NBER WORKING PAPER SERIES

# HOW TO EXAMINE EXTERNAL VALIDITY WITHIN AN EXPERIMENT

Amanda E. Kowalski

Working Paper 24834
http://www.nber.org/papers/w24834

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2018

I thank Pauline Mourot, Ljubica Ristovska, Sukanya Sravasti, Rae Staben, and Matthew Tauzer for excellent research assistance. NSF CAREER Award 1350132 provided support. I thank Magne Mogstad, Edward Vytlacil, and three anonymous referees for helpful feedback. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

How to Examine External Validity Within an Experiment
Amanda E. Kowalski
NBER Working Paper No. 24834
July 2018
JEL No. C9,C93,H0

## ABSTRACT

A fundamental concern for researchers who analyze and design experiments is that the experimental result might not be externally valid for all policies. Researchers often attempt to assess external validity by comparing data from an experiment to external data. In this essay, 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 figures.

Amanda E. Kowalski
Department of Economics
University of Michigan
37 Hillhouse Avenue
611 Tappan Ave.
Lorch Hall 213
Ann Arbor, MI 48109-1220
and NBER
aekowals@umich.edu

# 1 Introduction

The traditional reason that a researcher runs an experiment is to address selection into treatment. For example, 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 experiments. In many experiments, some lottery losers receive treatment and some lottery winners forgo treatment. Throughout this essay, I consider experiments in which both occur, experiments with "two-sided noncompliance." In these experiments, individuals participate in a lottery. Individuals who win the lottery are in the intervention group. They receive an intervention that affects selection into treatment. Individuals who lose the lottery are in the control group, and they do not receive the intervention. However, all individuals can select to receive or forgo the treatment.

Some researchers view this type of selection as immaterial, and they discard information on it 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 it 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.

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. An experiment re-conceived as a tool that identifies heterogeneous treatment effects can itself inform external validity. If treatment effects vary across groups within an experiment, then there is no single treatment effect that is externally valid for all policies.

In this essay, I discuss techniques from the treatment effects literature that researchers can use to begin examination of external validity within an experiment. These techniques are useful because of the tight relationship between treatment effect homogeneity and external validity. I do not break new ground in terms of methodology, and I do not aim to be comprehensive. Rather, I aim to present some existing methods simply using figures, making them readily accessible to researchers who evaluate and design experiments.

One of the virtues of experiments is that traditional analysis is straightforward, and it relies on assumptions that are well-known. Throughout this essay, I proceed under the well-

known local average treatment effect (LATE) assumptions of independence and monotonicity proposed by Imbens and Angrist (1994). Vytlacil (2002) constructs a model of selection into treatment that assumes no more than the LATE assumptions, and I use it as the foundation for my analysis. 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 Heckman and Vytlacil (1999, 2001b, 2005), Carneiro et al. (2011), Brinch et al. (2017) and Kowalski (2016, 2018). Therefore, the model that serves as the foundation for my analysis also serves at the foundation for the LATE and MTE approaches within the treatment effects literature. I do not present the model here. Instead, I focus on depicting its implications graphically.

In Section 2, I begin by depicting information that is necessary for traditional analysis of an experiment. Next, I include additional information that is available under the model. This additional information consists of shares and outcomes of always takers, compliers, and never takers, using the terminology of Angrist et al. (1996), obtained following Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003).

In Section 3, I depict a test for heterogeneous selection into treatment that uses a subset of the additional information and no ancillary assumptions. This test is equivalent to tests proposed in the econometric literature by Bertanha and Imbens (2014); Guo et al. (2014); Black et al. (2015), and Mogstad et al. (2017). Under some circumstances, it is also equivalent to the Einav et al. (2010) test from the insurance literature. In Kowalski (2016, 2018), I refer to this test as the "untreated outcome test," and my innovation is in the interpretation – I show that it identifies heterogeneous selection without any assumptions beyond the LATE assumptions. This test for heterogeneous selection is a natural precursor to a test for external validity because outcomes can differ across groups due to heterogeneous selection and heterogeneous treatment effects. The key to isolating heterogeneous treatment effects, which inform external validity, is to first isolate heterogeneous selection.

In Section 4, I depict a test for external validity proposed by Brinch et al. (2017) and applied in Kowalski (2016). The Mogstad et al. (2017) approach can be used for inference. Brinch et al. (2017) conduct this test under two ancillary assumptions. As I show in Kowalski (2016), it is possible to conduct the test under only one of their ancillary assumptions; either one will suffice. I also show that each ancillary assumption implies an upper or lower bound on the average treatment effect for always or never takers. These bounds help me to demonstrate how the ancillary assumptions combine with information on always and never takers to test external validity. If either bound does not include the LATE, then the LATE cannot be externally valid for all policies. Bounds on average treatment effects for always and never takers are also of interest in their own right. When the LATE is not externally

valid, these bounds inform the signs and magnitudes of the heterogeneous treatment effects induced by hypothetical policies.

Other tests proposed by Hausman (1978); Heckman (1979); Willis and Rosen (1979); Angrist (2004); Huber (2013); Bertanha and Imbens (2014); Guo et al. (2014); Black et al. (2015) and Brinch et al. (2017) rely on stronger assumptions to conduct more powerful tests of external validity. In Section 5, I engage with these tests by discussing how stronger assumptions yield estimates of treatment effects in lieu of bounds. I conclude by discussing implications for experimental design in Section 6.

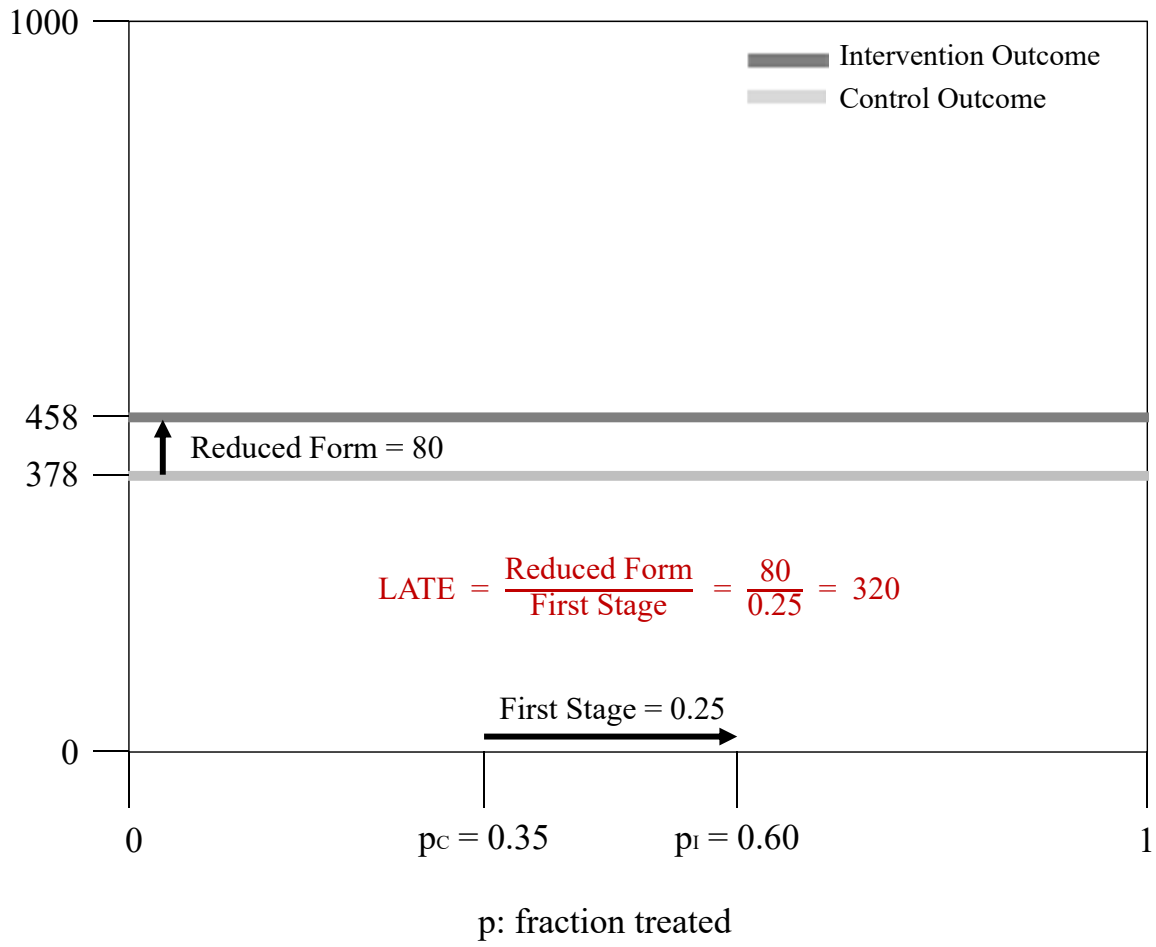
# 2 An Experiment under the LATE Assumptions

In the data from an experiment, suppose that researchers can observe whether each individual won the lottery, whether each individual received the treatment, and an outcome for each individual. Traditional analysis of an experiment begins by comparing the average outcomes of the intervention group and the control group. In Figure 1, I depict results from a hypothetical experiment in which the average outcome in the intervention group is 80 units higher than the average outcome in the control group. This difference in average outcomes is often called the "reduced form," as labeled along the vertical axis. It gives an estimate of the impact of the intervention that lottery winners receive 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 on the outcome. Calculation of the reduced form does not even require data on treatment. Some researchers report only the reduced form.

Traditional analysis of an experiment next compares the average treatment probabilities for lottery losers and winners. By the LATE independence assumption, lottery status is independent of treatment, so I can depict the average treatment probabilities for lottery losers and winners along the same horizontal axis in Figure 1. As depicted, $p_C$ represents the probability of treatment in the control group, and $p_I$ represents the probability of treatment in the intervention group. 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 the example depicted in Figure 1, 35% of lottery losers receive treatment and 60% of lottery winners receive treatment, so the first stage implies that winning the lottery increases the fraction treated by 25 percentage points.

To obtain an estimate of the impact of the treatment on the outcome, traditional analysis of an experiment divides the reduced form by the first stage. This quotient gives the local

Figure 1: Average Outcomes of Intervention and Control Groups
Under LATE Assumptions




average treatment effect (LATE) of Imbens and Angrist (1994). Traditional analysis of an experiment reports the LATE as the single treatment effect that the experiment isolates. In the example depicted in Figure 1, the LATE is equal to 320 (=80/0.25). Under the LATE assumptions, the LATE gives the average treatment effect on "compliers," individuals whose treatment status is determined by their random assignment, in the terminology of Angrist et al. (1996).

Experiments with two-sided noncompliance also include two other 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 assumptions rule out the presence of "defiers" who take up treatment if and only if they lose the lottery, so there are only always takers, compliers, and never takers. In experiments with two-sided noncompliance, researchers cannot identify whether each individual is an always taker, never taker, or complier: lottery winners who take up treatment could be always takers

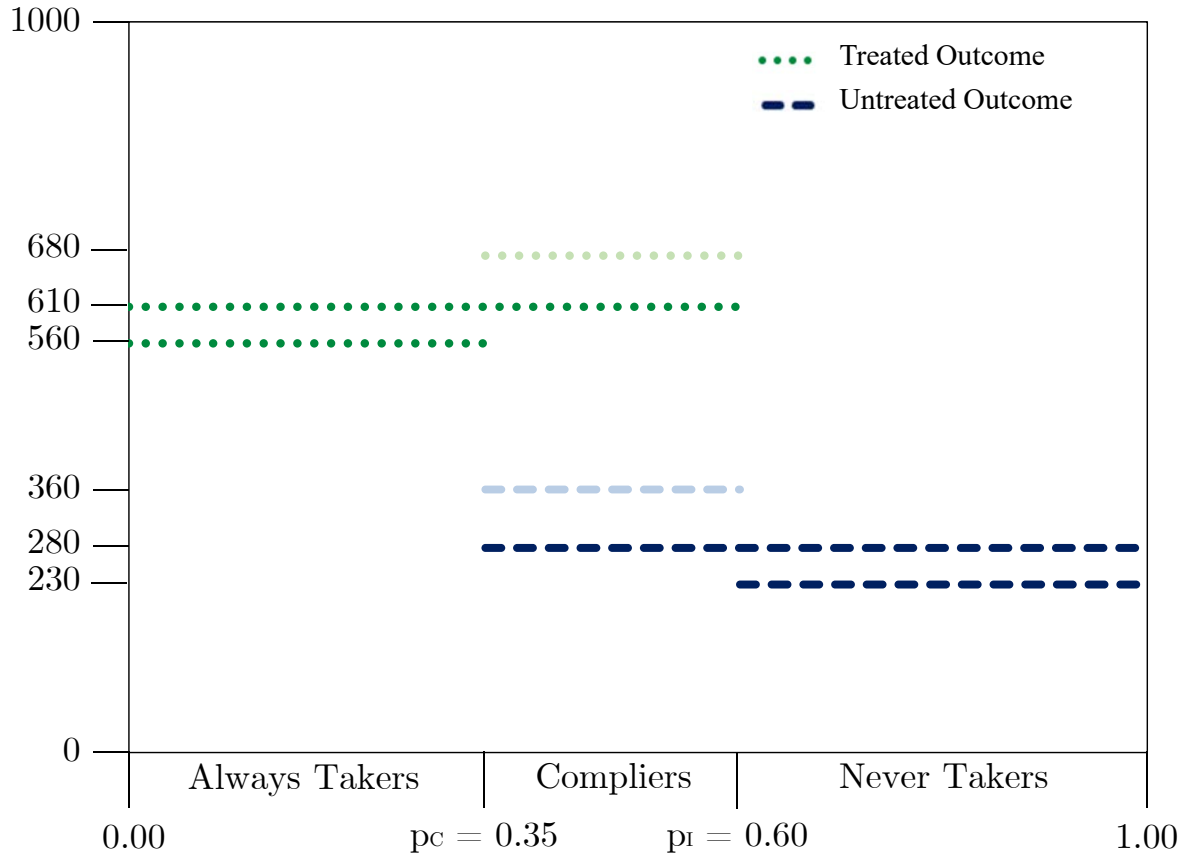
or compliers; lottery losers who do not take up treatment could be compliers or never takers. However, researchers can identify some individuals as always takers and other individuals as never takers. Lottery losers who take up treatment must be always takers; lottery winners who do not take up treatment must be 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 gives the share of always takers in the full sample, and the observed share of untreated lottery winners gives the share of never takers in the full sample. Furthermore, because always and never takers do not change their treatment decisions 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 calculate the average outcomes of treated and untreated compliers, as demonstrated by Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003).

To illustrate the calculation of average outcomes of always takers, compliers, and never takers graphically, I continue the hypothetical example in Figure 2. As originally shown by 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, in the order that they receive treatment. 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.

Along the vertical axis of Figure 2, I plot the average treated and untreated outcomes of the intervention and control groups over the relevant ranges of the horizontal axis. As shown, the average treated outcome in the intervention group is 610, which represents a weighted average of the treated outcomes of always takers and compliers. The average treated outcome in the control group is 560, which represents the average treated outcome of always takers. Because always takers make up 35% of the full sample and always takers combined with compliers make up 60% of the full sample, the average treated outcome of compliers is 680 (= (0.6/(0.6-0.35))*610 - (0.35/(0.6-0.35))*560), as depicted in light shading. Similar logic using the untreated outcomes implies that the average untreated outcome of never takers is 230 and that the average untreated outcome of compliers is 360 (= ((1-0.35)/(0.6-0.35))*280 - ((1-0.60)/(0.6-0.35))*230), as depicted in light shading. Researchers who would like to replicate the calculations in this paper can use the Stata command *mtebinary*, which

Figure 2: Average Treated and Untreated Outcomes of Intervention and Control Groups and Average Treated and Untreated Outcomes of Compliers Under LATE Assumptions




includes examples based on the same hypothetical data that I use here (Kowalski et al., 2016).

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 calculated 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 calculate the LATE via the Wald (1940) approach using separate datasets for the reduced form and first stage. Because the LATE can be calculated using limited data, it stands to reason that it does not capture all available information. Accordingly, Figure 3 provides additional information relative to

Figure 3: Average Outcomes of Always Takers, Compliers, and Never Takers Under LATE Assumptions

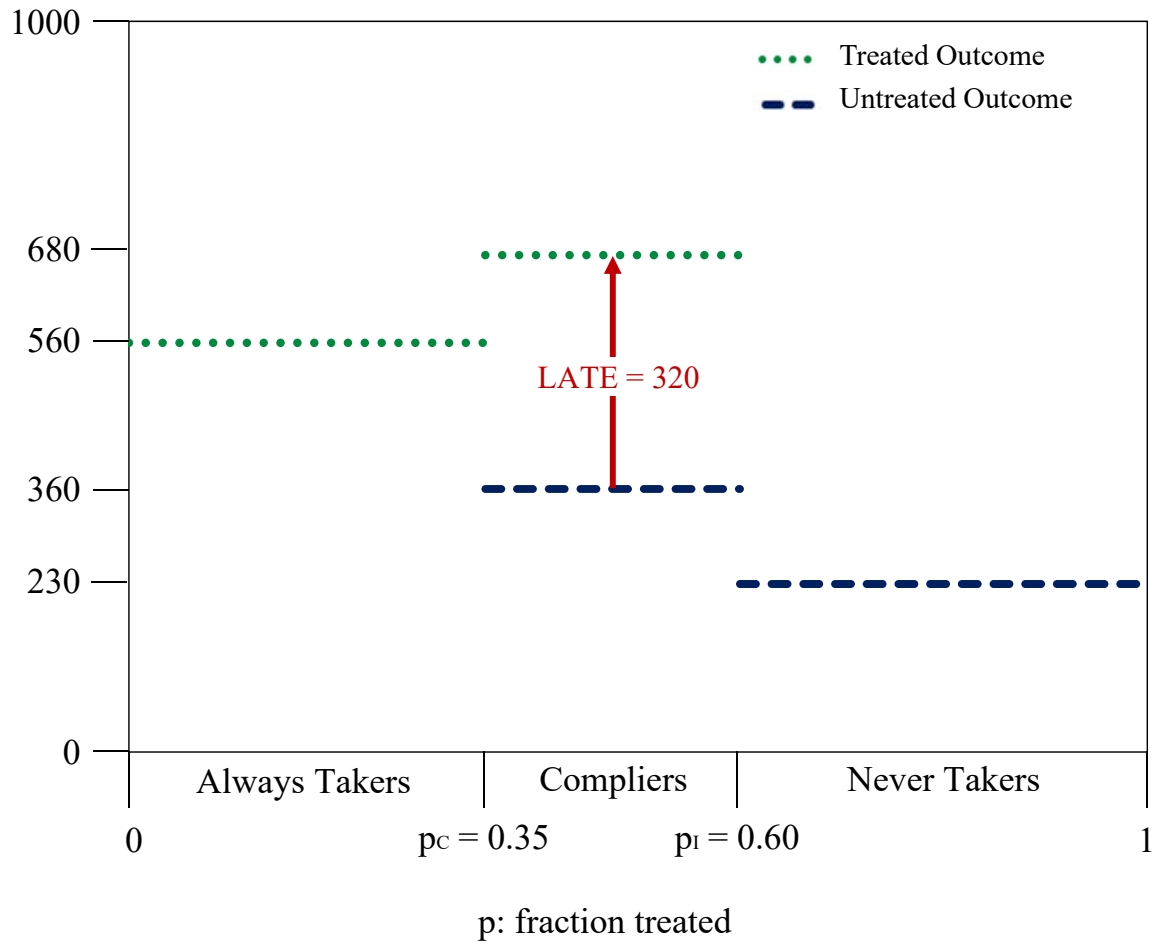



Figure 1.

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 hypothetical example, these groups are sizeable. Furthermore, the average treated outcome of always takers is known, and the average untreated outcome of never takers is known. The average untreated outcome of always takers is not known, and the average treated outcome of never takers is 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 the average treated and untreated outcomes for each group. Similarly, if they could be bounded, then would be possible to bound on the average treatment effect on each group. Such bounds could be implied by natural bounds on the range of outcomes in the tradition of Manski (1990), or they could be implied by ancillary assumptions.

Even in the absence of ancillary assumptions, a researcher examining the hypothetical example depicted in Figure 3 might question whether the LATE is likely to be equal to the

average treatment effect for always and never takers, given that average outcomes for always takers, compliers, and never takers appear to be so different. I formalize that intuition in the next sections. I begin by testing whether the average outcomes are statistically different, and then I use the differences to inform ancillary assumptions that allow for tests of external validity.

# 3 Test for Heterogeneous Selection under the LATE Assumptions

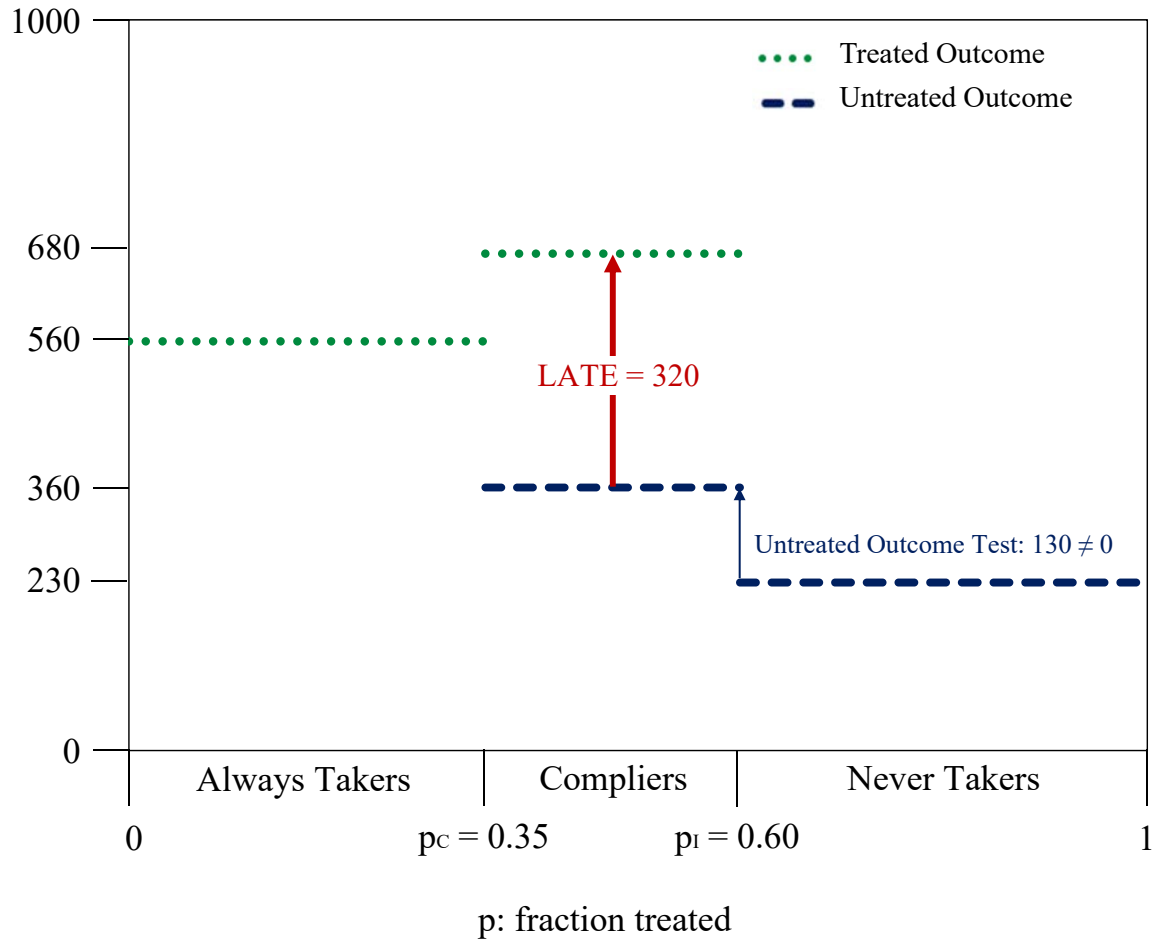
As I discuss in Kowalski (2016, 2018), the test of the null hypothesis that the difference in average untreated outcomes between compliers and never takers is equal to zero can be interpreted as a test for heterogeneous selection that does not require any assumptions beyond the LATE assumptions. If the difference in average untreated outcomes between compliers and never takers is statistically different from zero, then the test rejects selection homogeneity. Because it compares untreated outcomes, I refer to the test as the "untreated outcome test." This test is equivalent to tests proposed in the econometric literature by Bertanha and Imbens (2014); Guo et al. (2014); Black et al. (2015), and Mogstad et al. (2017). It is also equivalent to the "cost curve" test of Einav et al. (2010) from the insurance literature when the untreated outcome is uninsured costs.

The logic behind why the untreated outcome test identifies heterogeneous selection 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 hypothetical example, Figure 4 shows that the average untreated outcome of compliers is 130 higher than the average outcome of never takers. If this difference is statistically different from zero, then the test rejects selection homogeneity. Compliers select into treatment before never takers, as shown along the horizontal axis. Therefore, individuals with higher average outcomes select into treatment before individuals with lower average outcomes, and the untreated outcome test statistic provides evidence of positive selection.

Empirically, the untreated outcome test can show positive or negative selection. If the average untreated outcome of compliers were lower than the average untreated outcome of never takers, then the untreated outcome test statistic would be negative, indicating negative selection. Within the same experiment, the untreated outcome test can show positive selection on some outcomes while showing negative selection on others. If the outcome is insurance, then a positive value of the untreated outcome test indicates "adverse selection" into insurance and negative value indicates "advantageous selection" into insurance, per the

Figure 4: Untreated Outcome Test Rejects
Test Statistic Shows Positive Selection
Under LATE Assumptions




cost curve test of Einav et al. (2010).

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 proposed in the econometric literature by Bertanha and Imbens (2014); Guo et al. (2014); Black et al. (2015), and Mogstad et al. (2017). In the insurance literature, the treated outcome test is equivalent to the "cost curve" test of Einav et al. (2010) when the untreated outcome is *insured* costs. In Kowalski (2016, 2018), I emphasize that the treated outcome test does not isolate heterogeneous selection. Intuitively, treated outcomes can reflect selection *and* treatment effects. Therefore, a difference in treated outcomes can reflect heterogeneous selection and heterogeneous treatment effects.

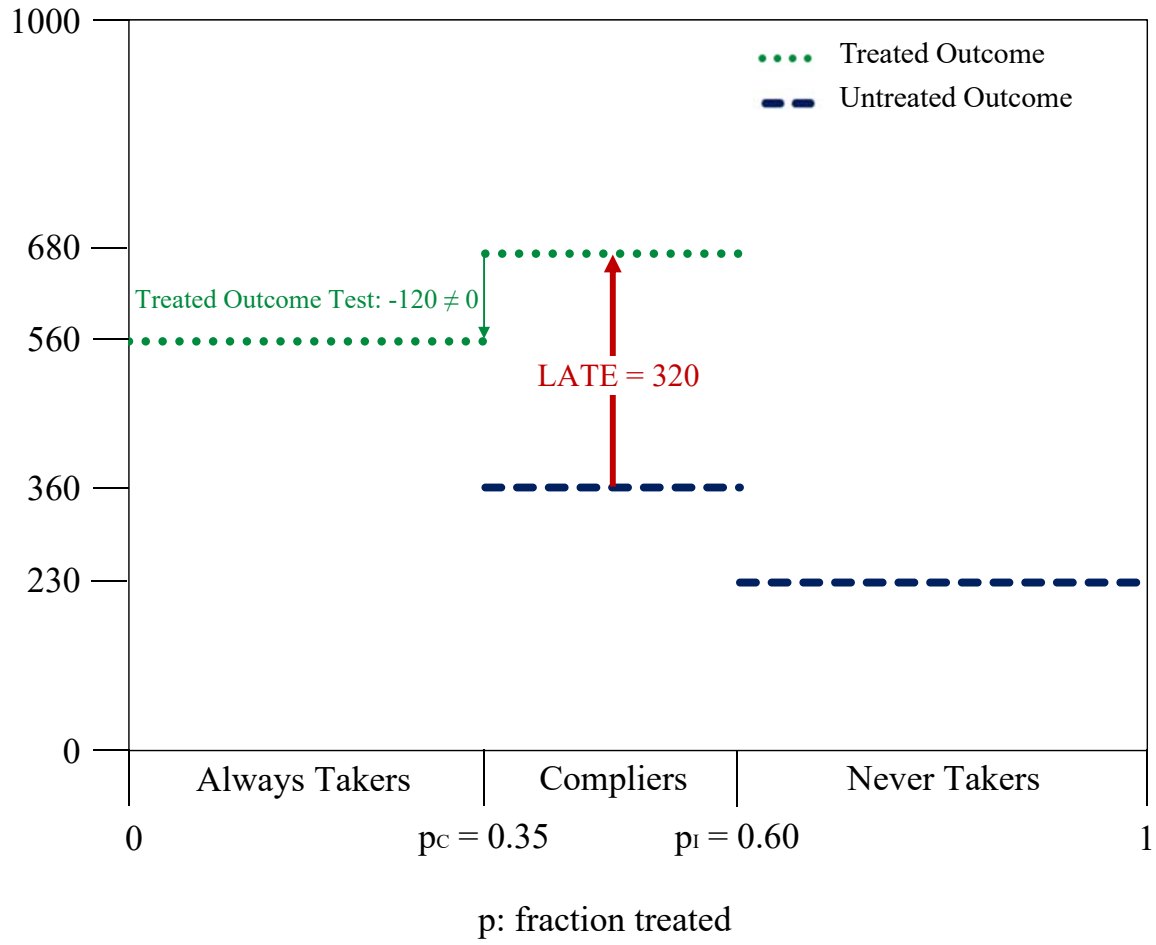
Continuing the hypothetical example, consider the implications of the treated outcome test depicted in Figure 5. The treated outcome test shows that the average outcome of always takers is 120 lower than the average outcome of compliers. As stated, the treated

outcome test statistic is statistically different from zero, so the treated outcome test rejects. This result could be entirely due to heterogeneous selection from always takers to never takers, 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 320 because the LATE is the average treatment effect for compliers. Therefore, the average untreated outcome of always takers would be 240 (=560-320). Alternatively, the result of the treated outcome test could be entirely due to treatment effect heterogeneity from always takers to never takers, 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 360. It is also possible that the treated outcome test could detect a combination of selection and treatment effect heterogeneity, which would be the case if the average untreated outcome of always takers were anything other than 240 or 360. As this example demonstrates, the treated outcome test can reflect various combinations of selection and 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 (2016, 2018), the distinction between treated and untreated is material to the definition of the treatment effect. 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 (LATE) with an arrow in the figures. It is tempting to think that renaming the treated the untreated and vice versa would have no consequence, but such a swap would change the direction of the arrow. In that case, the treated outcome test would detect only selection, and the untreated outcome test would detect various combinations of selection and treatment effect heterogeneity, creating a different but no less material distinction between the tests.

The distinction between heterogeneity in treated and untreated outcomes forms the foundation for tests for external validity. In this essay, tests for treatment effect homogeneity are tests for external validity. The key to testing for treatment effect homogeneity is to first test for selection heterogeneity using the untreated outcome test and to then impose ancillary assumptions to purge selection heterogeneity from the treated outcome test so the only treatment effect heterogeneity remains.

Figure 5: Treated Outcome Test Rejects
Test Statistic Shows Negative Selection and/or Treatment Effect Heterogeneity
Under LATE Assumptions

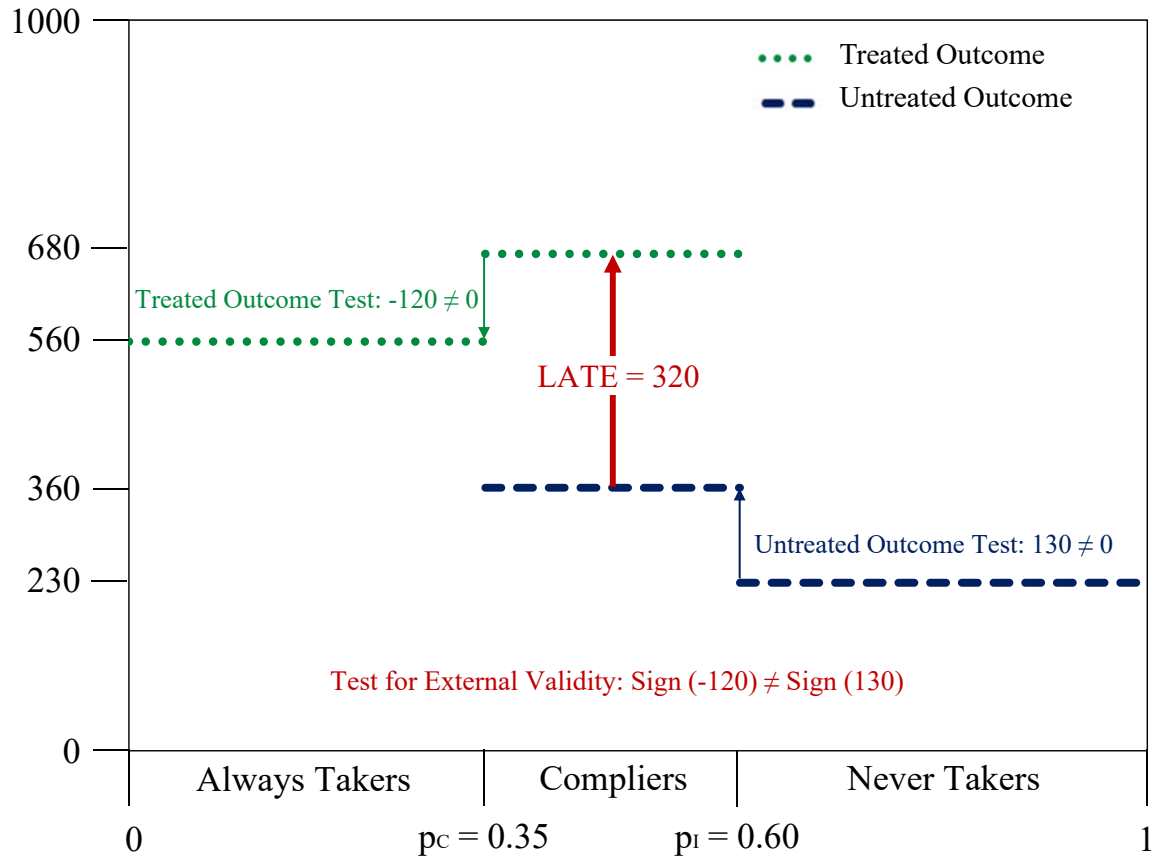



# 4 Test for External Validity under Ancillary Assumptions

In Figure 6, I depict a test for external validity proposed by Brinch et al. (2017) and applied in Kowalski (2016). The Mogstad et al. (2017) approach can be used for inference. The test rejects the null hypothesis of treatment effect homogeneity if the sign of the untreated outcome test statistic is not equal to the sign of the treated outcome test statistic. The intuition behind why this test for treatment effect homogeneity is also a test for external validity is that the LATE can only be externally valid for all policies if the treatment effect is homogeneous. If the treatment effect is homogeneous, then the treated outcome test and the untreated outcome test reflect only selection. If the untreated outcome test implies positive selection but the treated outcome test implies negative selection in the absence of treatment effect heterogeneity, then the treatment effect cannot be homogeneous, and the LATE cannot

be externally valid for all policies.

Figure 6: Test for External Validity Rejects Under Ancillary Assumptions



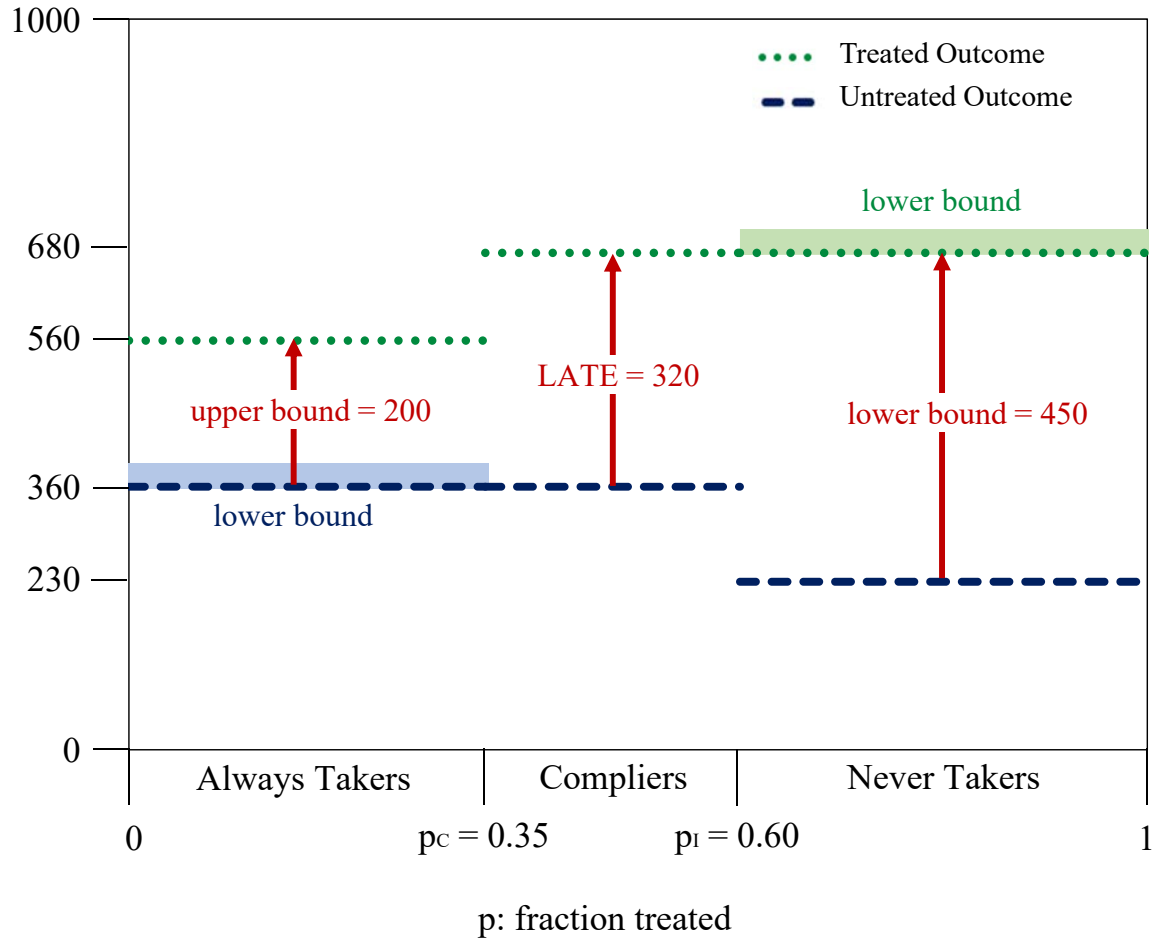

Brinch et al. (2017) conduct this test under two ancillary assumptions: 1) weak monotonicity of the untreated outcomes in the fraction treated $p$, and 2) weak monotonicity of the treated outcomes in the fraction treated $p$. As I show in Kowalski (2016) and demonstrate here, the test only requires one of their ancillary assumptions; either one is sufficient. I also show that each ancillary assumption implies an upper or lower bound on the average treatment effect for always or never takers. These bounds help me to demonstrate how the ancillary assumptions combine with information on always and never takers to test external validity. Intuitively, if the bounds on the average treatment effects of always and never takers do not include the LATE, then the LATE cannot be externally valid for all policies.

Bounds on treatment effects for always and never takers are also interesting in their own right. When the test shows that the LATE is not externally valid, the bounds demonstrate the magnitude and direction of variation in the average treatment effect across always takers, compliers, and never takers. The average treatment effects on always and never takers can

be particularly policy-relevant. Suppose that a policy assigns treatment using a lottery, but always and never takers are possible. If a hypothetical future policy were to prohibit treatment for everyone, then its effect would depend on the average treatment effect on always takers. On the other end of the spectrum, if a hypothetical future policy were to mandate treatment for everyone, then its effect would depend on the average treatment effect on never takers.

In Figure 7, I depict the bounds that result from applying the ancillary assumptions to the hypothetical example. The LATE assumptions imply an ordering from always takers to compliers to never takers along the horizontal axis. The ancillary assumptions imply the same ordering along the vertical axis. Because the average untreated outcome of compliers is larger than the average untreated outcome of never takers, yielding a positive untreated outcome test statistic in Figure 6, the ancillary assumption on the untreated outcomes implies a lower bound on the average untreated outcome of always takers in Figure 7. A negative untreated outcome test statistic would imply an upper bound.

Figure 7: Test for External Validity Rejects Under Ancillary Assumptions: Treatment Effect Bounds



The average treatment effect for a group is the difference between the average treated and untreated outcomes for that group. As depicted in Figure 7, for always takers, the difference between the observed average treated outcome and the lower bound on the average untreated outcome implies an upper bound on the average treatment effect. As shown, the upper bound on the average treatment effect for always takers is 200, which is *less* than the LATE of 320. Therefore, the test rejects the external validity of the LATE under the single assumption that untreated outcomes are weakly monotonic in the fraction treated $p$.

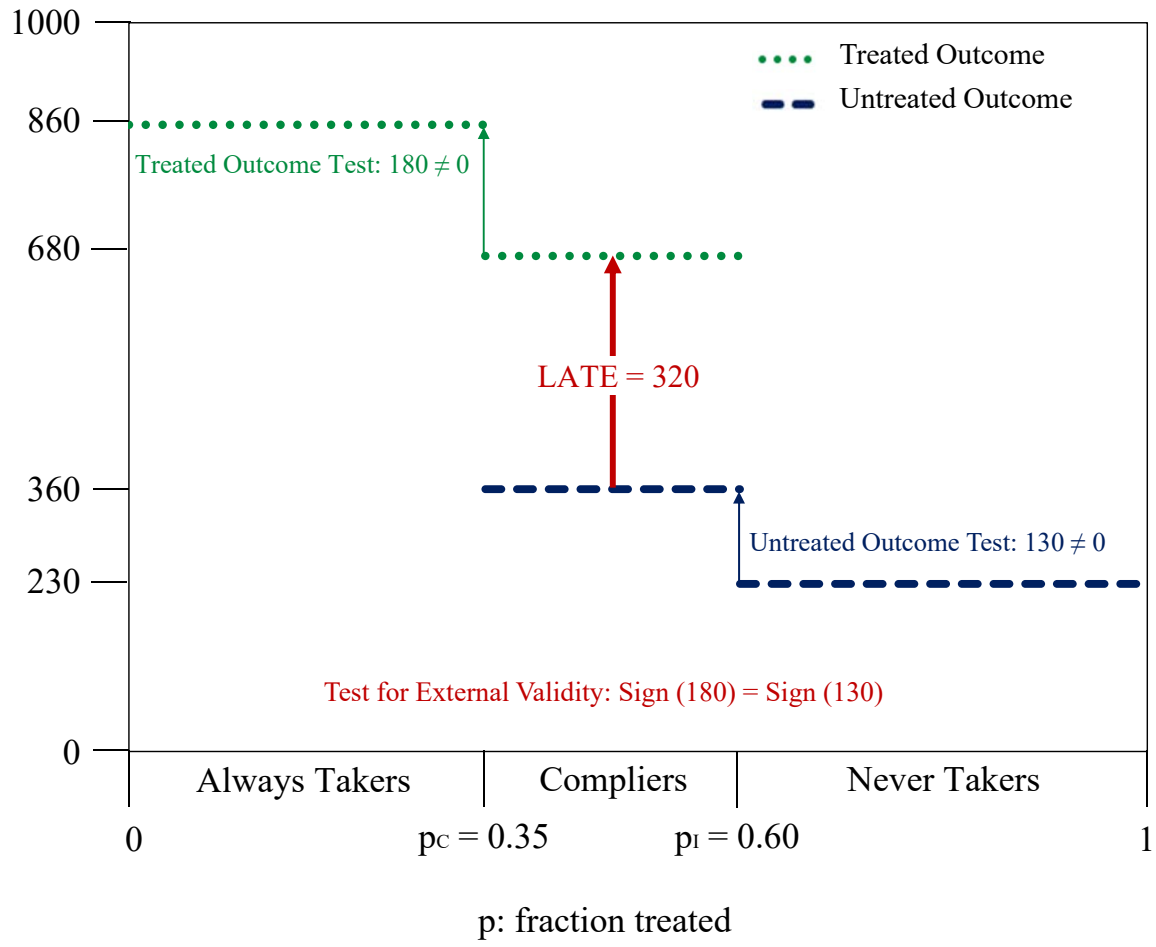
In Figure 7, I also depict the implications of the alternative ancillary assumption that *treated* outcomes are weakly monotonic in the fraction treated $p$. This assumption implies an upper or lower bound on the average treated outcome for never takers, depending on the sign of the treated outcome test statistic. The treated outcome test statistic is negative in Figure 6, so the assumption implies a lower bound on the average untreated outcome for never takers in Figure 7. As shown, the lower bound on the average treated outcome for never takers of 680 implies that the average treatment effect for never takers must be greater than or equal to 450. However, the LATE is equal to 320, so the test also rejects the external validity of the LATE under the alternative ancillary assumption.

In experiments with two-sided noncompliance, the test for external validity always yields the same result under either ancillary assumption. To demonstrate, Figure 8 depicts a different hypothetical example in which the test for external validity does not reject under either ancillary assumption. The only change in the hypothetical data is the average treated outcome of always takers, which changes from 560 in Figure 7 to 860 in Figure 8. This simple change reverses the sign of the treated outcome test statistic. Under this simple change, neither ancillary assumption rules out the external validity of the LATE, as demonstrated by the bounds depicted in Figure 9.

# 5 Tests For External Validity and Estimates of Treatment Effect Heterogeneity under Stronger Ancillary Assumptions

In cases where the test of external validity does not reject under the ancillary assumptions of weak monotonicity of the treated or untreated outcomes in the fraction treated $p$, researchers can impose stronger assumptions to generate more powerful tests. In the process, these stronger assumptions can be used to obtain estimates of average treatment effects on always and never takers in lieu of bounds. Although it is natural to progress from weaker assumptions to stronger assumptions in empirical work, the stronger assumptions were pro-

Figure 8: Test for External Validity Does Not Reject Under Ancillary Assumptions

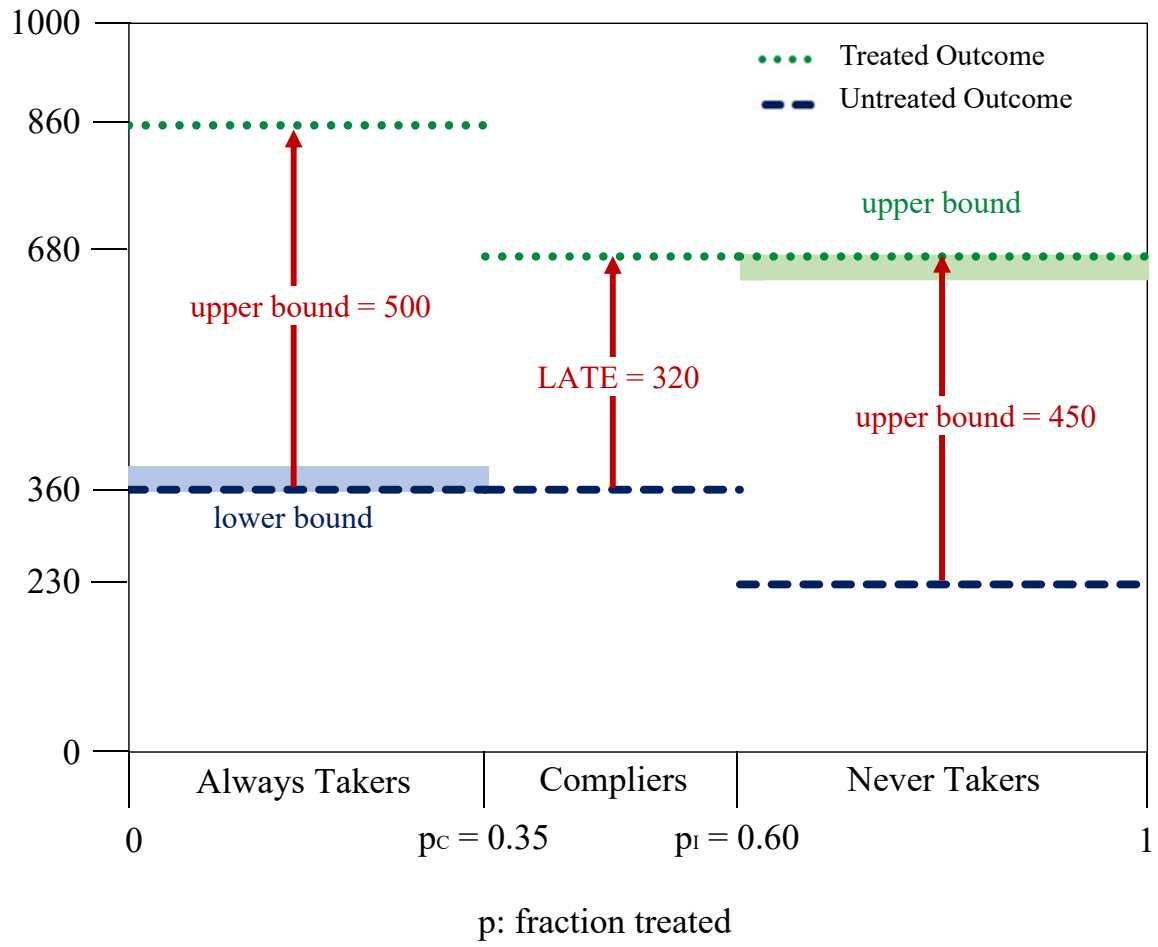



posed first.

One such set of stronger assumptions is linearity of the treated and untreated outcomes in the fraction treated $p$. Olsen (1980) imposes linearity of the treated outcomes only. Brinch et al. (2017) impose both ancillary linearity assumptions simultaneously. They show that under both ancillary linearity assumptions, the test of the null hypothesis that the untreated outcome test statistic is equal to the treated outcome test statistic is a test for external validity. Hausman (1978); Angrist (2004); Huber (2013); Bertanha and Imbens (2014); Guo et al. (2014) and Black et al. (2015) propose tests that are tests of external validity if both ancillary linearity assumptions hold, but they do not all state these assumptions.

Figure 10 demonstrates the implications of the ancillary linearity assumptions using the same hypothetical data as Figures 8 and 9. As in Kowalski (2016, 2018), I refer to the function that specifies how treated outcomes vary with the fraction treated $p$ as the marginal treated outcome function MTO(p), and I refer to the corresponding function for untreated outcomes as the marginal untreated outcome function MUO(p). Linearity of the treated and

Figure 9: Test for External Validity Does Not Reject Under Ancillary Assumptions: Treatment Effect Bounds

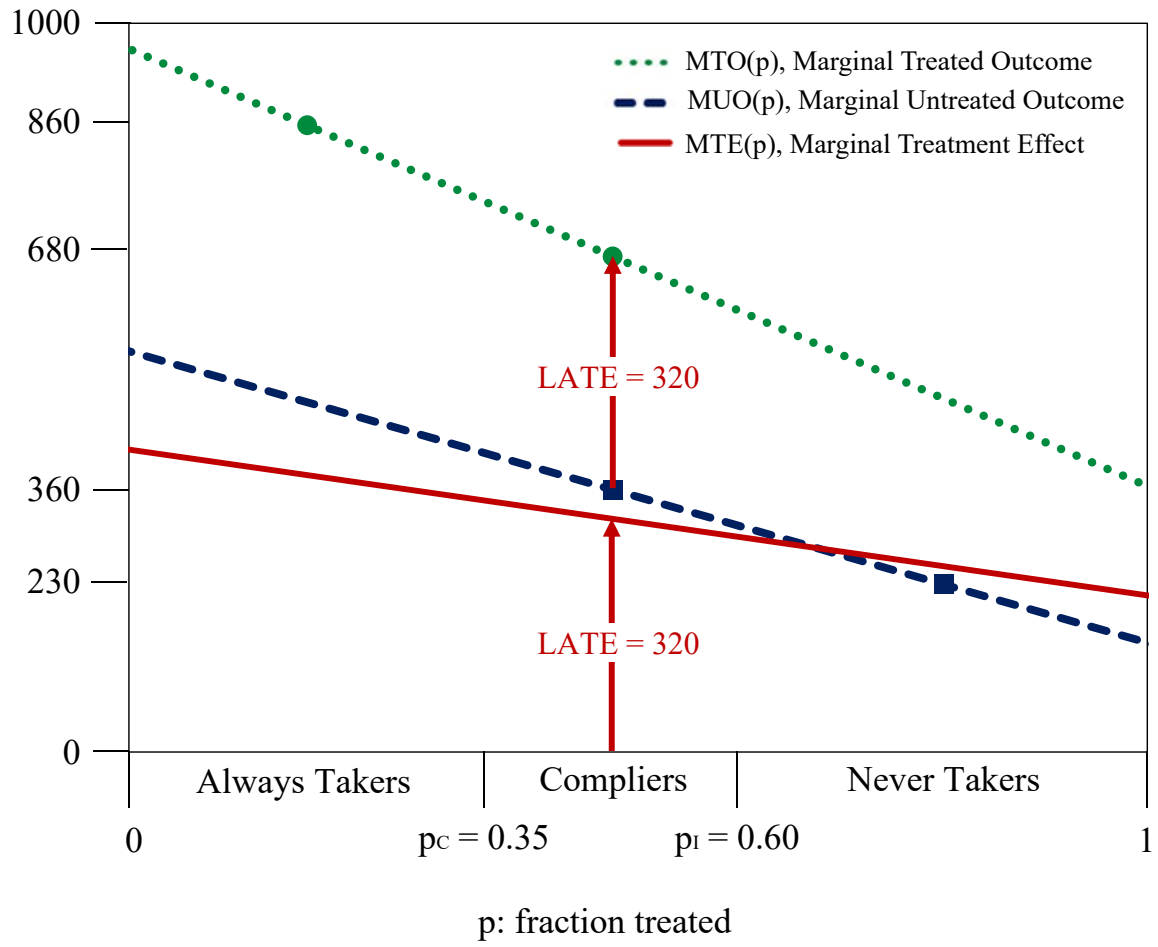



untreated outcomes in the fraction treated $p$ implies that the MTO and MUO functions are linear, as depicted in Figure 10. The difference between the MTO and MUO functions yields the marginal treatment effect function MTE(p) from the literature. The MTE function is linear in Figure 10 because the MTO and MUO functions are linear.

If the MTE function has a nonzero slope, then the treatment effect varies with the fraction treated $p$, and the LATE cannot be externally valid for all policies. Thus, a test for external validity under the ancillary linearity assumptions tests whether the slope of the MTE function is zero. In the hypothetical example depicted in Figure 10, the test for external validity rejects under the ancillary linearity assumptions. In contrast, the test for external validity does not reject under the ancillary weak monotonicity assumptions, as depicted in Figures 8 and 9. The comparison of the results under both sets of assumptions demonstrates that the stronger ancillary assumptions are more powerful, as discussed in Brinch et al. (2017).

As depicted in Figure 10, the ancillary linearity assumptions preserve the LATE of 320

Figure 10: Test for External Validity Rejects Under Stronger Ancillary Assumptions: Treatment Effect Estimates




while also yielding an estimate of the treatment effect at every fraction treated $p$, as depicted by the marginal treatment effect function MTE(p). The marginal treatment effect function $MTE(p)$ can be weighted to recover many average treatment effects of interest following Heckman and Vytlacil (1999, 2001b, 2005), Carneiro et al. (2011), Brinch et al. (2017) and Kowalski (2016, 2018). These average treatment effects of interest include the average treatment effects for always and never takers.

Any alternative ancillary assumptions that identify the MTE function at every fraction treated $p$ also allow for tests of external validity and estimates of average treatment effects for always and never takers. For example, Kline and Walters (2018) 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, then it is also possible to incorporate shape restrictions on how covariates enter the MTO and MUO functions to allow for more flexible functional forms for the MTE function, as in Carneiro and Lee (2009); Carneiro et al. (2011); Maestas et al. (2013); Kline and Walters (2016); Brinch et al. (2017); Kowalski (2016, 2018)

Researchers can determine which ancillary assumptions they are willing to impose based on the institutional features of their experiments. For example, in some experiments, it could be plausible that participants select into treatment based on underlying differences in their untreated outcomes, motivating assumptions on the untreated outcomes. In other experiments, it could be plausible that participants select into treatment based on their anticipated treated outcomes, motivating assumptions on the treated outcomes. The set of plausible assumptions could vary within an experiment across outcomes.

Researchers can also determine which ancillary assumptions they are willing to impose through examination of available covariates. For example, monotonicity in baseline covariates across always takers, compliers, and never takers can lend support to the assumption of monotonicity in untreated outcomes. The assumption of monotonicity of treated outcomes is harder to defend based on baseline covariates, unless there is an institutional reason to believe that the covariates affect selection *and* treatment effect heterogeneity. Researchers can also use shape restrictions on how covariates enter the MTO and MUO functions to test alternative assumptions.

# 6 Implications for Experimental Design

The examination of external validity in this essay reinforces a counter-intuitive insight: researchers should consider designing experiments to allow for always and never takers if the policy of interest would also entail always and never takers. Heckman and Vytlacil (2001a, 2007) make this insight clear with the concept of "policy-relevant treatment effects." If researchers are interested in treatment effects from a policy that would allow for always and never takers, then they should consider designing experiments with interventions to yield the same always or never takers that they would expect under the policy.

Sometimes researchers force all individuals to comply with random assignment with the goal of estimating a LATE equal to the average treatment effect in the entire sample. 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. In fact, by forcing perfect compliance when the policy of interest would not force all individuals to receive treatment, researchers not only risk reducing the applicability of their

results, but also they eliminate useful information. Such information can be used to examine heterogeneous selection under the given policy, and it can be combined with assumptions to examine the heterogeneous selection and treatment effects that would be induced by a range of hypothetical policies.

If researchers are primarily interested in the impact of a range of hypothetical policies, then they should consider designing experiments with a range of interventions instead of a single intervention. For example, researchers can offer a range of randomized prices for a treatment instead of simply offering the treatment for free to lottery winners. Ashraf et al. (2010), Chassang et al. (2012), Berry et al. (2015), and Narita (2018) present experimental designs that involve a range of interventions. These designs potentially involve a loss of power, but they have important advantages. Experiments with a range of interventions can inform selection and treatment effect heterogeneity even if always and never takers are not possible. Furthermore, if the range of interventions induces a continuous fraction treated over some range, then researchers can identify the selection and treatment effect heterogeneity over that range without any ancillary assumptions.

Finally, researchers should collect data to facilitate examination of external validity within and across experiments. To apply the approaches discussed in this essay, it is imperative to collect data such that it is possible to perform tabulations of outcomes by lottery status *and* treatment. It is also useful to collect data such that it is possible to perform similar tabulations of covariates. Data on covariates also facilitate comparisons across experiments. Approaches to assess external validity across experiments are even more powerful when used in concert with approaches to assess external validity within experiments.

# References


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

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

Joshua D Angrist. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336, 1990.

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

Joshua D Angrist and Alan B Krueger. 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, 1992.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Nava Ashraf, James Berry, and Jesse M Shapiro. Can higher prices stimulate product use? evidence from a field experiment in zambia. *The American economic review*, 100(5):2383–2413, 2010.

James Berry, Greg Fischer, and Raymond P Guiteras. Eliciting and utilizing willingness to pay: evidence from field trials in northern ghana. 2015.

Marinho Bertanha and Guido W. Imbens. External validity in fuzzy regression discontinuity designs. Working Paper 20773, National Bureau of Economic Research, December 2014.

Anders Björklund and Robert Moffitt. The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, pages 42–49, 1987.

Dan A Black, Joonhwi Joo, Robert LaLonde, Jeffrey A Smith, and Evan J Taylor. Simple tests for selection bias: Learning more from instrumental variables. 2015.

Christian N Brinch, Magne Mogstad, and Matthew Wiswall. Beyond late with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039, 2017.

Pedro Carneiro and Sokbae Lee. 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, 2009.

Pedro Carneiro, James J. Heckman, and Edward J. Vytlacil. Estimating marginal returns to education. *American Economic Review*, 101(6):2754–81, October 2011. doi: 10.1257/aer.101.6.2754. URL `http://www.aeaweb.org/articles/?doi=10.1257/aer.101.6.2754`.

Sylvain Chassang, Gerard Padro I Miquel, and Erik Snowberg. Selective trials: A principal-agent approach to randomized controlled experiments. *American Economic Review*, 102(4):1279–1309, 2012. doi: 10.1257/aer.102.4.1279. URL `http://www.aeaweb.org/articles.php?doi=10.1257/aer.102.4.1279`.

Liran Einav, Amy Finkelstein, and Mark R Cullen. Estimating welfare in insurance markets using variation in prices. *The Quarterly Journal of Economics*, 125(3):877, 2010.

Zijian Guo, Jing Cheng, Scott A Lorch, and Dylan S Small. Using an instrumental variable to test for unmeasured confounding. *Statistics in medicine*, 33(20):3528–3546, 2014.

Jerry A Hausman. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, pages 1251–1271, 1978.

James J Heckman and Edward Vytlacil. Policy-relevant treatment effects. *American Economic Review*, 91(2):107–111, 2001a.

James J. Heckman and Edward Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738, 05 2005. URL `http://ideas.repec.org/a/ecm/emetrp/v73y2005i3p669-738.html`.

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

James J. Heckman and Edward J. Vytlacil. Local instrumental variables. In Cheng Hsiao, Kimio Morimune, and James L. Powell, editors, *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pages 1–46. Cambridge University Press, 2001b.

James J Heckman and Edward J Vytlacil. 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, 2007.

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

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

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

Guido W. Imbens and Donald B. Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997. ISSN 00346527, 1467937X. URL `http://www.jstor.org/stable/2971731`.

Lawrence F Katz, Jeffrey R Kling, Jeffrey B Liebman, et al. Moving to opportunity in boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics*, 116(2):607–654, 2001.

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

Patrick M. Kline and Christopher R. Walters. On heckits, late, and numerical equivalence. Working Paper 24477, National Bureau of Economic Research, April 2018. URL `http://www.nber.org/papers/w24477`.

Amanda Kowalski. 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, June 2016. URL `http://www.nber.org/papers/w22362`.

Amanda Kowalski, Yen Tran, and Ljubica Ristovska. MTEBINARY: Stata module to compute Marginal Treatment Effects (MTE) With a Binary Instrument. Statistical Software Components, Boston College Department of Economics, December 2016. URL `https://ideas.repec.org/c/boc/bocode/s458285.html`.

Amanda E. Kowalski. Extrapolation using selection and moral hazard heterogeneity from within the oregon health insurance experiment. Working Paper 24647, National Bureau of Economic Research, May 2018. URL `http://www.nber.org/papers/w24647`.

Nicole Maestas, Kathleen J Mullen, and Alexander Strand. Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *The American Economic Review*, 103(5):1797–1829, 2013.

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

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment effects. Working Paper 23568, National Bureau of Economic Research, July 2017.

Thomas A Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society*, pages 765–799, 1987.

Yusuke Narita. Toward an ethical experiment. Working Paper 2127, Cowles Foundation, 2018.

Randall J. Olsen. A least squares correction for selectivity bias. *Econometrica*, 48(7):1815–1820, 1980. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1911938.

Andrew Donald Roy. Some thoughts on the distribution of earnings. *Oxford economic papers*, 3(2):135–146, 1951.

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

Abraham Wald. The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, 11(3):284–300, 09 1940.

Robert J Willis and Sherwin Rosen. Education and self-selection. *Journal of political Economy*, 87(5, Part 2):S7–S36, 1979.