right) + \boldsymbol{\pi_R^{-0.5}}\hat{\boldsymbol{\delta}}^{\boldsymbol{un}}\right] - 1.96 \tag{6}$$

iv. Specify a joint distribution about $\pi_R$ and $\hat{\delta}^{un}$: At this stage, one should form a joint distribution about the two focal parameters $\pi_R$ and $\hat{\delta}^{un}$, typically via detailed thought experiments where one would ask himself/herself that how representative the observed sample can be (in other words, what is the relative size of the observed part of the target population) and what the average treatment effect estimates would be based on the unobserved sample from the unobserved part of the target population. Typically, one should first choose the ranges of $\pi_R$ and $\hat{\delta}^{un}$ based on domain knowledge and literature, and those ranges should only cover the unfavorable scenarios where the values of $\pi_R$ and $\hat{\delta}^{un}$ would make the original inference less robust. The distribution of $\pi_R$ and $\hat{\delta}^{un}$ should also be chosen with explicit justifications, however, as a rule of thumb, one can choose a joint uniform distribution defined by the ranges of $\pi_R$ and $\hat{\delta}^{un}$.

For Borman et al. (2008), we carried out the thought experiments by conceptualizing the proportion of all American schools that can be represented by the observed sample (i.e., the five volunteered schools) and the possible estimates of the average treatment effect of OCR based on the unobserved sample, which can be thought as an imaginary random sample of schools from all non-volunteered schools in the U.S. The range of the proportion of all American schools that can be represented by the observed sample could be rather wide, since there is no direct literature or evidence to justify the choices of such proportions. We chose [0.2, 0.8] as the range of $\pi_R$, that is, we think the observed sample can represent at least 20% and at most 80% of all American schools. For the purpose of illustration, we compared three joint distributions which mainly differed in the range of $\hat{\delta}^{un}$ as well as its distribution.

The first joint distribution: Although Borman et al. (2008) concluded that the effect of OCR on reading achievement was significantly positive, a follow-up analysis found that OCR might have differential effects in the subgroups of students defined by gender, SES, race, grade or English language learner status (Vaden-Kiernan et al., 2018). Specifically, Vaden-Kiernan et al. (2018) reported that OCR had negative effects in some of those subgroups, and the effect size could be as large as −0.189 for the subgroup of their Year 2 students who did not receive free or reduced-price lunch. This effect size (−0.189) was tantamount to an effect of −3.33 in terms of raw scores, and therefore we chose the range of $\hat{\delta}^{un}$ as [−4, 8] to include all possible treatment effects implied by Borman et al. (2008) and Vaden-Kiernan et al. (2018). As a result, we assumed $\pi_R$ followed the uniform distribution in [0.2, 0.8] and $\hat{\delta}^{un}$ followed the uniform distribution in [−4, 8].

The second joint distribution: We intended to narrow the range of $\hat{\delta}^{un}$ to illustrate the impact of tighter bounds about $\hat{\delta}^{un}$ on the inference regarding the PEV. Given the strong positive effect reported by Borman et al. (2008), it might be reasonable to think OCR had less sizeable (but still positive) effect on reading achievement under most scenarios. Therefore, we assumed $\pi_R$ followed the uniform distribution in [0.2, 0.8] and $\hat{\delta}^{un}$ followed the uniform distribution in [0, 8].

The third joint distribution: We recognized that most empirical distributions were not uniform and likely to be modeled parametrically (such as normal distribution). Therefore, we assumed $\hat{\delta}^{un}$ followed the normal distribution with mean as 2 and standard deviation as 3 (thus the 95% confidence interval for a random value of $\hat{\delta}^{un}$ was [−4, 8], identical to the range of the first joint distribution), holding the distribution of $\pi_R$ the same as the previous two joint distributions.

v. Calculate the expected value and confidence interval for the PEV: Figure 2 is a contour plot that depicts the level of the PEV based on the first belief, i.e., a joint uniform distribution for $\pi_R$ in [0.2, 0.8] and $\hat{\delta}^{un}$ in [−4, 8]. The second belief is also indicated with a vertical dashed line at $\hat{\delta}^{un} = 0$ to delimit the ranges of the joint uniform distribution. The distribution of the PEV was approximated via Monte Carlo sampling, i.e., we first drew random values of $\pi_R$ as well as $\hat{\delta}^{un}$ from their joint distribution and then computed the PEV based on the Eq. (6). Consequently, we obtained the expected value of the PEV as 0.83 and the 95% confidence interval as [0.04, 1], which suggests the chance that the inference of Borman et al. (2008) is robust for external validity is expected to be 83% based on the first joint distribution. Likewise, the expected value of the PEV was calculated as 0.97 and its 95% confidence interval was [0.77, 1] based on the second joint distribution. It is clear that, by having tighter bounds on $\hat{\delta}^{un}$ (from [−4, 8] to [0, 8]), the level of confidence about the robustness of the inference of Borman et al. (2008) was significantly increased. Lastly, the expected value of the PEV was 0.86 and its 95% confidence interval was [0.09, 1] conditional on the third joint distribution, which were slightly different from the results acquired based on the first joint distribution.

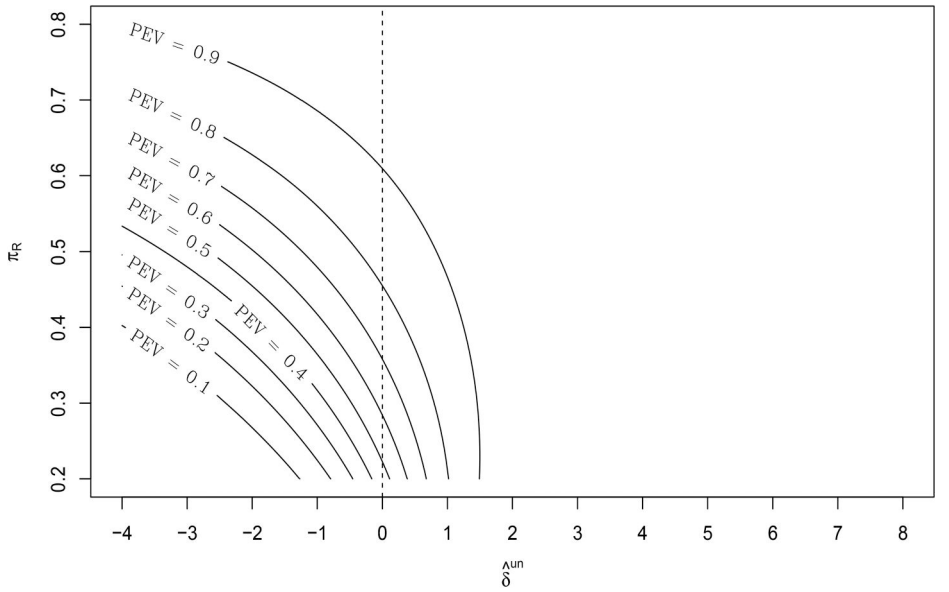


**Figure 2.** The contour plot of the PEV based on the first joint distribution of $\pi_R$ and $\hat{\delta}^{un}$ with the x axis representing $\hat{\delta}^{un}$ and the y axis representing $\pi_R$. The first joint distribution is defined based on the belief that the effect of OCR should be in the range [−4, 8] in the unobserved sample and the observed sample of Borman et al. (2008) represents about 20% to 80% of their target population. The second joint distribution is built on the belief that the effect of OCR should be non-negative in the unobserved sample and the observed sample of Borman et al. (2008) represents about 20% to 80% of their target population. Therefore, the right region (separated by the vertical dashed line) represents the second joint distribution of $\pi_R$ and $\hat{\delta}^{un}$.

vi. Evaluate the robustness based on the expected value of the PEV: As the PEV is essentially the statistical power for testing the null hypothesis $H_0 : \delta = 0$ versus the alternative hypothesis $H_a : \delta \neq 0$ based on the ideal sample, we checked to see if the expected value of the PEV exceeded 0.8 which was a common threshold for strong statistical power (Cohen, 1988, 1992). Therefore, we concluded that the inference of Borman et al. (2008) was robust for external validity based on the first belief/joint distribution of $\pi_R$ and $\hat{\delta}^{un}$ as the corresponding expected value of the PEV was 0.83. Likewise, based on the second and third beliefs/joint distributions of $\pi_R$ and $\hat{\delta}^{un}$, we concluded again that the inference of Borman et al. (2008) was robust for external validity as the corresponding expected values of the PEV were 0.97 and 0.86 respectively. It's noteworthy that the above conclusions may not hold for a different belief/joint distribution (of $\pi_R$ and $\hat{\delta}^{un}$) and/or a different threshold for strong statistical power.

Conceptually, the PEV calculations described in the step 4 and 5 can be thought of as testing the null hypothesis $H_0 : \delta = 0$ versus the alternative hypothesis $H_a : \delta \neq 0$ iteratively conditional on random values of $\pi_R$ and $\hat{\delta}^{un}$ drawn from their joint distribution, given one can straightforwardly compute the PEV with the values of $\pi_R$ and $\hat{\delta}^{un}$ according to Theorem 1. This is tantamount to a comprehensive power analysis regarding the null hypothesis $H_0 : \delta = 0$ versus the alternative hypothesis $H_a : \delta \neq 0$ in the ideal sample, across all plausible values of $\pi_R$ and $\hat{\delta}^{un}$. Such idea of power analysis is illustrated by Figure 3 where $\pi_R$ is assumed to be 0.46, a threshold derived by Frank et al. (2013). We caution readers that the PEV results from the interplay between $\pi_R$ and $\hat{\delta}^{un}$, as illustrated in Figure 2, and thus it's possible for a less consistent but more representative (i.e., larger gap between $\hat{\delta}^{un}$ and $\hat{\delta}^{ob}$ with a larger $\pi_R$) experiment to have a higher PEV than a more consistent but less representative experiment (i.e., smaller gap between $\hat{\delta}^{un}$ and $\hat{\delta}^{ob}$ with a smaller $\pi_R$). It's necessary to point out that one's belief/joint distribution about $\pi_R$ and $\hat{\delta}^{un}$ should be fairly inclusive and justified based on literature/domain knowledge, in order

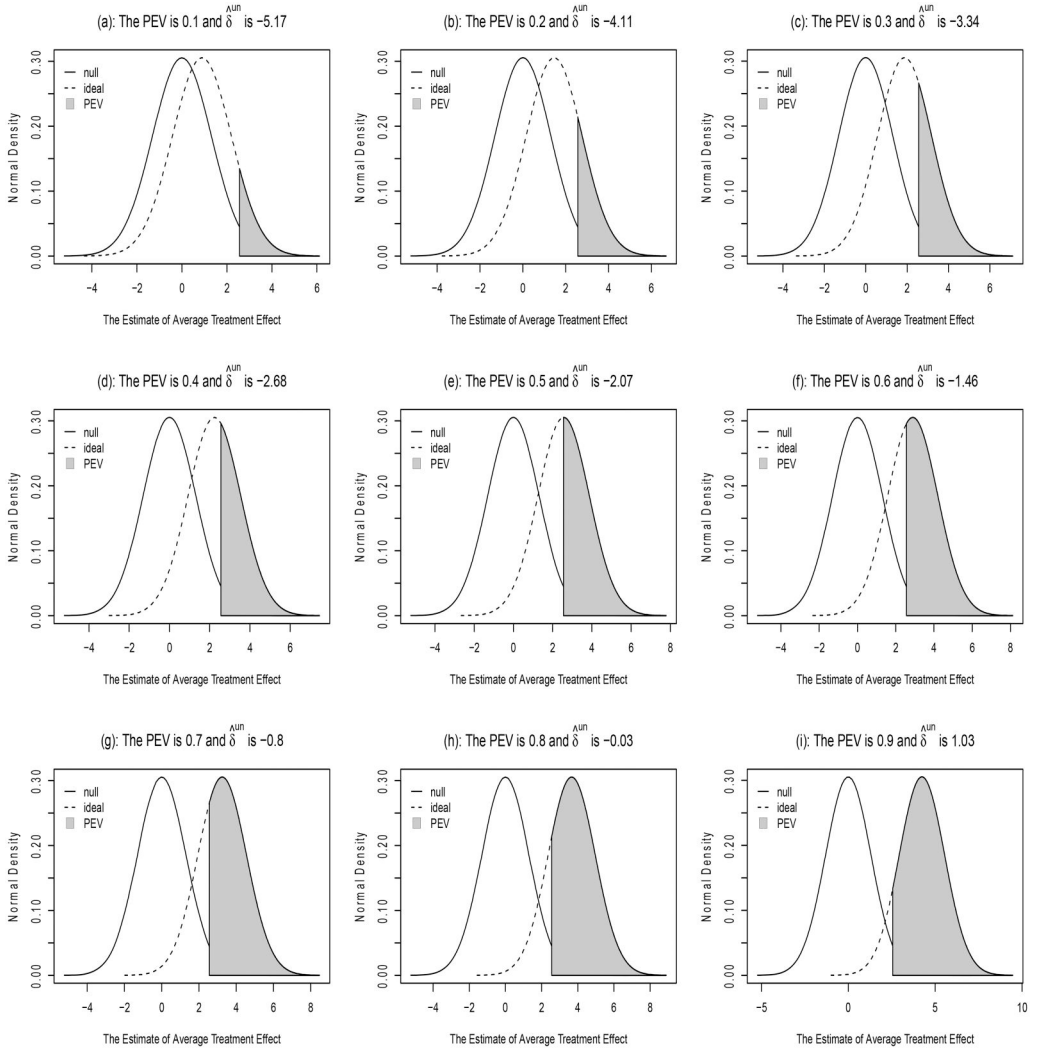


**Figure 3.** The relationship between the PEV and testing the null hypothesis based on the ideal sample for Borman et al. (2008), assuming $\pi_R = 0.46$. The solid curve represents the null hypothesis: $\delta = 0$ and the dashed curve represents the alternative hypothesis: $\delta \neq 0$. The grey shaded area symbolizes the statistical power for the above hypothesis testing, i.e., it represents the PEV of Borman et al. (2008).

to cover all possible unfavorable scenarios (characterized by $\pi_R$ and $\hat{\delta}^{un}$) that could compromise external validity. Correspondingly, an inference would be deemed robust for external validity if it on average possesses strong statistical power over all potential values of $\pi_R$ and $\hat{\delta}^{un}$.

## Discussion

This article recasts the problem of evaluating the external validity for a randomized experiment as a missing data problem where the unobserved sample is conceptualized as a random sample from the unobserved part of the target population. The same randomization procedure is also thought to be carried out in the unobserved sample to form the treatment and control groups, and thus the ideal sample, which refers to the combination of the observed sample and the unobserved sample, can be perceived as a random sample from the whole target population and

sufficient for claiming strong external validity. The probability of a robust inference for external validity, i.e., the PEV, is then defined as the probability of rejecting the null hypothesis $H_0 : \delta = 0$ (against $H_a : \delta \neq 0$) based on the ideal sample, provided the same null hypothesis has already been rejected for the observed sample. Holding the observed sample fixed, the PEV is proved to be a function of the parameters $\pi_R$ and $\hat{\delta}^{un}$, which characterize the unobserved sample. The process of evaluating external validity using the PEV is formalized as a six-step procedure, which essentially is power analysis across all plausible values of $\pi_R$ and $\hat{\delta}^{un}$ regarding the null hypothesis $H_0 : \delta = 0$ versus the alternative hypothesis $H_a : \delta \neq 0$. An inference that is robust for external validity should on average have a strong statistical power across all plausible values of $\pi_R$ and $\hat{\delta}^{un}$. We applied this six-step procedure to Borman et al. (2008) and found its inference was robust for external validity conditional on two different beliefs/joint distributions of $\pi_R$ and $\hat{\delta}^{un}$.

Other approaches/indices have been proposed for external validity evaluation. One notable example in the field of education is the robustness indices by Frank and Min (2007) and Frank et al. (2013), which focus on what the unobserved sample needs to be so that an inference would be invalidated due to limited external validity. As in this article, their robustness indices emphasize the necessity of conceptualizing the unobserved sample and thus preparing the ideal sample for inference. Probability of replication (Killeen, 2005) evaluates the probability of replicating a previous significant finding in the context of hypothesis testing, and thus it can be used to inform whether an inference is robust for external validity, provided a different sample is drawn for the purpose of study replication. Similar to probability of replication, reproducibility probability (Shao & Chow, 2002) is the estimated power of future trial based on the data from past trial(s), thus conceptually the reproducibility probability should be closely connected to the PEV as the future trial can be thought of as the unobserved sample (and the power pertains to testing the hypotheses based on the unobserved sample rather than the ideal sample). The generalizability index (O'Muircheartaigh & Hedges, 2014; Stuart et al., 2001; Tipton, 2014) is built on the definition of sampling propensity scores based on a key assumption of unconfounded sample selection and the comparison of sampling propensity scores between the observed sample and the target population. It further quantifies the distance between the sampling propensity score distributions in the population and the sample. The s-value is similarly motivated by the sensitivity of focal parameter to distributional changes, and it can be easily adapted for external validity evaluation (Gupta & Rothenhäusler, 2023).

The PEV, though shares similarities with the aforementioned indices, has its own unique features and thoughts evolved from the literature (Frank et al., 2013; Frank & Min, 2007; Li, 2018; Li et al., 2024; Li & Frank, 2022). Our work is founded on Frank's robustness analysis framework, specifically the conceptualization of the unobserved sample and the ideal sample (Frank & Min, 2007), as well as the definition of robust inference described in the third section (Frank et al., 2013). Centered on the philosophy of prompting one to characterize the unobserved sample for evaluating robustness of an inference, a distinct feature of Frank's robustness analysis framework, Li (2018) developed the probabilistic robustness indices for both internal validity and external validity, namely PIV (Li et al., 2024; Li & Frank, 2022) and PEV, and formalized the probabilistic robustness indices with frequentist (as statistical power) and Bayesian (as Bayesian sensitivity analysis) interpretations. Inherited from Frank's robustness analysis framework (Frank et al., 2013; Frank & Min, 2007; Li, 2018), the PEV asks one to characterize the unobserved sample (using $\pi_R$ and $\hat{\delta}^{un}$) in quantifying the robustness of external validity, which marks the main distinction between the PEV and the probability of replication (PR) as PR does not require one to conceptualize and characterize the unobserved sample that is unavoidable in assessing external validity. Compare with the generalizability index, the PEV does not require the unconfounded sample selection assumption given it directly models unobserved sample and its influence on hypothesis testing. Therefore, the PEV is less likely subject to the bias incurred by the unconfounded sample selection assumption and the estimation of sampling propensity scores. Lastly, from Bayesian

perspective, the PEV is a posterior probability and reproducibility probability (RP) is a posterior predictive probability, which makes RP focus on the impact of the unobserved sample on future hypothesis testing rather than on current hypothesis testing (i.e., considers the observed sample like the PEV does). Some similar work on robustness of external validity is also notable (Devaux & Egami, 2022; Jeong & Namkoong, 2020; Spini, 2021).

The scholarly significance of this study manifests in three aspects: First of all, this study helps promote critical thinking as well as scientific debate about external validity as one can evaluate external validity based on the domain knowledge and/or personal belief in the terms of a probabilistic index (i.e., the PEV) *via* an open process (Li & Frank, 2022). As shown earlier, external validity is evaluated mainly through the expected value of the PEV based on plausible values of $\pi_R$ and $\hat{\delta}^{un}$ that characterize the potential unobserved samples. This propels one to conduct thought experiments about the observed and unobserved parts of the target population. For example, one should conceptualize the proportion of American schools that can be represented by the sample of Borman et al. (2008) and the effect of OCR had the unobserved sample been drawn from the American schools that cannot be represented by the sample of Borman et al. (2008), with the same randomization process. Drawing on this thought, researchers are recommended to report both the ideal sample statistics (i.e., the treatment effect estimate and its standard error) under the assumptions they made and the PEV resulting from the open process demonstrated in this paper, to facilitate transparent scientific debate. Second, the PEV can be interpreted as the statistical power of testing the null hypothesis $H_0 : \delta = 0$ versus the alternative hypothesis $H_a : \delta \neq 0$ based on the ideal sample which is representative of the whole target population. The PEV also has an intuitive Bayesian interpretation that is tantamount to Bayesian sensitivity analysis (i.e., a fixed likelihood with varying prior built on the unobserved sample), which potentially appeals to Bayesian thinkers. Third, the PEV is rooted in decision making as both the decision threshold for the PEV and the critical value C can be pragmatically chosen based on the transaction costs for treatments in comparison.

Even though the PEV is a useful tool of quantifying the robustness of external validity, it has limitations in three aspects: First, it only addresses bias due to nonrandom sampling error and cannot inform bias due to other sources like measurement error or violation of SUTVA (Li & Lawson, 2024; Rubin, 1980, 1990). Therefore, the PEV is only appropriate for evaluating external validity when the observed sample is suspected to be non-representative of the whole target population. Second, the PEV does not inform true effect or model validity for a particular study, rather it is used to indicate under what circumstances and to what degree a significant finding from hypothesis testing can still hold based on one's belief about the unobserved sample which is incorporated to address external validity. Third, our approach requires one to specify a joint distribution of $\pi_R$ and $\hat{\delta}^{un}$, which sometimes could be inaccessible based on domain knowledge/literature. In this case, using an unjustified subjective belief haphazardly may endanger our proposed procedure for external validity evaluation and nullify the decision regarding the external validity of a study. It's recommended that one builds empirical distributions of $\pi_R$ and $\hat{\delta}^{un}$ based on meta-analysis or prior datasets when possible (such as mixture models) to reduce the risk of over-simplifying the specification of those parameters.

## Notes

1. We also consider regression model with additional controls and related theoretical results are provided in the appendix. The analytical framework (including the definitions, the PEV and the procedure) remains largely unchanged in this case.
2. Our framework can be easily modified for non-zero constants or one-sided hypothesis.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## References


Avellar, S. A., Thomas, J., Kleinman, R., Sama-Miller, E., Woodruff, S. E., Coughlin, R., & Westbrook, T. P. R. (2017). External validity: The next step for systematic reviews? *Evaluation Review*, *41*(4), 283–325. https://doi.org/10.1177/0193841X16665199

Borman, G. D., Dowling, B. M., & Schneck, C. (2008). A multi-site cluster randomized field trial of open court reading. *Educational Evaluation and Policy Analysis*, *30*(4), 389–407. https://doi.org/10.3102/0162373708326283

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates. 2.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199. https://doi.org/10.3102/01623737024003175

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *The ANNALS of the American Academy of Political and Social Science*, *589*(1), 114–149. https://doi.org/10.1177/0002716203254764

Cook, T. D. (2007). Randomized experiments in education: Assessing the objections to doing them. *Economics of Innovation and New Technology*, *16*(5), 331–355. https://doi.org/10.1080/10438590600982335

Cook, T. D., Scriven, M., Coryn, C. L., & Evergreen, S. D. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation*, *31*(1), 105–117. https://doi.org/10.1177/1098214009354918

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127. https://doi.org/10.1037/h0076829

Devaux, M., & Egami, N. (2022). Quantifying robustness to external validity bias. *SSRN 4213753*.

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd.

Frank, K. A., & Min, K. (2007). Indices of robustness for sample representation. *Sociological Methodology*, *37*(1), 349–392. https://doi.org/10.1111/j.1467-9531.2007.00186.x

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's Causal Model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437–460. https://doi.org/10.3102/0162373713493129

Gupta, S., & Rothenhäusler, D. (2023). The s-value: Evaluating stability with respect to distributional shifts. *Advances in Neural Information Processing Systems*, *36*, 72058–72070.

Garcia, M. F., & Wantchekon, L. (2010). Theory, external validity, and experimental inference: Some conjectures. *The Annals of the American Academy of Political and Social Science*, *628*(1), 132–147.

Hedges, L. V. (2013). Recommendations for practice: Justifying claims of generalizability. *Educational Psychology Review*, *25*(3), 331–337. https://doi.org/10.1007/s10648-013-9239-x

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *171*(2), 481–502. https://doi.org/10.1111/j.1467-985X.2007.00527.x

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Jeong, S., & Namkoong, H. (2020). *Assessing external validity over worst-case subpopulations.* arXiv preprint arXiv: 2007.02411.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*(5), 345–353. https://doi.org/10.1111/j.0956-7976.2005.01538.x

Li, T. (2018). *The Bayesian Paradigm of robustness indices of causal inferences* [Unpublished doctoral dissertation]. Michigan State University.

Li, T., & Frank, K. (2022). The probability of a robust inference for internal validity. *Sociological Methods & Research*, *51*(4), 1947–1968. https://doi.org/10.1177/0049124120914922

Li, T., & Lawson, J. (2024). A generalized bootstrap procedure of the standard error and confidence interval estimation for inverse probability of treatment weighting. *Multivariate Behavioral Research*, *59*(2), 251–265. https://doi.org/10.1080/00273171.2023.2254541

Li, T., Frank, K. A., & Chen, M. (2024). A conceptual framework for quantifying the robustness of a regression-based causal inference in observational study. *Mathematics*, *12*(3), 388. https://doi.org/10.3390/math12030388

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263–1282. https://doi.org/10.1214/12-AOS1008

National Reading Panel (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. (NIH Publication No.00-4769)*. U.S. Government Printing Office.

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *63*(2), 195–210. https://doi.org/10.1111/rssc.12037

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management: [the Journal of the Association for Public Policy Analysis and Management]*, *32*(1), 107–121. https://doi.org/10.1002/pam.21660

Orr, L. L. (2015). 2014 Rossi award lecture: Beyond internal validity. *Evaluation Review*, *39*(2), 167–178. https://doi.org/10.1177/0193841X15573659

Reichardt, C. S., & Gollob, H. F. (1999). Justifying the use and increasing the power of at test for a randomized experiment with a convenience sample. *Psychological Methods*, *4*(1), 117–128. https://doi.org/10.1037/1082-989X.4.1.117

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association*, *75*(371), 591–593. https://doi.org/10.2307/2287653

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*(4), 472–480.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*(484), 1350–1353. https://doi.org/10.1198/016214508000001011

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational design*. American Educational & Reseach Association.

Shao, J., & Chow, S. C. (2002). Reproducibility probability in clinical trials. *Statistics in Medicine*, *21*(12), 1727–1742. https://doi.org/10.1002/sim.1177

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, *16*(2), 179–191. https://doi.org/10.1037/a0023345

Spini, P. E. (2021). Robustness, heterogeneous treatment effects and covariate shifts. arXiv Preprint, arXiv:2112.09259.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2001). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Stuart, E. A., & Rhodes, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation Review*, *41*(4), 357–388. https://doi.org/10.1177/0193841X16660663

Thomas, G. (2016). After the gold rush: Questioning the "gold standard" and reappraising the status of experiment and randomized controlled trials in education. *Harvard Educational Review*, *86*(3), 390–411. https://doi.org/10.17763/1943-5045-86.3.390

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501. https://doi.org/10.3102/1076998614558486

Tipton, E., & Peck, L. R. (2017). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review*, *41*(4), 326–356. https://doi.org/10.1177/0193841X16655656

Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., Sullivan, K., Ruiz de Castilla, V., Grace, G., Rodriguez, D., Henry, C., Long, T., & Hughes Jones, D. (2018). Findings from a multiyear scale-up effectiveness trial of Open Court Reading. *Journal of Research on Educational Effectiveness*, *11*(1), 109–132. https://doi.org/10.1080/19345747.2017.1342886

What Works Clearinghouse (2014). *Procedures and standards handbook* (version 3.0). Retrieved June 22, 2018, from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf