Check for updates

Received: 7 August 2019 | Revised: 15 September 2021 | Accepted: 6 December 2021

DOI: 10.1111/jems.12468

ORIGINAL ARTICLE

Journal of Economics & Management Strategy WILEY

How to examine external validity within an experiment

Amanda E. Kowalski

Department of Economics, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Amanda E. Kowalski, Department of Economics, University of Michigan, Ann Arbor, MI, USA.
Email: aekowals@umich.edu

Funding information
National Science Foundation, Grant/Award Number: 1350132

Abstract

A fundamental concern for researchers who analyze and design experiments is that the estimate obtained from the experiment might not be externally valid for other policies of interest. Researchers often attempt to assess external validity by comparing data from an experiment to external data. In this paper, I discuss approaches from the treatment effects literature that researchers can use to begin the examination of external validity internally, within the data from a single experiment. I focus on presenting the approaches simply using stylized examples.

1 | INTRODUCTION

The traditional reason that a researcher runs an experiment is to address selection into treatment. A researcher might be worried that individuals with better outcomes regardless of treatment are more likely to select into treatment, so the simple comparison of treated to untreated individuals will reflect a selection effect as well as a treatment effect. By running an experiment, the reasoning goes, a researcher isolates a single treatment effect by eliminating selection. However, there is still room for selection within an experiment. In many experiments, some lottery losers receive treatment and some lottery winners do not. In this paper, I focus on experiments with "two-sided noncompliance," in which both occur.

Some researchers view this type of selection as immaterial, and they discard information on which individuals select into treatment by focusing on the comparison of all lottery winners to all lottery losers. Other researchers view this type of selection as a nuisance, and they alter information on which individuals select into treatment by encouraging all individuals to comply with random assignment. I view this type of selection as a useful source of information that can be combined with assumptions to learn about the external validity of an experiment in the tradition of Heckman et al. (2000). The ability to learn from information on selection gives a researcher new reasons to run an experiment. An experiment is no longer a tool that eliminates selection; it is a tool that identifies selection. Furthermore, under ancillary assumptions, an experiment is no longer a tool that isolates a single treatment effect; it is a tool that identifies a range of heterogeneous treatment effects.

The central idea of this paper is that examination of how treatment effects vary with selection into treatment can inform external validity. Instead of thinking about external validity with respect to a setting, it is useful to think about external validity with respect to a policy. An experiment introduces a new policy to lottery winners and retains the current policy for lottery losers. The treatment effect obtained within the experiment is the treatment effect induced by the shift from the current policy to the new policy. Different policies could induce different individuals to select into treatment and thus have different impacts. I illustrate how data on individuals who select into treatment and institutional details can be used to predict the impact of specific alternative policies. Alternative policies can be within the setting of the experiment or within an entirely different setting, although external validity in other settings requires stronger assumptions.

I focus on how researchers can use recent advances from the treatment effects literature to begin examination of external validity within an experiment. I do not break new ground in terms of methodology or substantive application,

 © 2022 Wiley Periodicals LLC

and I do not aim to be comprehensive. Rather, I aim to present some existing methods simply using stylized examples, making them accessible to researchers who analyze and design experiments.

One of the virtues of experiments is that standard analysis is straightforward and relies on well-known assumptions. The well-known local average treatment effect (LATE) assumptions of independence and monotonicity proposed by Imbens and Angrist (1994) serve as the foundation for my analysis. Vytlacil (2002) constructs a model of selection into treatment that assumes no more than the LATE assumptions. The model can be interpreted as a generalized Roy (1951) model of the marginal treatment effect (MTE) introduced by Björklund and Moffitt (1987), in the tradition of Brinch et al. (2017), Carneiro et al. (2011), Cornelissen et al. (2018), and Heckman and Vytlacil (1999, 2001b, 2005). It has been used to motivate and formalize ancillary assumptions beyond the LATE assumptions. Here, I depict implications of the LATE assumptions and ancillary assumptions graphically. In future work, I envision that some researchers will choose to frame their analyses around such assumptions as I have done in an application to the Canadian National Breast Screening Study (Kowalski, 2021), and others will prefer to provide an algebraic presentation of the MTE model as I have done in an application to the Oregon Health Insurance Experiment (Kowalski, forthcoming). The combination of both papers supercedes previous work in Kowalski (2016). I provide a Stata command in Kowalski et al. (2018).

In Section 2, I depict information from a hypothetical experiment run by a health insurer. I begin by presenting information required for standard analysis of an experiment. I then present additional information available under the LATE assumptions that is often unreported. This additional information consists of shares and outcomes of "always takers" who take up treatment regardless of random assignment, "compliers" who take up treatment according to random assignment, and "never takers" who do not take up treatment regardless of random assignment, using the terminology of Angrist et al. (1996), obtained following Abadie (2002, 2003), Imbens and Rubin (1997), and Katz et al. (2001). Although there can be many types of heterogeneity within an experiment, I focus on heterogeneity across always takers, compliers, and never takers and how it can inform the impact of particular policies. In one interpretation, an experiment introduces a new policy for lottery winners that expands the fraction treated in the "intervention arm." It continues the current policy for lottery losers that maintains the fraction treated in the "control arm." Policies that would contract the fraction treated below the level of treatment in the control arm would induce treatment effects on always takers, and policies that would expand the fraction treated above the level of treatment in the intervention arm would induce treatment effects on never takers. In this way, heterogeneity across always takers, compliers, and never takers informs external validity with respect to policies that expand and contract the fraction treated.

In Section 3, I depict a test for heterogeneous selection as the fraction treated expands and contracts that uses a subset of the information available on compliers and never takers under the LATE assumptions. In Kowalski (forthcoming), I refer to this test as the "untreated outcome test," and my innovation is in the interpretation—I show that under the LATE assumptions alone, it identifies one specific instance of how selection into treatment varies as the fraction treated varies, a concept that generalizes the notion of "selection bias" (Angrist, 1998; Heckman et al., 1998), which is not identified under the LATE assumptions alone when the intervention is binary. This test is equivalent to tests proposed by Black et al. (2017) and Guo et al. (2014), and generalized by Mogstad et al. (2018). It is also similar to the Bertanha and Imbens (2014) test proposed for the regression discontinuity context and to the Einav et al. (2010) test in the insurance literature. This test for heterogeneous selection is a natural precursor to a test for treatment effect heterogeneity because if outcomes do not differ across groups due to heterogeneous selection effects, then differences could reflect heterogeneous treatment effects.

In Section 4, I depict a test for treatment effect heterogeneity equivalent to a test proposed by Brinch et al. (2017) and applied in Kowalski (2021). Brinch et al. (2017) conduct this test under two ancillary assumptions. As I show in Kowalski (2021), it is possible to conduct the test under only one of their ancillary assumptions; either one will suffice. I implement the test using the more justifiable assumption, and I discuss how data on covariates can be used to assess its plausibility. The assumption implies an upper or lower bound on the average treatment effect for always takers, and an additional assumption can imply an upper or lower bound on the average treatment effect for never takers. Using the insurance experiment, I discuss how the implied bounds on always and never takers yield specific implications for how to scale up the treatment.

In Section 5, I demonstrate how stronger assumptions of Angrist (2004), Bertanha and Imbens (2014), Black et al. (2017), Brinch et al. (2017), Guo et al. (2014), Hausman (1978), Heckman (1979), Huber (2013), and Willis and Rosen (1979) yield estimates of treatment effects in lieu of bounds. I also discuss how data can inform the plausibility of the assumptions. Because I do not need stronger assumptions to draw meaningful conclusions about external validity from the insurance experiment, I introduce a stylized clinical trial for hip replacement surgery to illustrate the implications of stronger assumptions for external validity.

The approaches that I discuss here do not supplant other approaches to examine external validity such as subgroup analysis, LATE-reweighting, and causal forests (Angrist & Fernandez-Val, 2013; Hotz et al., 2005; Wager & Athey, 2018). However, I demonstrate that subgroup analysis alone would not be sufficient to reach the same conclusions in the surgery trial. I conclude by discussing implications for experimental design in Section 6.

2 | AN EXPERIMENT UNDER THE LATE ASSUMPTIONS

Consider the following stylized example. A health insurer rolled out a new wellness plan at one pilot site. For the same premium, beneficiaries chose between a traditional plan and a wellness plan that offered the same benefits plus gym membership reimbursement. Based on financials from the pilot site, the insurer thinks that the wellness plan lowers the average costs that it pays on behalf of beneficiaries. It is considering two options: (1) offering the wellness plan as a choice at all sites or (2) enrolling all beneficiaries in the wellness plan all sites. However, it recognizes that only 25% of beneficiaries at the pilot site chose the wellness plan, and those enrollees might not be representative.

Before pursuing either wellness plan expansion option, the insurer decides to gather more evidence by running a randomized experiment at the pilot site. It mails an informational brochure that promotes the wellness plan to lottery winners but does not mail a brochure to lottery losers. The treatment, defined relative to current policy, is enrollment in the wellness plan. There is two-sided noncompliance, as some lottery winners enroll in the traditional plan and some lottery losers enroll in the wellness plan. The insurer observes whether each individual wins the lottery and whether each individual enrolls in the wellness plan. It also observes an outcome for each individual: the average monthly health care costs that it pays on behalf of each beneficiary.

Standard analysis of an experiment begins by comparing average outcomes in the intervention and control arms. I depict these outcomes in Figure 1. As shown, the new policy seems to decrease cost because the average outcome (cost per month) is $17.5 lower in the intervention arm than it is in the control arm. This difference in average outcomes is often called the "reduced form," as labeled along the vertical axis, or the "intent to treat (ITT)." It gives an estimate of the impact of the new policy (the mailing of the informational brochure) on the outcome. In experiments with two-sided noncompliance, lottery status does not perfectly determine treatment, so the reduced form does not give an estimate of the impact of the treatment (enrollment in the wellness plan) on the outcome. Calculation of the reduced form does not even require data on treatment. Some researchers report only the reduced form.

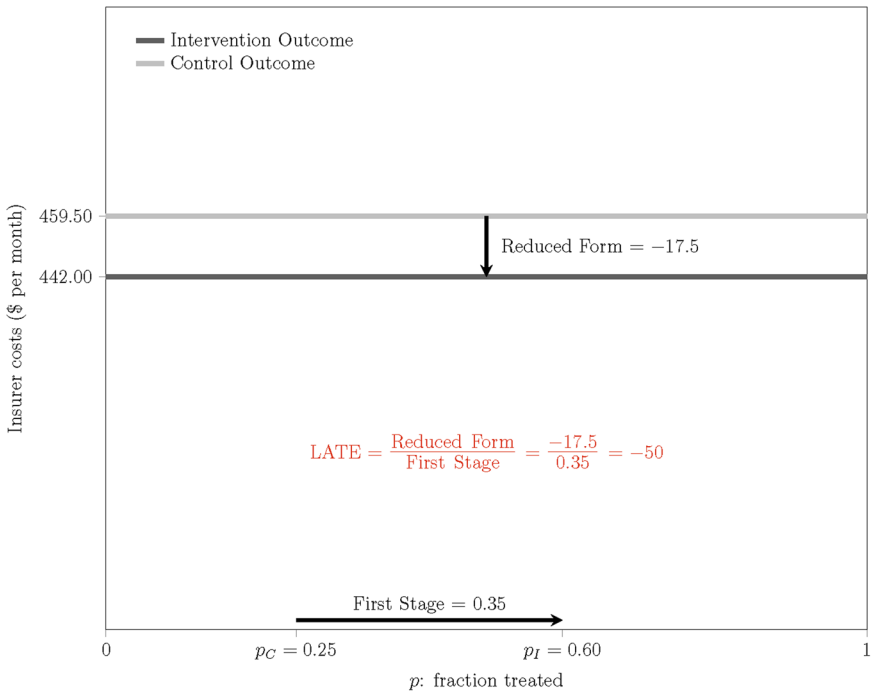

FIGURE 1 Average outcomes in intervention and control arms under LATE assumptions [Color figure can be viewed at wileyonlinelibrary.com]

Standard analysis of an experiment next compares the probability of treatment in the control and intervention arms. By the LATE independence assumption, lottery status is independent of treatment, so I can depict the probability of treatment in the control and intervention arms along the same horizontal axis in Figure 1. As shown, p_C represents the probability of treatment in the control arm, and p_I represents the probability of treatment in the intervention arm. The difference $p_I - p_C$ is often called the "first stage." It gives an estimate of the impact of winning the lottery on the fraction treated p. In experiments with two-sided noncompliance, the first stage is less than one. In Figure 1, 25% of lottery losers and 60% of lottery winners receive treatment, so the first stage implies that winning the lottery increases the fraction treated by 35 percentage points. That is, the mailing of the informational brochure increases enrollment in the wellness plan by 35 percentage points.

To obtain an estimate of the impact of the treatment on the outcome, standard analysis of an experiment divides the reduced form by the first stage. Under the LATE monotonicity assumption, which requires that the new policy either weakly increases treatment for all participants or weakly decreases treatment for all participants, this quotient gives the LATE of Imbens and Angrist (1994). Many researchers report the LATE as the single treatment effect that the experiment isolates. The LATE gives the average treatment effect for "compliers," individuals whose treatment status is determined by their random assignment, in the terminology of Angrist et al. (1996). In Figure 1, the LATE of -50 implies that the wellness plan reduces health care costs by \$50 per month on average among individuals who take up the plan if and only if the insurer mails them the informational brochure.

The evidence from the standard analysis of the experiment looks promising. The wellness plan decreases costs for compliers. But does it decrease costs for all individuals? Is the LATE useful to the insurer in deciding between the two wellness plan expansion options?

To inform the external validity of the LATE with respect to the expansion options, I consider two groups of individuals to which the LATE need not apply: "always takers" who take up treatment regardless of random assignment and "never takers" who do not take up treatment regardless of random assignment, in the terminology of Angrist et al. (1996). Under this terminology, the LATE monotonicity assumption rules out "defiers" who take up treatment if and only if they lose the lottery (Angrist et al., 1996; Balke & Pearl, 1993), so experiments with two-sided noncompliance involve only always takers, compliers, and never takers. Researchers cannot label each individual as an always taker, complier, or never taker: lottery winners who take up treatment could be always takers or compliers; lottery losers who do not take up treatment could be compliers or never takers. However, researchers can label lottery losers who take up treatment as always takers and lottery winners who do not take up treatment as never takers.

The ability to identify some individuals as always or never takers allows researchers to learn more about compliers. The LATE independence assumption implies that lottery status is independent of whether an individual is an always taker, complier, or never taker. Therefore, the observed share of treated lottery losers yields an estimate of the share of always takers in the full sample, and the observed share of untreated lottery winners yields an estimate of the share of never takers in the full sample. Furthermore, because always and never takers do not change their treatment status based on their lottery status, their average outcomes should not depend on their lottery status. Using the shares and average outcomes of always takers and never takers, researchers can estimate the average outcomes of treated and untreated compliers, as demonstrated by Abadie (2002, 2003), Imbens and Rubin (1997), and Katz et al. (2001).[1]

To illustrate the calculation of the average outcomes of always takers, compliers, and never takers graphically, I continue the insurance example in Figure 2. As shown by Imbens and Rubin (1997) and Vytlacil (2002), the LATE assumptions imply an ordering from always takers to compliers to never takers. Consistent with this ordering, I label ranges of the horizontal axis that correspond to the shares of each group. On the left, the fraction p_C of individuals who receive treatment regardless of their lottery status are always takers. In the middle, the fraction $(p_I - p_C)$ of individuals who receive treatment if and only if they win the lottery are compliers. On the right, the remaining fraction $(1 - p_I)$ of individuals who do not receive treatment regardless of their lottery status are never takers. The intuition behind the ordering is clear if we interpret the experiment in terms of a policy change that occurs within the intervention arm. In the experiment as a whole, always takers are the individuals who receive treatment under the current policy, compliers are the new individuals who can be induced to receive treatment by the policy change, and never takers are the remaining individuals who could be induced to receive treatment by a future policy change. In the insurance example, the policy change is the mailing of the informational brochure and the treatment is enrollment in the wellness plan.

Along the vertical axis of Figure 2, I plot average treated and untreated outcomes in the intervention and control arms over the relevant ranges of the horizontal axis in the dark shading. As shown, the average treated outcome in the intervention arm is \$410, which represents a weighted average of the treated outcomes of always takers and compliers. The average treated outcome in the control arm is \$452, which represents the average treated outcome of always takers.

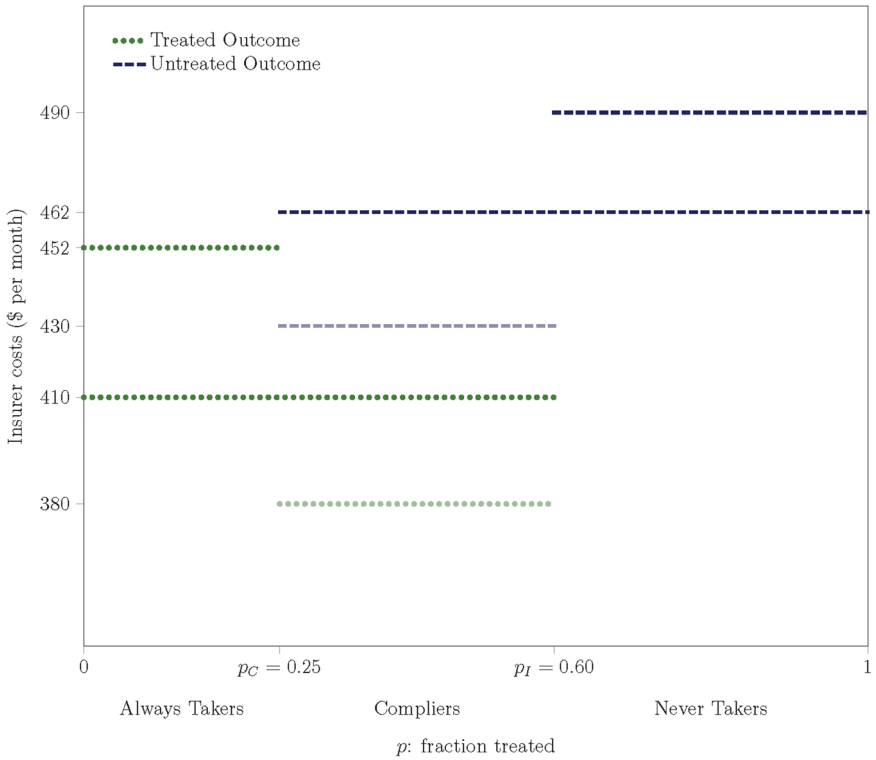

FIGURE 2 Average treated and untreated outcomes in intervention and control arms and average treated and untreated outcomes of compliers under LATE assumptions [Color figure can be viewed at wileyonlinelibrary.com]

Because always takers make up 25% of the full sample and always takers combined with compliers make up 60% of the full sample, the average treated outcome of compliers is \$380 $(= (0.6/(0.6 - 0.25)) \times 410 - (0.25/(0.6 - 0.25)) \times 452)$, as depicted in light shading. Similar logic using the untreated outcomes implies that the average untreated outcome of never takers is \$490 and that the average untreated outcome of compliers is \$430 $(=((1 - 0.25)/(0.6 - 0.25)) \times 462 - ((1 - 0.60)/(0.6 - 0.25)) \times 490)$, as depicted in light shading. Researchers who would like to replicate these calculations can use the Stata command *mtebinary* (Kowalski et al., 2018).

As shown by Imbens and Rubin (1997), the LATE is equal to the difference in the average treated and untreated outcomes of compliers. Accordingly, in Figure 3, I depict an arrow that gives the sign and magnitude of the LATE. However, I could have obtained the LATE using Figure 1 alone, even if my data would not allow me to construct Figures 2 and 3. Construction of Figures 2 and 3 requires data on outcomes by lottery status *and* treatment. In contrast, construction of Figure 1 only requires data on outcomes by lottery status (for the reduced form) and data on treatment by lottery status (for the first stage). As shown by Angrist (1990) and Angrist and Krueger (1992), it is possible to obtain the LATE via the Wald (1940) approach using separate datasets for the reduced form and first stage. Because the LATE can be obtained using limited data, it stands to reason that it does not capture all available information. Figure 3 provides additional information relative to Figure 1.

Using the additional information depicted in Figure 3, I emphasize that always and never takers are distinct groups to which the LATE need not apply. In the insurance example, these groups are sizeable. Furthermore, the average treated outcome of always takers and the average untreated outcome of never takers are known. The average untreated outcome of always takers and the average treated outcome of never takers are not known. If they could be identified, then it would be possible to estimate the average treatment effect for each group as the difference between their average treated and untreated outcomes. Similarly, if they could be bounded, then it would be possible to bound the average treatment effect for each group, as discussed by Imbens and Rubin (1997). Such bounds could be implied by natural bounds on the range of outcomes in the tradition of Balke and Pearl (1997), Manski (1990), and Robins (1989), or they could be implied by ancillary assumptions.

Even in the absence of ancillary assumptions, a researcher examining Figure 3 might conjecture that the LATE is not equal to the average treatment effect for always and never takers, and is thus not externally valid for all alternative policies. If the treatment effect were the same for everyone, then why do some individuals select into treatment while

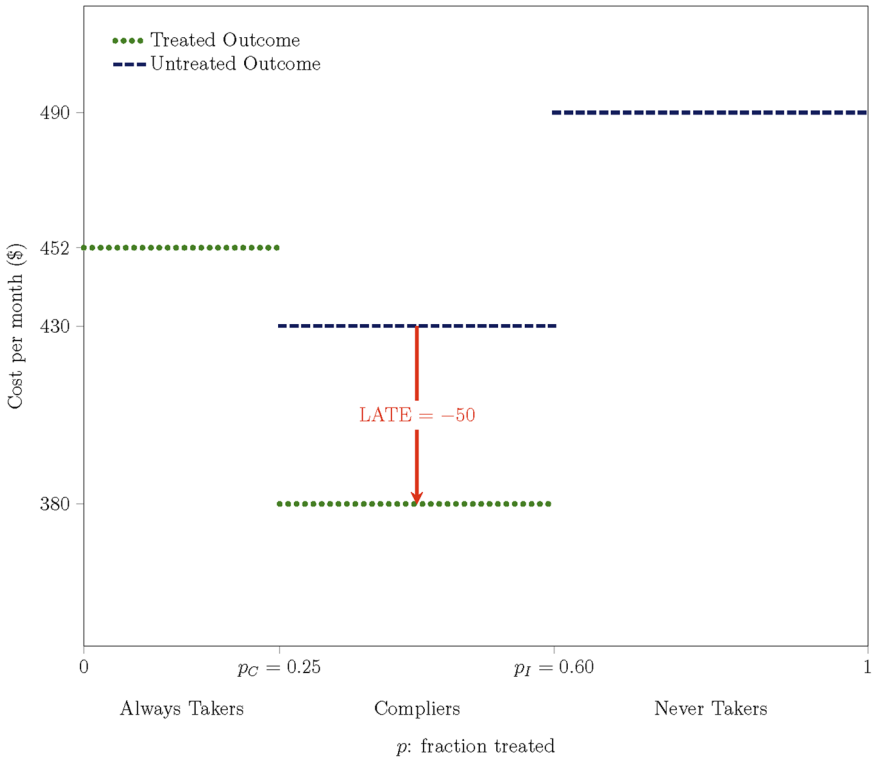

FIGURE 3 Average outcomes of always takers, compliers, and never takers under LATE assumptions [Color figure can be viewed at wileyonlinelibrary.com]

others do not, even within the same arm of the experiment? Also, why do the average treated and untreated outcomes differ across always takers, compliers, and never takers? Is it because their average treatment effects differ? In the next sections, I interpret the implications of these differences for external validity, first under the LATE assumptions, and then under ancillary assumptions.

3 | TEST FOR HETEROGENEOUS SELECTION UNDER THE LATE ASSUMPTIONS

Consider the test of the null hypothesis that the difference between the average untreated outcomes of compliers and never takers is equal to zero. This test is equivalent or similar to tests proposed by Bertanha and Imbens (2014), Black et al. (2017), Guo et al. (2014), and generalized by Mogstad et al. (2018). It is also related to the "cost curve" test of Einav et al. (2010) from the insurance literature when the untreated outcome is uninsured costs (or, in my stylized example, costs under the traditional plan). In Kowalski (forthcoming), I refer to this as the "untreated outcome test," and I provide a novel interpretation for it. I show that under the LATE assumptions, it identifies a specific instance of heterogeneous selection as the fraction treated p varies. The logic behind this interpretation is simple. Untreated compliers and never takers do not receive treatment. Therefore, a difference in their outcomes cannot reflect a difference in the treatment effect. It can only reflect a difference in selection.

Continuing the insurance example, Figure 4 shows that the average untreated outcome of compliers is $60 lower than the average outcome of never takers. If this difference is statistically different from zero, then the test rejects selection homogeneity. Statistical significance can be obtained via a variety of approaches, including bootstrapping, which can account for estimation of the average outcomes as well as their difference.

The sign of the untreated outcome test statistic, the difference in average untreated outcomes between compliers and never takers, indicates whether selection is positive or negative. In Figure 4, a negative untreated outcome test statistic indicates negative selection, such that individuals with higher average costs (never takers) select into treatment after individuals with lower average costs (compliers). When the treatment represents enrollment in an insurance plan, it is customary to refer to negative selection as "advantageous selection" and positive selection as "adverse selection."

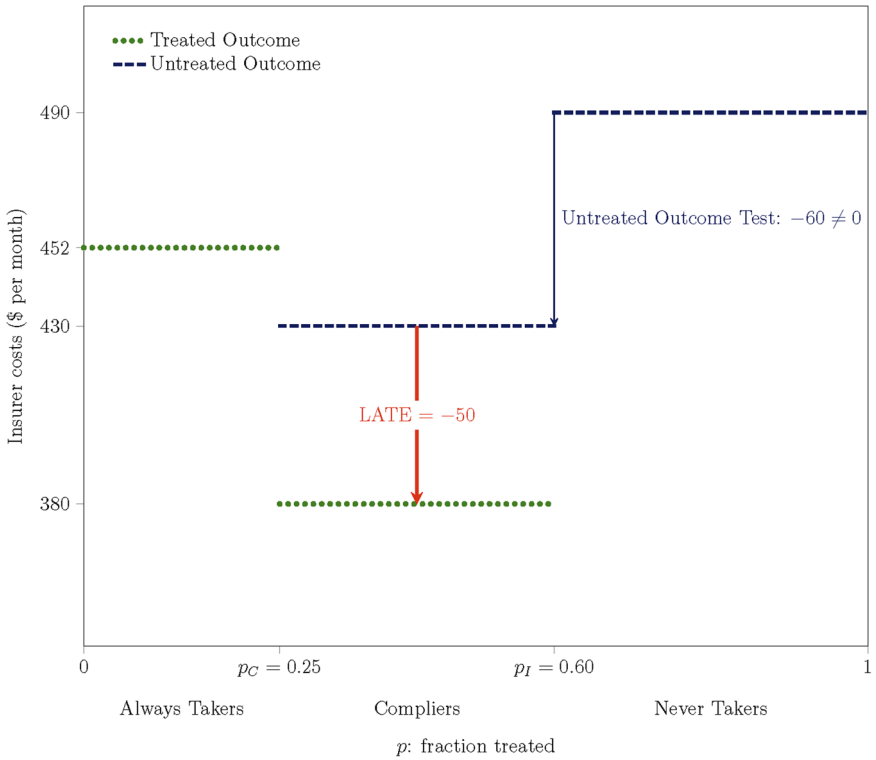

FIGURE 4 Untreated outcome test rejects: Test statistic shows negative selection under LATE assumptions [Color figure can be viewed at wileyonlinelibrary.com]

Although insurers often face adverse selection, there is a rationale for advantageous selection in the stylized insurance example: individuals with lower health care costs in the traditional plan have better health behaviors and are therefore more likely to enroll in the wellness plan that offers gym membership reimbursement.

There can be positive selection on some outcomes and negative selection on others within the same experiment. In the insurance example, consider a measure of how often beneficiaries go to the gym. Even though there is negative selection on health care costs, there could be positive selection on how often beneficiaries go to the gym. This positive selection on gym behavior could help to explain negative selection on health care costs. It could also offer a rationale for why the treatment effect on health care costs could vary with the fraction treated p. For example, there might be less scope for the wellness plan to reduce health care costs for beneficiaries who go to the gym more frequently.

The analogous *treated* outcome test, which tests the null hypothesis that the difference between the average treated outcomes of always takers and compliers is equal to zero, has also been considered by the literature that examines tests similar or equivalent to the untreated outcome test (Bertanha & Imbens, 2014; Black et al., 2017; Guo et al., 2014). In the insurance literature, the treated outcome test is related to the "cost curve" test of Einav et al. (2010) when the treated outcome is *insured* costs (or, in my stylized example, costs under the wellness plan). In Kowalski (forthcoming), I emphasize that the treated outcome test does not isolate heterogeneous selection. For this reason, I do not recommend running the treated outcome test, but I discuss it here to illustrate why it does not isolate heterogeneous selection.

Continuing the insurance example, consider the treated outcome test depicted in Figure 5. It shows that the average outcome of always takers is $72 higher than the average treated outcome of compliers. Assume that this difference is statistically different from zero, so the treated outcome test rejects. This result could be entirely due to heterogeneous selection from always takers to compliers, which would be the case if the average treatment effects for both groups were equal. In that case, the average treatment effect for always takers would be equal to the LATE of −$50, and the average untreated outcome of always takers would be $402 (=452 − 50). Alternatively, the result of the treated outcome test could be entirely due to treatment effect heterogeneity from always takers to compliers, which would be the case if there were no selection heterogeneity across the two groups. In that case, the average untreated outcome of always takers would be equal to the average untreated outcome of compliers of $430, and the average treatment effect for always takers would be an increase of $22 (=452 − 430). It is also possible that the treated outcome test could reflect a

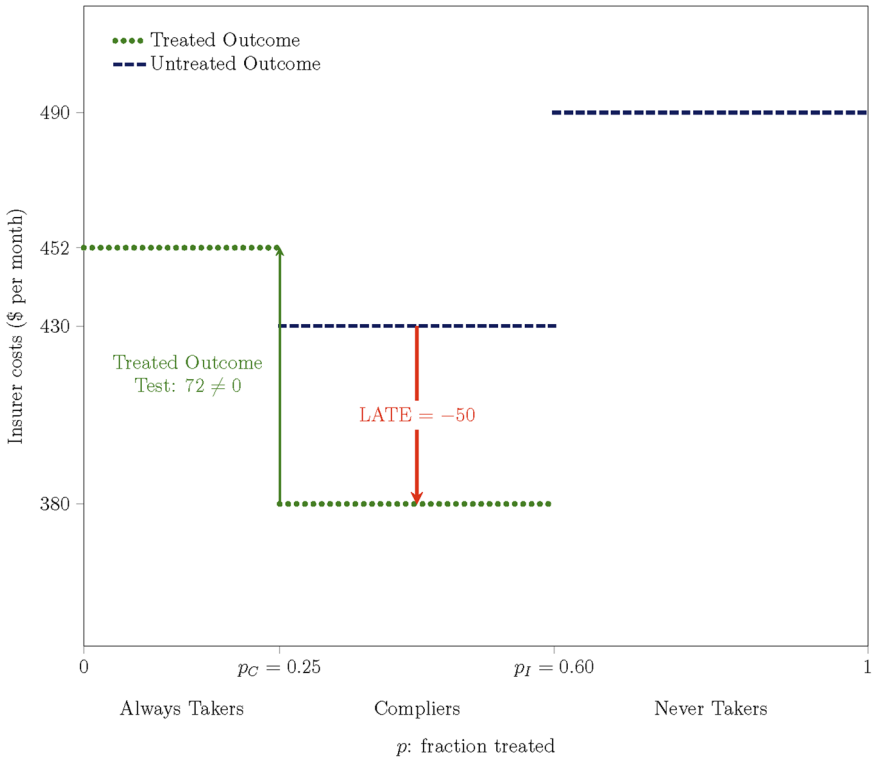

FIGURE 5 Treated outcome test rejects: Test statistic shows positive selection and/or treatment effect heterogeneity under LATE assumptions [Color figure can be viewed at wileyonlinelibrary.com]

combination of nonzero selection heterogeneity and nonzero treatment effect heterogeneity, which would be the case if the average untreated outcome of always takers were anything other than $402 or $430. As shown, the treated outcome test reflects selection heterogeneity *plus* treatment effect heterogeneity, while the untreated outcome test can only reflect selection heterogeneity.

It is tempting to think that the treated outcome test should have the same implications as the untreated outcome test because the distinction between treated and untreated should be immaterial. However, as I discuss in Kowalski (forthcoming), the distinction between treated and untreated is material to the definition of the treatment and thus to the definitions of the treatment effect, treatment effect heterogeneity, and selection heterogeneity. The treatment effect is defined as the treated outcome minus the untreated outcome, not the untreated outcome minus the treated outcome. Therefore, the treatment effect has magnitude *and* direction, which is why I depict the local average treatment effect with an arrow in the figures. Renaming the treated the untreated and vice versa might seem inconsequential, but such a swap would change the direction of the arrow. The swap would also change the definitions of selection and treatment effect heterogeneity. The treated outcome test would detect only selection, and the untreated outcome test would detect a combination of selection and treatment effect heterogeneity, creating a different but no less material distinction between the tests.

The distinction between treated and untreated outcomes forms the foundation for the tests for heterogeneous treatment effects that I present in the next sections. The tests use differences in untreated outcomes to motivate ancillary assumptions. The assumptions purge selection heterogeneity from treated outcomes to isolate treatment effect heterogeneity.

4 | TEST FOR HETEROGENEOUS TREATMENT EFFECTS UNDER ANCILLARY ASSUMPTIONS

Brinch et al. (2017) propose a test for treatment effect heterogeneity that relies on two ancillary assumptions beyond the LATE assumptions: (1) weak monotonicity of untreated outcomes in the fraction treated p, and (2) weak monotonicity of treated outcomes in the fraction treated p. As I show in Kowalski (2021) and demonstrate here, the test only requires one

of their ancillary assumptions; either one is sufficient. It is harder to justify the assumption of weak monotonicity of treated outcomes because variation in treated outcomes can reflect treatment effect heterogeneity as well as selection heterogeneity. I prefer not to make an opaque assumption about treatment effect heterogeneity to test for treatment effect heterogeneity. I therefore proceed under the assumption of weak monotonicity of untreated outcomes, which is easier to justify because variation in untreated outcomes only reflects selection heterogeneity.

Researchers can justify the assumption of weak monotonicity of untreated outcomes using institutional details about their experiments. The assumption is easiest to justify if there is a plausible mechanism for differential gain from selection that is correlated with untreated outcomes. Such a gain need not be measured in terms of the main outcome of interest. In the insurance example, it is plausible that individuals with lower costs under the traditional plan are healthier and more likely to enroll in the wellness plan because they will be more likely to gain financially from the gym membership reimbursement. This gain to the enrollee is a loss to the insurer in terms of the main outcome of interest: its monthly costs.

Covariates collected at baseline before the experiment can also justify or call into question the assumption of weak monotonicity of untreated outcomes. Unlike untreated outcomes, which are not observed for always takers, baseline covariates can be observed for always takers, compliers, and never takers. Baseline covariates can thus serve as a proxy for untreated outcomes for all individuals. In the insurance example, baseline body mass index (BMI) can serve as a proxy for costs under the traditional plan because both are likely correlated with underlying health. The monotonic relationship in baseline BMI shown in Figure 6 lends support to the assumption of weak monotonicity of untreated outcomes. It shows that in terms of baseline BMI, healthier people are more likely to enroll in the wellness plan.

Figure 7 depicts the implications of the assumption of weak monotonicity of untreated outcomes in the insurance example. As discussed previously, the LATE monotonicity assumption implies an ordering from always takers to compliers to never takers along the horizontal axis. The ancillary assumption of weak monotonicity of untreated outcomes implies the same ordering along the vertical axis. Because the average untreated outcome of compliers is smaller than the average untreated outcome of never takers, yielding a negative untreated outcome test statistic, the ancillary assumption of weak monotonicity of untreated outcomes implies an upper bound on the average untreated outcome of always takers.

As depicted, the upper bound on the average untreated outcome of always takers implies a lower bound on the average treatment effect for always takers, which indicates that the wellness plan increases their average costs by at least $22. In contrast, the LATE indicates that the wellness plan *decreases* average costs by $50 for compliers. Because the bound excludes

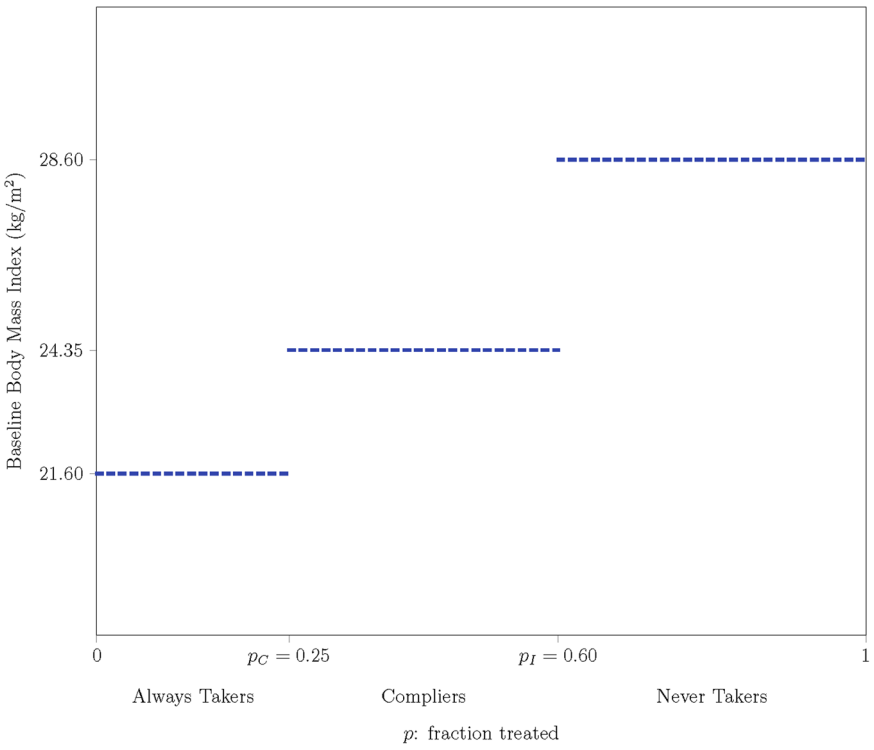

FIGURE 6 Monotonicity of baseline body mass index lends support to the ancillary assumption of weak monotonicity of untreated outcomes [Color figure can be viewed at wileyonlinelibrary.com]

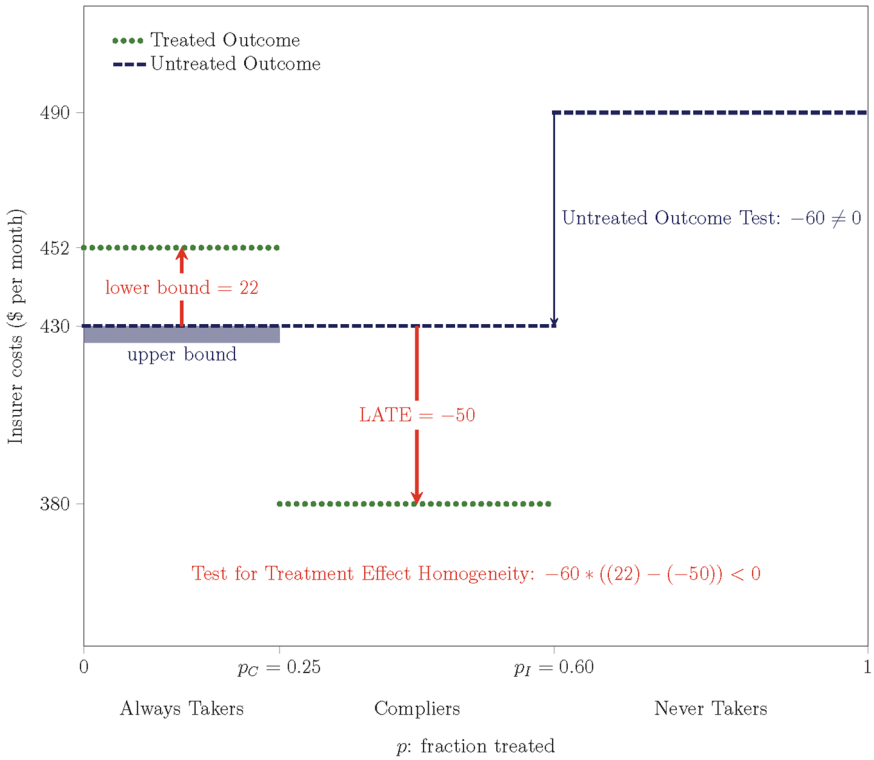

FIGURE 7 Test for treatment effect homogeneity rejects under ancillary assumption of weak monotonicity of untreated outcomes: Average treatment effect bound for always takers [Color figure can be viewed at wileyonlinelibrary.com]

the LATE, the test rejects the null hypothesis of treatment effect homogeneity under the single assumption that untreated outcomes are weakly monotonic in the fraction treated p. Allowing for settings in which the lower bound could instead be an upper bound because of selection heterogeneity in the opposite direction, the test rejects if the product of the untreated outcome test statistic and the difference between the bound on the average treatment effect for always takers and the LATE is negative, as shown algebraically in Figure 7.[2] It is straightforward to visualize that the test would also reject under the single assumption that treated outcomes are weakly monotonic in the fraction treated p, demonstrating that either of the Brinch et al. (2017) ancillary assumptions is sufficient to test the null of treatment effect homogeneity.

What could explain the pattern of treatment effect heterogeneity shown in Figure 7? Baseline covariates show that always takers have lower body mass index than compliers, and the assumption of weak monotonicity of untreated outcomes in the fraction treated p implies that always takers would have lower insurer costs in the traditional plan. Why then, would the wellness plan increase insurer costs for always takers but decrease them for compliers? Recall that the wellness plan offers gym membership reimbursement. It is plausible that the wellness plan increases insurer costs for always takers because the gym membership reimbursement crowds out an expense that they were already making. Therefore, insurer costs increase through the reimbursement but do not decrease on other dimensions through improved health. In contrast, the wellness plan decreases insurer costs for compliers because it induces them to take up a new gym membership, which decreases insurer costs on net.

The treatment effect heterogeneity between always takers and compliers depicted in Figure 7 has important implications for external validity. The LATE implies that the wellness plan decreases costs by $50 for compliers, but costs must increase by at least $22 for always takers, so the LATE cannot be externally valid for all alternative policies. However, the insurer is interested in whether the LATE is externally valid with respect to two specific policy options that would expand the wellness plan. The first option is to rollout the wellness plan at all sites just as it did last year at the pilot site, without the informational mailing. Figure 7 implies that the rollout of the wellness plan increased costs at the pilot site because the only enrollees were always takers. Based on the narrative for the patterns of selection and treatment effect heterogeneity observed at the pilot site, it could be reasonable for the insurer to assume that those patterns would also apply at other sites: individuals with lower insurer costs in the traditional plan and lower body mass index would enroll, and insurer costs would increase because of the gym reimbursement. Therefore, the first policy option, rolling out the wellness plan as an option at all sites, would not make sense.

However, the design of the policy in the intervention arm within the experiment suggests an alternative policy option: rolling out the wellness plan at other sites in conjunction with the informational mailing. At the pilot site, costs increased by at least \$22 for always takers, who represent 25% of beneficiaries, decreased by \$50 for compliers, who represent 35% of beneficiaries, and remained unchanged for never takers, who represent 40% of beneficiaries. Therefore, assuming negligible costs for the informational mailing, as long as the insurer believes that costs did not increase by more than \$70 $(=(50 \times 0.35)/0.25)$ for always takers, then the combination of the wellness plan and the informational mailing decreased its costs at the pilot site, and it could decrease costs at other sites.

The insurer also has its second wellness plan expansion option to consider: would it make more sense to enroll all beneficiaries at all sites in the wellness plan? To understand the impact of this option, the insurer needs information on the average treatment effect for never takers. Weak monotonicity of untreated outcomes is not sufficient to yield any meaningful conclusions about the average treatment effect for never takers because it is their *treated* outcomes that are unobserved. Although weak monotonicity of treated outcomes would yield meaningful conclusions, such an assumption is harder to justify as discussed. An alternative is to impose an additional ancillary assumption of weak monotonicity of treatment effects in conjunction with weak monotonicity of untreated outcomes. This approach is transparent because it makes assumptions about selection and treatment effect heterogeneity separately.

Researchers can again turn to institutional details of their experiments and data on covariates to justify an additional ancillary assumption of weak monotonicity of treatment effects. Suppose that before the experiment began, the insurer asked all beneficiaries at the pilot site how likely they would be to join a gym if offered gym membership reimbursement. It found that on average, always takers report that they would be the most likely to join a gym and never takers report that they would be the least likely to do so. However, given that the wellness plan decreases average insurer costs by more for compliers than for always takers, it is plausible that average costs would decrease by even more for never takers. Recall that never takers are the least healthy as measured by their baseline body mass index, so they have the biggest potential for insurer cost reductions through improved health. Thus, the variation in the likelihood of joining a gym observed across always takers, compliers, and never takers can support an additional ancillary assumption of weak monotonicity of treatment effects in the fraction treated p.

Figure 8 depicts the implications of an additional ancillary assumption of weak monotonicity of treatment effects. Under the assumption of weak monotonicity of untreated outcomes alone, the average treatment effect for always

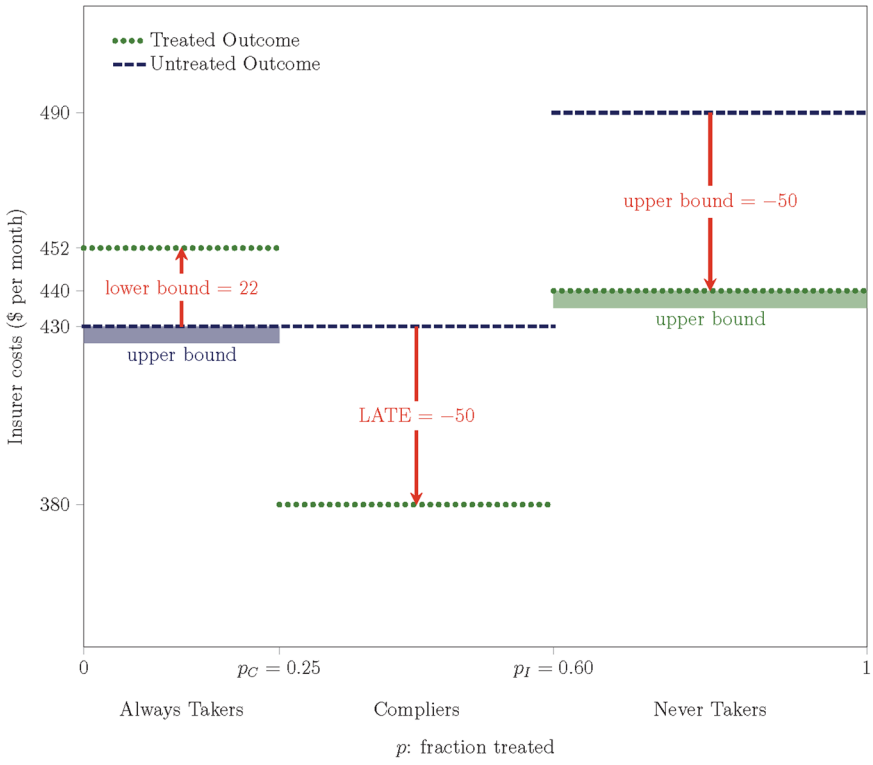

FIGURE 8 Treatment effect bounds for always takers and never takers under ancillary assumptions of weak monotonicity of untreated outcomes and weak monotonicity of treatment effects [Color figure can be viewed at wileyonlinelibrary.com]

takers must be weakly greater than the average treatment effect for compliers. Weak monotonicity of treatment effects then implies that the average treatment effect for compliers is weakly greater than the average treatment effect for never takers. Therefore, the LATE of −$50 gives an upper bound on the average treatment effect for never takers.

This upper bound helps to inform external validity with respect to the second expansion option of enrolling all beneficiaries in the wellness plan. It implies that the wellness plan decreases insurer costs by at least $50 per person for the 40% of beneficiaries that are never takers. At the same time, the wellness plan increases insurer costs by at least $22 per person for the 25% of beneficiaries that are always takers and decreases costs by $50 per person for the 35% of beneficiaries that are compliers. These bounds do not imply a bound on the average treatment effect across all beneficiaries. However, it is possible to calculate the average cost increase for always takers that would make the insurer indifferent between enrolling all beneficiaries in the wellness plan and enrolling all beneficiaries in the traditional plan at the pilot site. As long as the average cost does not increase by more than $150 $(=(50 \times 0.4 + 50 \times 0.35)/0.25)$ per month for always takers, the insurer is better off enrolling all beneficiaries in the wellness plan. An extra $150 per month for always takers would bring their average cost to $580, which is much higher than the average cost for never takers, who have the highest costs, so it is plausible that the best wellness expansion option for the insurer is to enroll all beneficiaries in the wellness plan. This option strictly dominates the option of simply keeping the wellness plan as a choice at the pilot site.

But how should the insurer think about the merits of the expansion options beyond the pilot site? One traditional approach is to estimate LATEs within subgroups determined by covariates available at the pilot site and the site of interest and then re-weight those LATEs. Such an approach only informs the impact of the wellness plan for compliers who respond to the informational mailing, which is not necessarily relevant for either expansion option. To augment such an approach, the insurer could examine covariates at the additional sites that vary with the fraction treated p at the pilot site. For example, if a site of interest has an average body mass index that is comparable to that of the always takers, then the insurer might be wary of expanding the wellness program to all beneficiaries at that site.

5 | TESTS FOR HETEROGENEOUS TREATMENT EFFECTS UNDER STRONGER ANCILLARY ASSUMPTIONS

Thus far, I have demonstrated how researchers can reject the null of treatment effect homogeneity under the ancillary assumption of weak monotonicity of untreated outcomes in the fraction treated p. They can impose stronger assumptions to generate more powerful tests and to obtain estimates of average treatment effects for always and never takers instead of bounds. Although it is natural to progress from weaker assumptions to stronger assumptions in empirical work, stronger assumptions were proposed first. To illustrate stronger assumptions, I present a second stylized example in which I cannot reject treatment effect homogeneity under weak monotonicity of untreated outcomes and treatment effects, but I can under stronger assumptions.

Suppose researchers at a large hospital want to evaluate the impact of outpatient hip replacement surgery as compared to inpatient hip replacement surgery. Unlike inpatient surgery, outpatient surgery aims to send patients home on the same day. Potential benefits include reduced infection risk due to decreased time at the hospital, increased patient satisfaction from rehabilitation at home, and reduced health care costs. Data from the hospital records show that the rate of subsequent hip fractures is lower among patients who undergo outpatient surgery, consistent with higher quality and reduced cost. However, the researchers are concerned that patients who receive outpatient surgery might be systematically different from patients who receive inpatient surgery. To gather more evidence, they conduct a randomized trial. They choose trial participants from the population of patients scheduled to have hip replacement surgery and randomly assign them to a default of outpatient surgery (the new policy in the intervention arm) or inpatient surgery (the current policy in the control arm). Patients can consult with their doctors before choosing whether to proceed with their default surgery type, which leads to two-sided noncompliance with the treatment, receipt of outpatient surgery. The main outcome of interest is hip fracture within 6 months after surgery.

Results from the trial show that the average subsequent hip fracture rate is 2.75 percentage points lower in the intervention arm than it is in the control arm. The corresponding LATE depicted in Figure 9 implies that the treatment reduces the subsequent hip fracture rate for compliers by 5 percentage points on average. Moreover, the researchers find that the LATE is negative within all subgroups that they define by age, sex, and comorbidity score. Based on this promising evidence, should they go beyond recommending outpatient surgery as the default and mandate it for everyone? Or should they instead provide a subsidy to patients who undergo outpatient surgery?

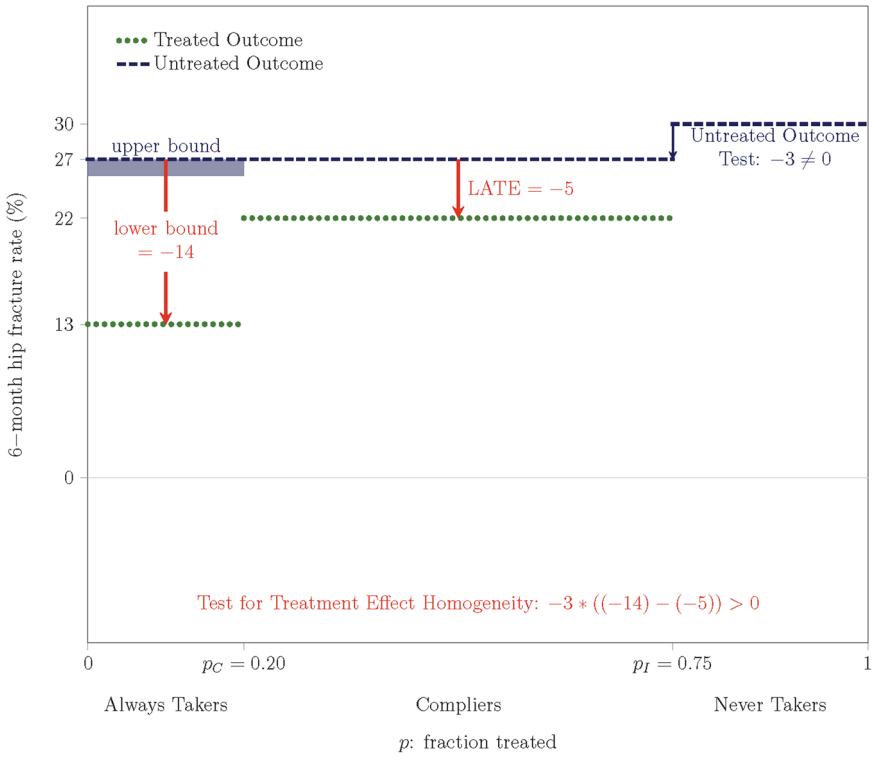

FIGURE 9 Test for treatment effect homogeneity does not reject under ancillary assumption of weak monotonicity of untreated outcomes: Bound on average treatment effect for always takers [Color figure can be viewed at wileyonlinelibrary.com]

To answer these questions, the researchers can begin by testing for heterogeneous selection within the trial under the LATE assumptions. The untreated outcome test statistic in Figure 9 is negative, implying selection heterogeneity such that the average subsequent hip fracture rate is 3 percentage points lower for untreated compliers than it is for never takers. This selection heterogeneity is consistent with the explanation that patients who are sicker are more likely to receive inpatient surgery because it allows for more direct supervision, which could justify an ancillary assumption of monotonicity of untreated outcomes. Baseline comorbidity score, which serves as a proxy for underlying health, also provides support for such an assumption. The average baseline comorbidity score increases from always takers to compliers to never takers. This evidence can also provide support for an assumption of weak monotonicity of treatment effects if there is reason to believe that patients with fewer comorbidities will respond better to outpatient surgery.

As shown in Figure 9, the ancillary assumption of weak monotonicity of untreated outcomes implies a lower bound on the average treatment effect for always takers. The product of the untreated outcome test statistic and the difference between the lower bound on the average treatment effect for always takers and the LATE is positive, so the test fails to reject the null of treatment effect homogeneity. Moreover, the additional ancillary assumption of weak monotonicity of treatment effects does not provide an informative bound on the average treatment effect for never takers in the sense that the average treatment effect for never takers could be greater or less than the LATE.

Even though the researchers cannot reject the null hypothesis of treatment effect homogeneity under weak monotonicity assumptions, they might be able to do so under stronger assumptions. One such set of stronger assumptions is linearity of untreated outcomes and treatment effects in the fraction treated p. Linearity assumptions are established in the literature. Olsen (1980) imposes linearity of treated outcomes, while Brinch et al. (2017) impose linearity of untreated and treated outcomes to test the null hypothesis of treatment effect homogeneity. Angrist (2004), Bertanha and Imbens (2014), Black et al. (2017) Guo et al. (2014), Hausman (1978), and Huber (2013) propose equivalent or similar tests, but they do not all explicitly state linearity assumptions. Linearity of untreated and treated outcomes is equivalent to linearity of untreated outcomes and treatment effects. However, I prefer to make assumptions on untreated outcomes and treatment effects because they are more transparent to motivate and assess.

Data on covariates can lend support to linearity assumptions, as I demonstrate in Kowalski (forthcoming). Continuing the surgery example, suppose the observed monotonic relationship in the average baseline comorbidity score is

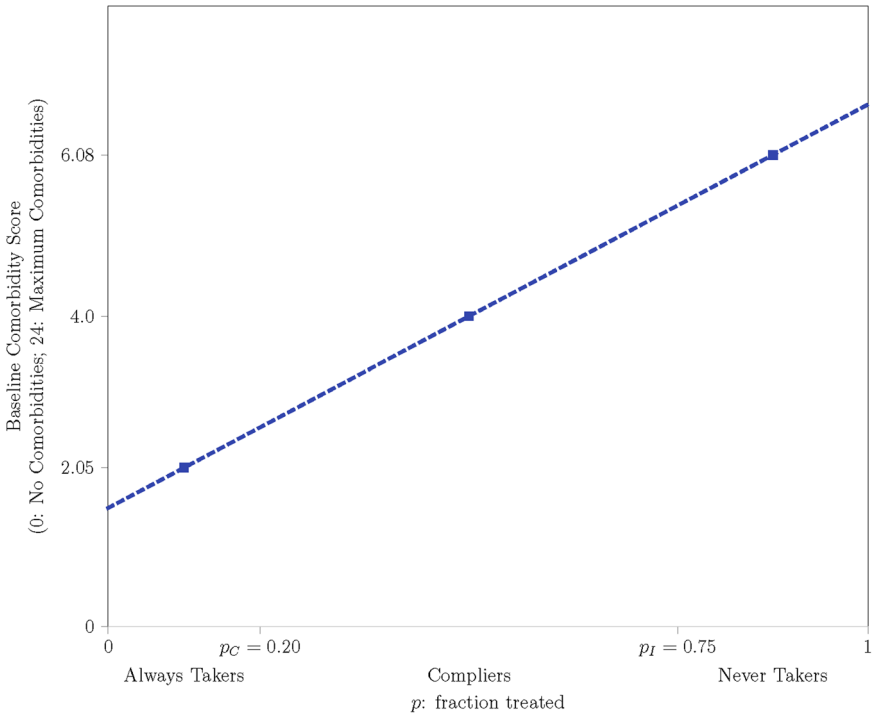

FIGURE 10 Linearity of baseline comorbidity score lends support to the ancillary assumptions of linearity of untreated outcomes and linearity of treatment effects [Color figure can be viewed at wileyonlinelibrary.com]

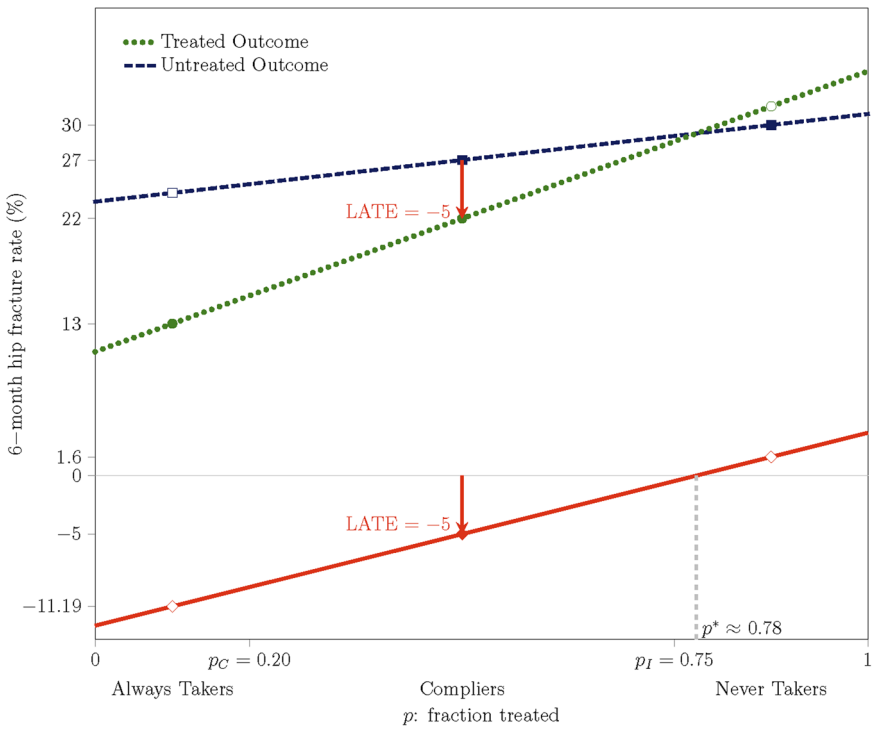

FIGURE 11 Test for treatment effect homogeneity rejects under stronger ancillary assumptions of linearity of untreated outcomes and linearity of treatment effects: Treatment effect estimates [Color figure can be viewed at wileyonlinelibrary.com]

linear across always takers, compliers, and never takers, as depicted in Figure 10. Linear variation in the comorbidity score can support ancillary assumptions of linearity of untreated outcomes and treatment effects.

Figure 11 demonstrates the implications of the ancillary linearity assumptions in the surgery example. As in Kowalski (forthcoming), I refer to the function that specifies how untreated outcomes vary with the fraction treated p

as the marginal untreated outcome function MUO(p) for consistency with the corresponding function for treatment effects, which is commonly called the marginal treatment effect function MTE(p). These functions are marginal functions in the sense that it is possible to obtain average quantities from them using weights over the fraction treated p following Brinch et al. (2017), Carneiro et al. (2011), Heckman and Vytlacil (1999, 2001b, 2005), and Kowalski (forthcoming). Linearity of untreated outcomes and treatment effects in the fraction treated p implies that the MUO and MTE functions are linear. The sum of the MUO and MTE functions yields the marginal treated outcome function MTO(p). The MTO function is also linear in Figure 11 because the MUO and MTE functions are linear.

The ancillary linearity assumptions preserve the LATE of −5 while also yielding an estimate of the treatment effect at every fraction treated p, as depicted by the marginal treatment effect function MTE(p) in Figure 11. As shown, the MTE function has a positive slope. Statistical significance of the slope can be obtained via bootstrap as in Brinch et al. (2017) and Kowalski (forthcoming). I do not recommend plotting confidence intervals for the MTE function, as doing so would convey statistical significance at a particular fraction treated p, which could be misleading about the statistical significance of quantities derived from the MTE function, such as its slope. Instead, I recommend reporting confidence intervals on those quantities directly. If the slope of the MTE function is statistically significant, the test for treatment effect homogeneity rejects under the ancillary linearity assumptions.

The positively sloped MTE function in Figure 11 implies that the LATE cannot be externally valid for all policies. However, the researchers are interested in whether the LATE is externally valid with respect to two specific options. The first option is to mandate outpatient surgery for everyone. This option would enroll all never takers. The average treatment effect obtained by averaging the MTE function over the support of the fraction treated p for never takers shows that an outpatient surgery mandate would increase the average subsequent hip fracture rate for never takers by approximately 1.6 percentage points. Therefore, it is not sensible to choose a mandate over the policy of recommending outpatient surgery as the default.

However, instead of a mandate, should the researchers provide a subsidy to patients who undergo outpatient surgery? To understand the impact of this option, the researchers assume that anyone who is shifted into treatment by the recommendation cannot be shifted out by the subsidy and vice versa.[3] Given the slope of the MTE function, the researchers want to set the subsidy just large enough to expand the fraction treated to the optimal level of approximately 78%, denoted by p^* in Figure 11, so that only the patients who benefit from outpatient surgery are induced to receive it. In practice, they can adjust the subsidy over time if they overshoot or undershoot.

Covariates can also offer guidance on how to set the subsidy. To determine which covariates explain treatment effect heterogeneity, researchers can incorporate them into an MTE function under shape restrictions, as demonstrated in Kowalski (forthcoming). In the surgery example, linearity of the baseline comorbidity score in the fraction treated p implies that incorporating it into the linear MTE function under an additive separability restriction will make the MTE function flat, indicating that this covariate explains all treatment effect heterogeneity. The increase in the comorbidity score in the fraction treated p implies that the level of the MTE function for each value of the comorbidity score will increase as the comorbidity score increases. As long as the MTE function for a particular value of the comorbidity score is negative, patients with that value of the comorbidity score will benefit. This result could be useful to the researchers pursuing the subsidy option because instead of guessing the right subsidy level, they could offer outpatient surgery only to those patients who have low enough comorbidity scores to benefit.

Subgroup analysis does not offer similar insights into the optimal fraction treated p. It does not require ancillary assumptions or shape restrictions, but it delivers more limited results. In the surgery example, although the researchers estimated a negative LATE within all subgroups that they defined by age, sex, and comorbidity score, these LATEs only apply to compliers. The LATEs can possibly be negative within all subgroups even if there are positive treatment effects for always or never takers. Furthermore, subgroup analysis and LATE-reweighting are limited by which covariates the researchers have and how they use them to form subgroups. Researchers relying solely on subgroup analysis might erroneously conclude that treatment effects are homogeneous if they lack the right covariates. In contrast, if the linear MTE function has a nonzero slope and none of the available covariates make it flat, then they will know to keep looking for other covariates that can explain treatment effect heterogeneity. Even in the extreme case where values of the comorbidity score perfectly distinguish between always takers, compliers, and never takers, subgroup analysis need not completely uncover treatment effect heterogeneity because the LATEs cannot be estimated within subgroups in which all participants are treated or untreated. If researchers only report subgroup analysis among subgroups in which they can obtain estimates, they can erroneously conclude that all treatment effects are negative, even if treatment effects are positive within subgroups determined by a covariate that completely explains treatment effect heterogeneity.

Alternative ancillary assumptions that identify the MTE function at every fraction treated p also allow for tests of treatment effect heterogeneity and estimates of average treatment effects for always and never takers. For example, Kline and Walters (2019) show that the distributional assumptions made by the "Heckit" estimator of Heckman (1979) and the estimator used by Mroz (1987) identify the MTE function at every fraction treated p. The assumptions made by Willis and Rosen (1979) also identify the MTE function at every fraction treated p. As another example, Brinch et al. (2017) propose that MUO and MTO functions are quadratic and monotonic over the fraction treated from 0 to 1, and those assumptions identify the MTE function at every fraction treated p. If covariates are available, it is also possible to use them to estimate more flexible forms for the MTE function, as in Brinch et al. (2017), Carneiro et al. (2011), Carneiro and Lee (2009), Kline and Walters (2016), Kowalski (2016), and Maestas et al. (2013) under additional shape restrictions.

While I consider expansion options within the hospital in the surgery example, researchers can also use the MUO and MTE functions from one hospital to examine potential expansion options at other hospitals. Evaluating whether a treatment effect estimated in one context is externally valid to policies in another context requires an additional assumption that both contexts have the same underlying MUO and MTE functions. I refer interested readers to Kowalski (forthcoming), where I demonstrate how examination of covariates, institutional details, and related outcomes can motivate such an assumption.

6 | IMPLICATIONS FOR EXPERIMENTAL DESIGN

To strengthen the case for external validity, researchers should design the new policy implemented within the intervention arm to be as similar as possible to the policy of interest. Heckman and Vytlacil (2001a, 2007) make this insight clear with the concept of "policy-relevant treatment effects." For any treatment, there can be multiple policies that affect takeup, and each policy can generate different sets of always takers, compliers, and never takers. For example, never takers in an experiment that involves a mailing might be compliers in an experiment that involves a phone call. Therefore, the average treatment effects for always takers, compliers, and never takers—and more generally, the shape of the MTE function—depend on the new policy introduced in the intervention arm.

Researchers risk weakening the case for external validity if they design the new policy to force perfect compliance when the policy of interest would not force all individuals to receive treatment. Some researchers do so to increase power. Others do so with the goal of estimating a LATE that can be interpreted as the average treatment effect in a given population. However, unless the policy of interest would also force all individuals to receive treatment, an experiment with perfect compliance is not superior to an experiment with noncompliance for purposes of external validity.

Researchers also risk weakening the case for external validity if they design the new policy to generate always and never takers that would not arise under the policy of interest, either because the policy of interest would involve *different* always or never takers or because it would not involve *any* always or never takers. Some policies naturally do not involve any always takers, especially if they introduce treatments that are not available otherwise. If there are no always takers, the untreated outcome test can still identify heterogeneous selection. Without always and never takers, the tests for heterogeneous treatment effects discussed in this paper cannot be applied without further assumptions.

Researchers should only design experiments to apply the tests discussed in this paper if those tests will inform external validity to policies of interest. In the insurance example, because the insurer was considering a policy that would enroll everyone in the wellness plan, it would have been better served by an experiment that would do just that, even though there would have been no never takers. Instead of introducing a new policy within the experiment that involved an informational mailing, it could have tested its two main policy options directly.

Researchers interested in external validity to a range of policies might want to consider designing experiments that introduce a range of policies. Several experimental designs involve a range of policies, including those discussed in Ashraf et al. (2010), Basu (2015), Berry et al. (2020), Burtless and Hausman (1978), Chassang et al. (2012), and Narita (2018), among others. In the insurance example, researchers could implement such a design by offering a range of subsidies for enrollment in the wellness plan.

Designs that involve a range of policies potentially involve a loss of power, but they have important advantages for the examination of external validity. Specifically, they have the potential to identify selection and treatment effect heterogeneity even if there are no always and never takers. Under monotonicity conditions recently formalized by Mogstad et al. (2020), researchers can identify selection and treatment effect heterogeneity over a range of the fraction treated. Such experimental variation in the fraction treated can also allow researchers to investigate the concern that

treatment effects change as the fraction treated changes because of general equilibrium factors, especially if the variation is across experiments as in Lee et al. (2020).

Beyond designing the policies introduced within experiments thoughtfully, researchers should collect data to facilitate examination of external validity. To apply the approaches discussed in this paper, it is imperative to collect data that allow tabulations of outcomes by lottery status *and* treatment. It is also useful to collect data that allow similar tabulations of covariates. Data on covariates can also facilitate comparisons across experiments (Angrist & Fernandez-Val, 2013; Hotz et al., 2005; Wager & Athey, 2018). Approaches to assess external validity across experiments are even more powerful when used in concert with approaches to assess external validity within experiments.

ACKNOWLEDGMENTS

I thank Neil Christy, Simon Essig Aberg, Ryan Fraser, Katie Laursen, Pauline Mourot, Srajal Nayak, Ljubica Ristovska, Sukanya Sravasti, Rae Staben, and Matthew Tauzer for excellent research assistance. NSF CAREER Award 1350132 provided support. I thank Magne Mogstad, Jeffrey Smith, Edward Vytlacil, the editor, coeditor, and anonymous referees for helpful feedback.

ORCID

Amanda E. Kowalski https://orcid.org/0000-0002-0221-8826

ENDNOTES

[1] These approaches also allow researchers to estimate the *distributions* of outcomes of treated and untreated compliers, which paves the way for examination of treatment effect heterogeneity *within* compliers, as in Heckman et al. (1997). Here, I focus on treatment effect heterogeneity across always takers, compliers, and never takers.

[2] In Kowalski (2021), I propose to implement the test in this way, which is straightforward to implement with a bootstrap procedure. Brinch et al. (2017) take a different approach. They test the signs of the treated and untreated outcome test statistics separately and then test if they are equal by accounting for multiple hypothesis testing using a Bonferroni correction, which could decrease power relative to the implementation I propose.

[3] I thank an anonymous referee for suggesting this characterization of a "common monotonicity" assumption.

REFERENCES

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, *97*(457), 284–292.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, *113*(2), 231–263.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, *80*, 313–336.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, *66*(2), 249–288.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal*, *114*(494), C52–C83.

Angrist, J. D., & Fernandez-Val, I. (2013). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In *Advances in economics and econometrics: Volume 3, econometrics: Tenth world congress* (Vol. 51, pp. 401). Cambridge University Press.

Angrist, J. D., & Krueger, A. B. (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, *87*(418), 328–336.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Ashraf, N., Berry, J., & Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *The American Economic Review*, *100*(5), 2383–2413.

Balke, A. and Pearl, J. (1993). *Nonparametric bounds on causal effects from partial compliance data*. Working paper. https://escholarship.org/uc/item/2rn0420q

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Basu, A. (2015). Welfare implications of learning through solicitation versus diversification in health care. *Journal of Health Economics*, *42*, 165–173.

Berry, J., Fischer, G., & Guiteras, R. (2020). Eliciting and utilizing willingness to pay: Evidence from field trials in Northern Ghana. *Journal of Political Economy*, *128*(4), 1436–1473.

Bertanha, M., & Imbens, G. W. (2014). *External validity in fuzzy regression discontinuity designs* (Working Paper 20773). National Bureau of Economic Research. https://www.nber.org/papers/w20773

Björklund, A., & Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 42–49.

Black, D. A., Joo, J., LaLonde, R., Smith, J. A., & Taylor, E. J. (2017). *Simple tests for selection bias: Learning more from instrumental variables* (Working Paper 6932). CESifo. https://www.cesifo-group.de/DocDL/cesifo1_wp6392.pdf

Brinch, C. N., Mogstad, M., & Wiswall, M. (2017). Beyond LATE with a discrete instrument. *Journal of Political Economy*, *125*(4), 985–1039.

Burtless, G., & Hausman, J. A. (1978). The effect of taxation on labor supply: Evaluating the gary negative income tax experiment. *Journal of Political Economy*, *86*(6), 1103–1130.

Carneiro, P., & Lee, S. (2009). Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, *149*(2), 191–208.

Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, *101*(6), 2754–81.

Chassang, S., Miquel, G. P. I., & Snowberg, E. (2012). Selective trials: A principal-agent approach to randomized controlled experiments. *American Economic Review*, *102*(4), 1279–1309.

Cornelissen, T., Dustmann, C., Raute, A., & Schönberg, U. (2018). Who benefits from universal child care? Estimating marginal returns to early child care attendance. *Journal of Political Economy*, *126*(6), 2356–2409.

Einav, L., Finkelstein, A., & Cullen, M. R. (2010). Estimating welfare in insurance markets using variation in prices. *The Quarterly Journal of Economics*, *125*(3), 877.

Guo, Z., Cheng, J., Lorch, S. A., & Small, D. S. (2014). Using an instrumental variable to test for unmeasured confounding. *Statistics in Medicine*, *33*(20), 3528–3546.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251–1271.

Heckman, J., Hohmann, N., Smith, J., & Khoo, M. (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *The Quarterly Journal of Economics*, *115*(2), 651–694.

Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, *96*(8), 4730–4734.

Heckman, J. J., & Vytlacil, E. J. (2001a). Policy-relevant treatment effects. *American Economic Review*, *91*(2), 107–111.

Heckman, J. J., & Vytlacil, E. J. (2001b). Local instrumental variables. In C. Hsiao, K. Morimune and J. L. Powell (Eds.), *Nonlinear statistical modeling: Proceedings of the thirteenth international symposium in economic theory and econometrics: Essays in honor of takeshi amemiya* (pp. 1–46). Cambridge University Press.

Heckman, J. J., & Vytlacil, E. J. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, *73*(3), 669–738.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of Econometrics*, *6*, 4875–5143.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, *64*(4), 487–535.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*(5), 1017–1098.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–162.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*(1), 241–270.

Huber, M. (2013). A simple test for the ignorability of non-compliance in experiments. *Economics Letters*, *120*(3), 389–391.

Imbens, G. W., and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–75.

Imbens, G. W., & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, *64*(4), 555–574.

Katz, L. F., Kling, J. R., & Liebman, J. B. (2001). Moving to opportunity in Boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics*, *116*(2), 607–654.

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics*, *131*(4), 1795–1848.

Kline, P., & Walters, C. R. (2019). On heckits, LATE, and numerical equivalence. *Econometrica*, *87*(2), 677–696.

Kowalski, A. (2016). *Doing more when you're running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments* (Working Paper 22362). National Bureau of Economic Research. http://www.nber.org/papers/w22362

Kowalski, A., Tran, Y., & Ristovska, L. (2018). *MTEBINARY: Stata module to compute Marginal Treatment Effects (MTE) With a Binary Instrument*. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458285.html

Kowalski, A. E. (2021). *Behavior within a clinical trial and implications for mammography guidelines* (Working Paper 25049). National Bureau of Economic Research. http://www.nber.org/papers/w25049

Kowalski, A. E. (forthcoming). Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform. *Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_01069

Lee, K., Miguel, E., & Wolfram, C. (2020). Experimental evidence on the economics of rural electrification. *Journal of Political Economy*, *128*(4), 1523–1565.

Maestas, N., Mullen, K. J., & Strand, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *The American Economic Review*, *103*(5), 1797–1829.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, *80*(2), 319–323.

Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment effects. *Econometrica*, *86*(5), 1589–1619.

Mogstad, M., Torgovitsky, A., & Walters, C. R. (2020). *Policy evaluation with multiple instrumental variables* (Working Paper 27546). National Bureau of Economic Research. http://www.nber.org/papers/w27546

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, *55*(4), 765–799.

Narita, Y. (2018). *Toward an ethical experiment* (Working Paper 2127). Cowles Foundation. https://cowles.yale.edu/sites/default/files/files/pub/d21/d2127.pdf

Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica*, *48*(7), 1815–1820.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, & A. Mulley (Eds.), *Health service research methodology: A focus on AIDS* (pp. 113–159). Public Health Service.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, *70*(1), 331–341.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, *11*(3), 284–300.

Willis, R. J., & Rosen, S. (1979). Education and self-selection. *Journal of Political Economy*, *87*(5, Part 2), S7–S36.

How to cite this article: Kowalski, A. E. (2022). How to examine external validity within an experiment. *Journal of Economics & Management Strategy*, 1–19. https://doi.org/10.1111/jems.12468