How to Examine External Validity Within an Experiment*

Amanda E. Kowalski
Associate Professor, Department of Economics, Yale University
Faculty Research Fellow, NBER

August 31, 2016

Prepared for the *Journal of Economic Perspectives.*

Abstract

A fundamental concern for researchers who design and analyze experiments is that the experimental result might not be externally valid in another context. Researchers have traditionally attempted to assess external validity by comparing data from an experiment to other data. In this essay, I use insights from my recent work to show researchers how to begin the examination of external validity internally, within the data from a single experiment. My insights rely on overlooked information and minimal assumptions.

*I thank Ljubica Ristovska, Rae Staben, and Matthew Tauzer for excellent research assistance. NSF CAREER Award 1350132 provided support.

1 Introduction

The traditional reason that a researcher runs an experiment is to eliminate 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, the experiment isolates a single treatment effect by eliminating selection.

However, there is still room for selection within an experiment. In many experiments, lottery losers can decide to receive treatment, and lottery winners can decide to forgo treatment. Some researchers view this type of selection as immaterial, and they discard information on selection 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 reduce information on selection into treatment by encouraging all individuals to comply with random assignment. I view this type of selection as a useful source of overlooked information, and I combine information on selection into treatment with minimal assumptions to learn more from an experiment.

The ability to learn from overlooked information on selection into treatment gives a researcher new reasons to run an experiment. An experiment is no longer just a tool that eliminates selection; it is also a tool that identifies selection. Furthermore, an experiment is no longer just a tool that isolates a single treatment effect; it is also a tool that identifies treatment effect heterogeneity.

A fundamental concern for a researcher who examines an experiment has been whether the single treatment effect that the experiment isolates will be externally valid in other contexts. However, an experiment re-conceived as a tool that identifies treatment effect heterogeneity can itself inform external validity. If the treatment effect varies across individuals within an experiment, then there is no single treatment effect that is externally valid in all contexts.

In this essay, I show researchers how to examine external validity within an experiment. I focus heavily on key insights from my own recent work in Kowalski [2016]. After all, my thesis is that the analysis of external validity should begin internally!

In the next section, I use a graphical approach to highlight how traditional analysis of an experiment overlooks information on selection into treatment. In Section 3, I use that information to construct the marginal untreated outcome test for selection introduced in Kowalski [2016]. In Section 4, I impose minimal assumptions to construct a test of external validity introduced in Kowalski [2016]. I briefly discuss extensions in Section 5. Finally, discuss implications for how to design experiments to assess selection, treatment

effect heterogeneity, and thus external validity.

2 Overlooked Information in Traditional Analysis of an Experiment

Throughout this essay, I consider a specific type of experiment. In this type of experiment, a sample of individuals participates in a lottery. Individuals who win the lottery receive an intervention that affects selection into treatment; individuals who lose the lottery do not. However, all individuals can decide if they want to select into treatment. Some individuals who lose the lottery receive treatment, and other individuals who win the lottery forgo treatment. The researcher can observe whether each individual won the lottery, whether each individual received the treatment, and an outcome for each individual.

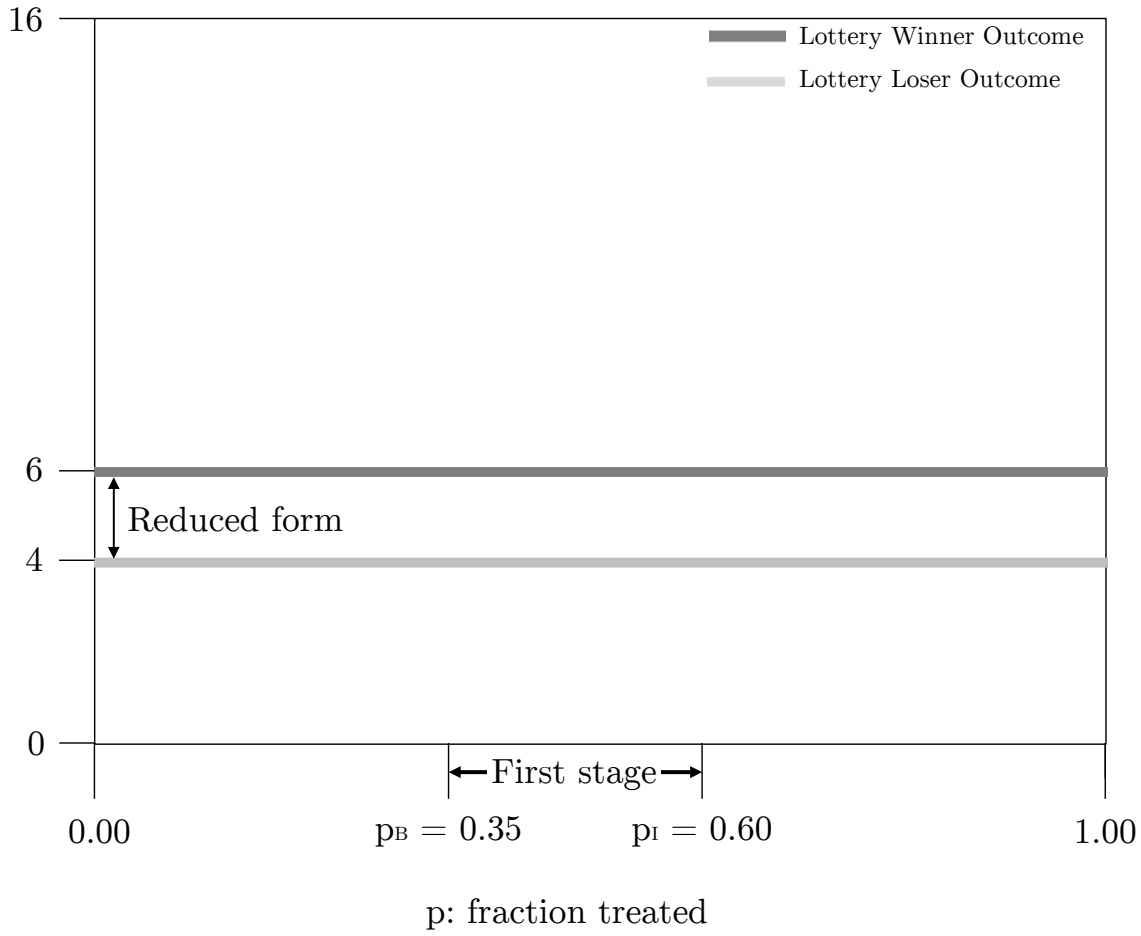
The traditional analysis of an experiment begins by comparing the average outcomes of lottery winners and losers. In Figure 1, I depict results from a hypothetical experiment in which lottery winners have an average outcome of 6 and lottery losers have an average outcome of 4. The difference in their outcomes is 2. This difference is often called the "reduced form," as labeled along the vertical axis of Figure 1. It gives an estimate of the impact of the intervention that lottery winners receive on the outcome. In the type of experiment that I consider, 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. The reduced form does not even use data on treatment. Some researchers report only the reduced form.

The traditional analysis of an experiment next compares the average treatment probability of lottery losers p_B to the average treatment probability of lottery winners p_I.[1] The difference $p_I - p_B$ is often called the "first stage." It gives an estimate of the impact of winning the lottery on p, the fraction of the sample that receives treatment. In the type of experiment that I consider, 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 estimate implies that winning the lottery increases the probability of treatment by 25 percentage points.

To obtain an estimate of the impact of the treatment on the outcome, the traditional analysis of an experiment divides the reduced form by the first stage. This quotient is often called the "local average treatment effect" (LATE). Under the assumptions of monotonicity

[1] In this notation, p_B represents the probability of treatment at "baseline," which is observed among lottery losers, and p_I represents the probability of treatment under the "intervention," which is observed among lottery winners.

Figure 1: Traditional Analysis of an Experiment



and independence set forth by Imbens and Angrist [1994], the LATE gives the average treatment effect on "compliers," individuals whose treatment status is determined by their random assignment. The 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 8 (=2/0.25).

Selection into treatment generates two groups of individuals that the LATE overlooks: "always takers" who take up treatment regardless of random assignment, and "never takers" who do not take up treatment regardless of random assignment. The LATE monotonicity assumption implies that all participants in the lottery are either always takers, compliers, or never takers; there are no "defiers" who select treatment if and only if they lose the lottery. In the type of experiment that I consider, researchers cannot identify whether any individual

is a complier: lottery winners who take up treatment could be compliers or always takers; lottery losers who do not take up treatment could be compliers or never takers. However, researchers can identify some individuals as always takers or never takers. Lottery losers who take up treatment are always takers; lottery winners who do not take up treatment are never takers. These individuals provide useful information traditionally discarded by researchers.

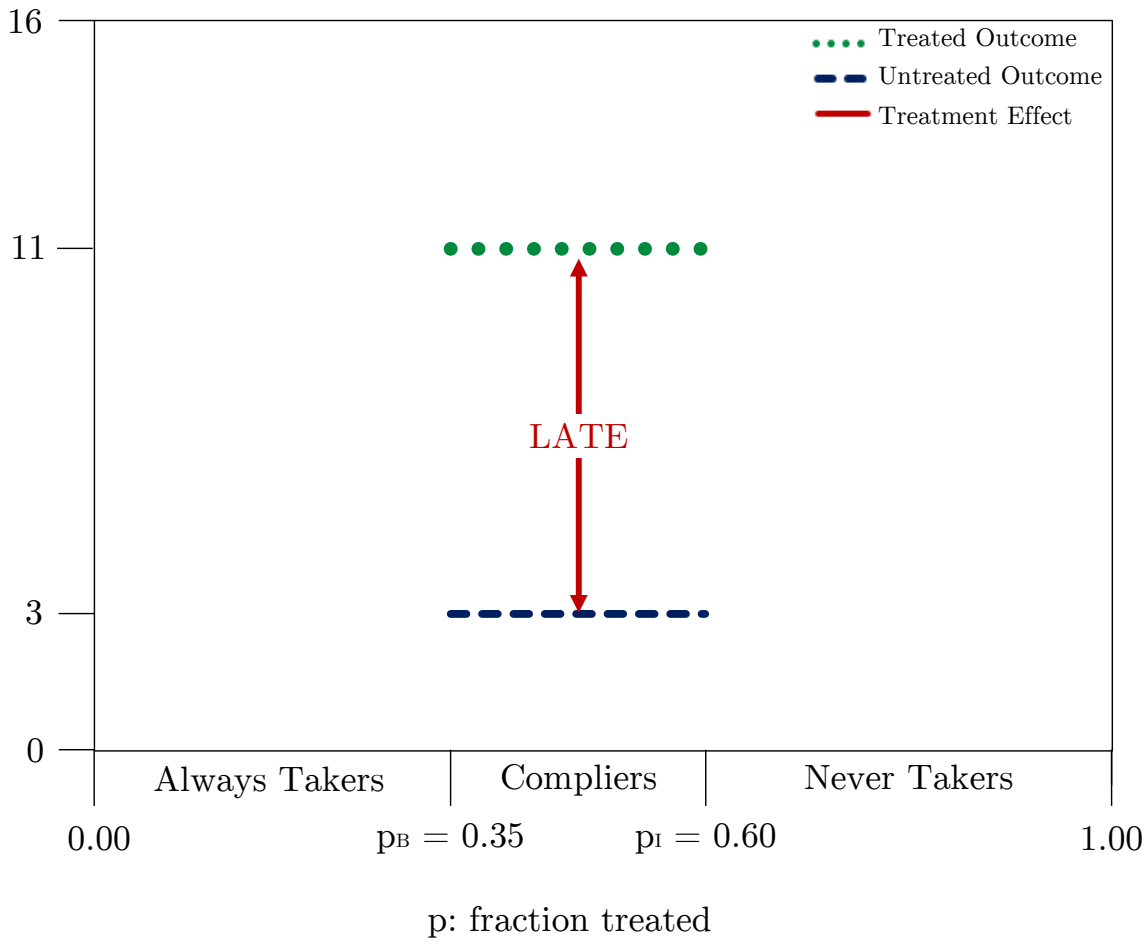
The ability to identify some always and never takers allows researchers to learn more about compliers. The LATE independence assumption guarantees that lottery status is independent of whether an individual is an always taker, complier, or never taker. Therefore, the observed share of treated lottery 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 takers and never takers do not change their treatment decisions based on the lottery, their average outcomes should not depend on their lottery status. Using the shares and average outcomes of always takers and never takers, researchers can recover the average outcomes of treated and untreated compliers, as demonstrated by Katz et al. [2001] and Abadie [2003]. The LATE is equal to the difference in the average outcomes of treated and untreated compliers.

Using the same hypothetical data presented in Figure 1, in Figure 2, I depict the LATE as the difference between the average treated and untreated outcomes of compliers along the vertical axis. The LATE assumptions imply an ordering along the horizontal axis that I introduce in Kowalski (2016). By the LATE independence assumption, lottery status is independent of treatment, so the horizontal axis p can depict the fraction treated among either lottery winners or losers. By the LATE monotonicity assumption, the horizontal axis can be separated into three groups. First, the individuals who receive treatment when the fraction treated is in the range $0 < p \leq p_B$ must be always takers because the p_B of individuals who receive treatment when they lose the lottery are always takers. Second, the individuals who receive treatment when the fraction treated is in the range $p_B < p \leq p_I$ must be compliers because the $(p_I - p_B)$ individuals who gain treatment if they win the lottery but do not gain treatment if they lose the lottery are compliers. Third, the remaining individuals who would receive treatment when $p_I < p \leq 1$ must be never takers.

In Figure 2, I plot the average treated and untreated outcomes of compliers only in the range of the horizontal axis relevant to the compliers, making clear that average outcomes of always takers or never takers are not required to identify the LATE.[2] In the example

[2]To obtain the LATE, it is not even necessary to have data on the treatment and the outcome in the same dataset. As long as one data set contains information on the lottery and the treatment and a separate dataset contains information on the lottery and the outcome, then it is possible to construct the exact LATE estimate that would be obtained in a single dataset via "two sample IV." In contrast, the construction of average outcomes for always and never takers requires data on the lottery, the treatment, and the outcome in a single dataset.

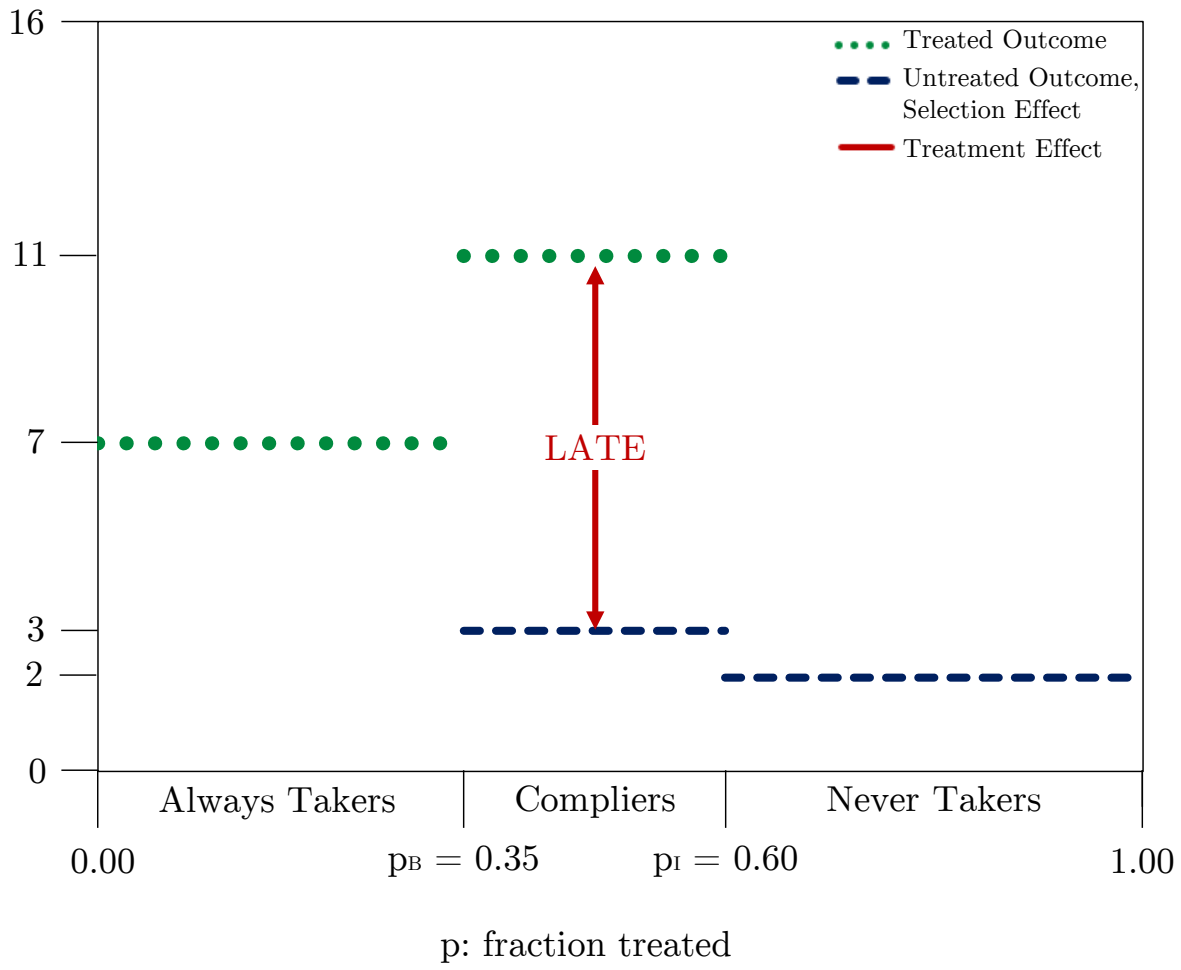
Figure 2: Traditional Analysis of an Experiment: Alternative Depiction



depicted, treated compliers have an outcome of 11 and untreated compliers have an outcome of 3, so traditional analysis of an experiment implies that treatment increases the outcome by 8.

Using the same hypothetical data, in Figure 3, I incorporate information from always takers and never takers, plotting their average outcomes on the vertical axis in the relevant ranges of the horizontal axis. Always takers have an average outcome of 7, and never takers have an average outcome of 2. The information on always takers and never takers depicted along the vertical and horizontal axes, in combination with the information on compliers, can inform selection and treatment effect heterogeneity.

Figure 3: Analysis of an Experiment with Overlooked Information:
Marginal Untreated Outcome Test Shows Positive Selection



3 Marginal Untreated Outcome Test for Selection

The marginal untreated outcome test that I introduce in Kowalski [2016] tests for selection into treatment by comparing the average outcome of untreated compliers to the average outcome of never takers. Untreated compliers and never takers do not receive the treatment. Therefore, a difference in their outcomes cannot reflect a differential treatment effect; it can only reflect a differential selection effect. By the same logic, an untreated outcome reflects a selection effect, as I show in the legend of Figure 3.

In the hypothetical example in Figure 3, along the vertical axis, untreated compliers have a higher average outcome than never takers. Compliers always select into treatment

before never takers, as shown along the horizontal axis. Therefore, in Figure 3, individuals with higher average outcomes select into treatment before individuals with lower average outcomes, and the marginal untreated outcome test provides evidence of positive selection.

Empirically, untreated compliers could have a lower average outcome than never takers. In that case, the marginal untreated outcome test would provide evidence of negative selection. In general, the marginal untreated outcome test can show positive or negative selection without violating the LATE assumptions. Vytlacil [2002] shows that the LATE assumptions are equivalent to the assumptions of the Heckman [1976] model in which individuals select into treatment based on their net benefit of treatment. Therefore, the marginal untreated outcome test should show positive selection if the outcome measures the net benefit of treatment. However, if the outcome measures anything else, such as a component of the net benefit of treatment, then the marginal untreated outcome test can show positive or negative selection without violating the LATE assumptions. For example, if the outcome measures insurer costs, then the marginal untreated outcome test can detect adverse or advantageous selection into insurance, generalizing the "cost curve" test of Einav et al. [2010]. Furthermore, within the same experiment, the marginal untreated outcome test can show positive selection on some outcomes while showing negative selection on others.

The analogous marginal treated outcome test that compares the average outcome of always takers to the average outcome of treated compliers does not isolate selection because it can also reflect a heterogeneous treatment effect. Always takers always select into treatment before compliers, as shown along the horizontal axis of Figure 3. In the hypothetical example depicted, always takers have a lower average outcome than treated compliers, as shown along the vertical axis. Therefore, in Figure 3, always takers could be negatively selected relative to compliers, or the treatment could have a smaller average effect on always takers than it has on compliers, or there could be some interplay between selection and treatment effect heterogeneity. The key to isolating treatment effect heterogeneity is to first isolate selection using the marginal untreated outcome test and then purge that selection from the marginal treated outcome test so that only treatment effect heterogeneity remains.

4 Test of External Validity

If the treatment does not have the same average effect on always takers that it has on compliers, then the LATE, which is the average treatment effect on compliers, cannot be a "globally externally valid" treatment effect that applies in all contexts.[3] Similarly, the LATE

[3]The LATE can still be "locally externally valid" for another context of interest if the LATE is equal to the treatment effect in the context of interest, even if the treatment effect varies.

cannot be a globally externally valid treatment effect if it does not apply to never takers. In Kowalski [2016], I use these insights, overlooked information on always takers and never takers, and assumptions to test for global external validity within an experiment.

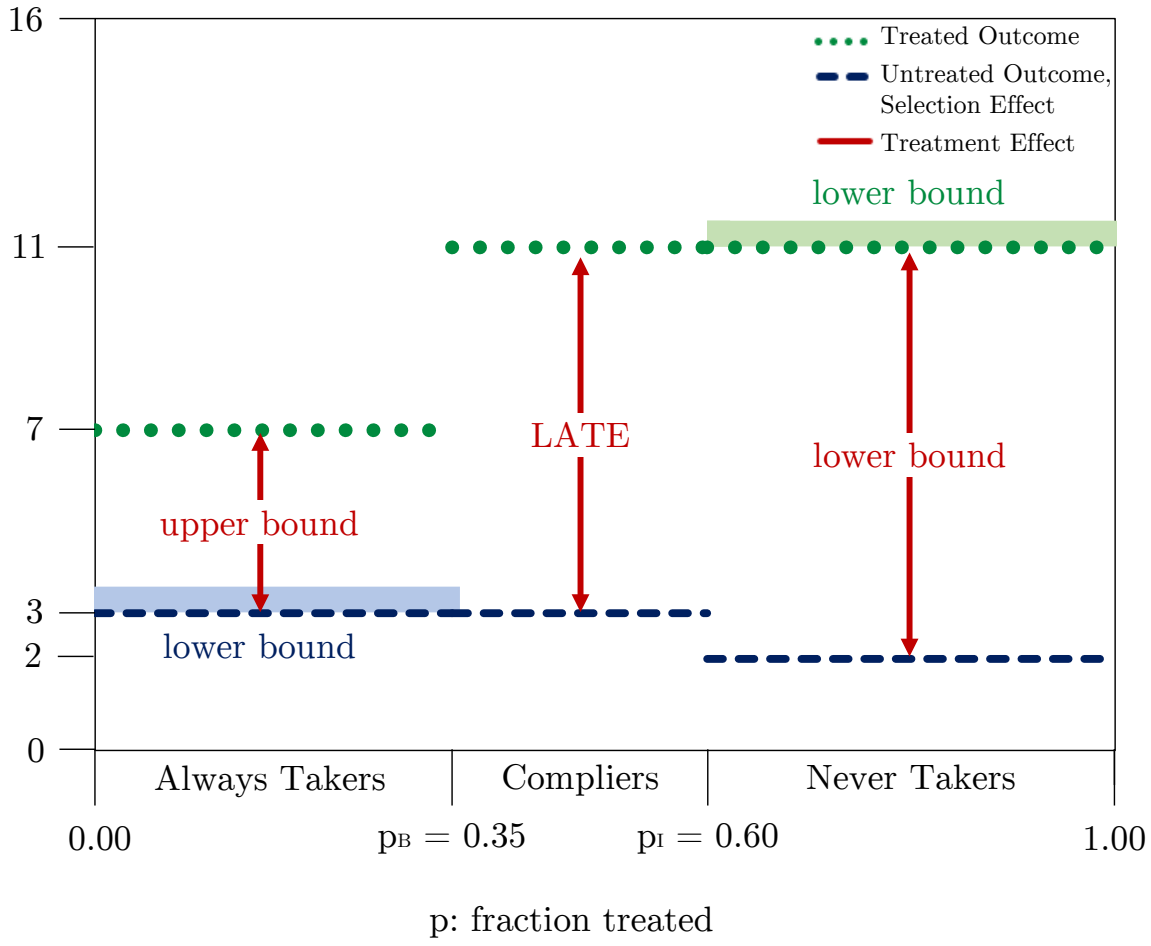
A single assumption that allows for a test of global external validity is that selection into treatment is weakly monotonic from always takers to compliers to never takers. This assumption implies an upper or lower bound on the untreated outcome of always takers, which is not observed because always takers always receive treatment in the experiment. Whether the marginal untreated outcome tests shows positive or negative selection determines whether the assumption implies an upper or lower bound. As depicted in Figure 4, which reflects the same hypothetical data as the previous figures, the marginal untreated outcome test shows positive selection, so the assumption of weak monotonicity in the untreated outcomes implies a lower bound on the average untreated outcome of always takers. If the marginal untreated outcome test were to show negative selection, then the untreated outcome of compliers would form an upper bound on the average untreated outcome of always takers.

The bound on the average untreated outcome of always takers implies a bound on the average treatment effect for always takers. The average treatment effect for a group is the difference between the average treated and untreated outcomes for that group. As depicted in Figure 4, the difference between the observed average treated outcome and the bound on the average untreated outcome implies a maximum average treatment effect for always takers. As shown, the maximum average treatment effect for always takers of 4 is less than the average treatment effect for compliers of 8. Therefore, the test rejects the global external validity of the LATE.

An alternative assumption that allows for a test of global external validity is that treated outcomes are weakly monotonic from always takers to compliers to never takers. This assumption yields an upper or lower bound on the treated outcome of never takers, which is not observed because never takers never receive treatment in the experiment. Under this assumption, in Figure 4, the average treated outcome of always takers is smaller than the average treated outcome of compliers, so the average treated outcome of compliers forms a lower bound on the average treated outcome of never takers. As shown, the lower bound on the average treated outcome of never takers of 11 implies that the average treatment effect for never takers must be greater than or equal to 9. However, the LATE is equal to 8, so the test also rejects the global external validity of the LATE under the alternative assumption.

In the type of experiment that I consider, the test of global external validity always yields the same result under both assumptions. Figure 5 depicts a different hypothetical example in which the test does not reject global external validity. The only change in hypothetical data from Figure 4 to Figure 5 is the average treated outcome of always takers, which changes

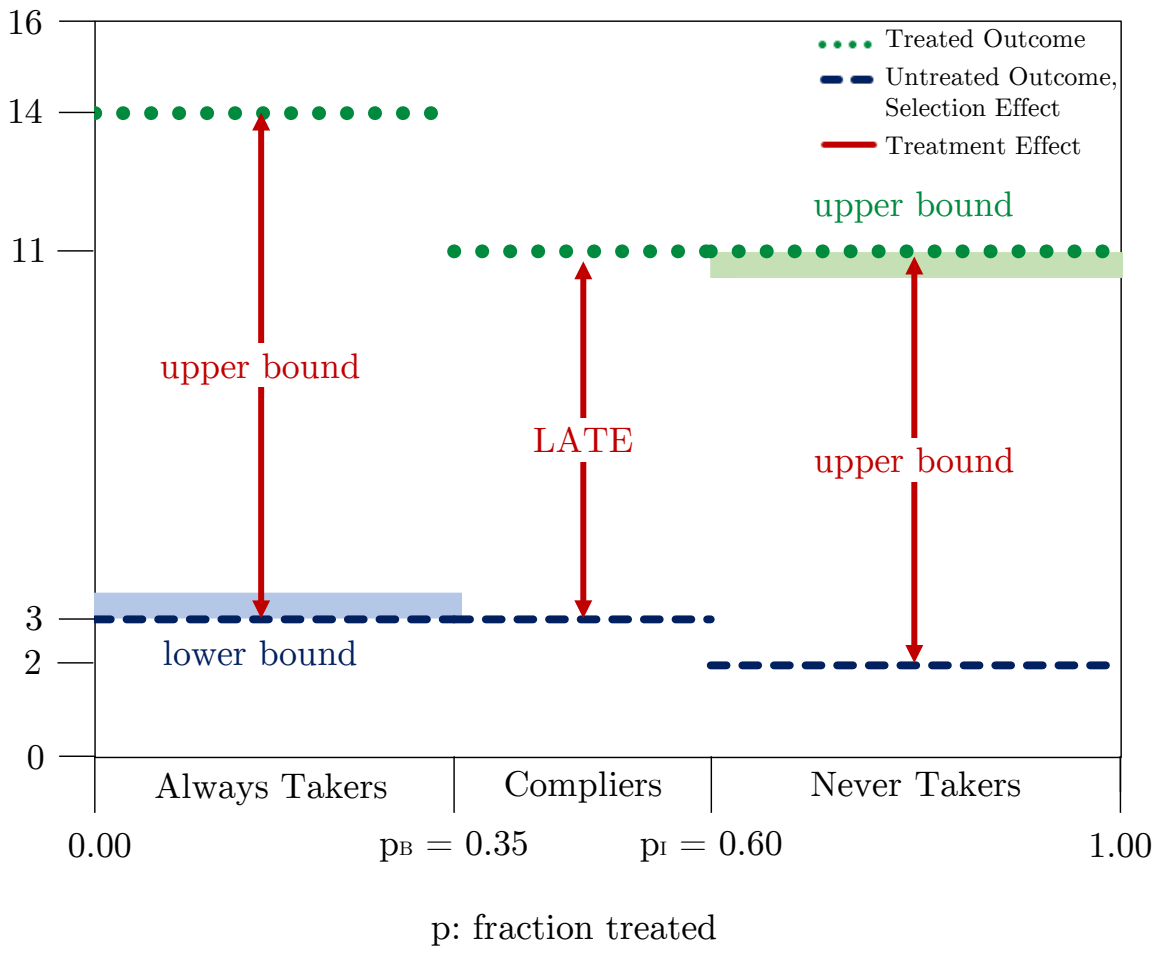
Figure 4: Analysis of an Experiment with Overlooked Information and Weak Monotonicity Assumptions: Test Rejects External Validity



from 7 in Figure 4 to 14 in Figure 5. This simple change reverses the monotonic relationship between the average treated outcomes of always takers and compliers. Weak monotonicity of the untreated outcomes rules out the global external validity of the LATE in Figure 4 but does not in Figure 5. Similarly, weak monotonicity of the treated outcomes rules out the global external validity of the LATE in Figure 4 but does not in Figure 5

The intuition behind why both assumptions yield the same result is that the LATE can only be globally externally valid if the treatment effect is constant. If the treatment effect is constant, then the marginal treated outcome test and the marginal untreated outcome test reflect only selection. If the one test implies positive selection and the other test implies negative selection, then the treatment effect cannot be constant, and the LATE cannot be globally externally valid. In cases where the test does not reject external validity, the

Figure 5: Analysis of an Experiment with Overlooked Information and Weak Monotonicity Assumptions: Test Does Not Reject External Validity



weak monotonicity assumptions still imply bounds that can be informative for researchers interested in predicting behavior.

Researchers can determine if they are willing to impose each weak monotonicity assumption based on the institutional features of their experiments. In some experiments, it will be plausible that participants select into treatment based on underlying differences in their untreated outcomes, motivating weak monotonicity in untreated outcomes. In other experiments, it will be plausible that participants select into treatment based on underlying differences in their untreated outcomes along with an assessment of how much their outcomes will change with treatment, motivating weak monotonicity in treated outcomes. The set of plausible assumptions could vary within an experiment across different outcomes.

5 Extensions

Additional data and assumptions can allow researchers to do even more to assess external validity within an experiment. I discuss several extensions in Kowalski [2016]. Here, I highlight the usefulness of data on individual characteristics and stronger assumptions about heterogeneity in the treated and untreated outcomes.

Thus far, I have not required any data on individual characteristics. However, those data are often available. Suppose that the marginal untreated outcome test shows selection and the test of external validity shows that there must be some treatment effect heterogeneity. With data on individual characteristics, researchers can begin to characterize which types of individuals select into treatment and which individuals have the largest treatment effects.

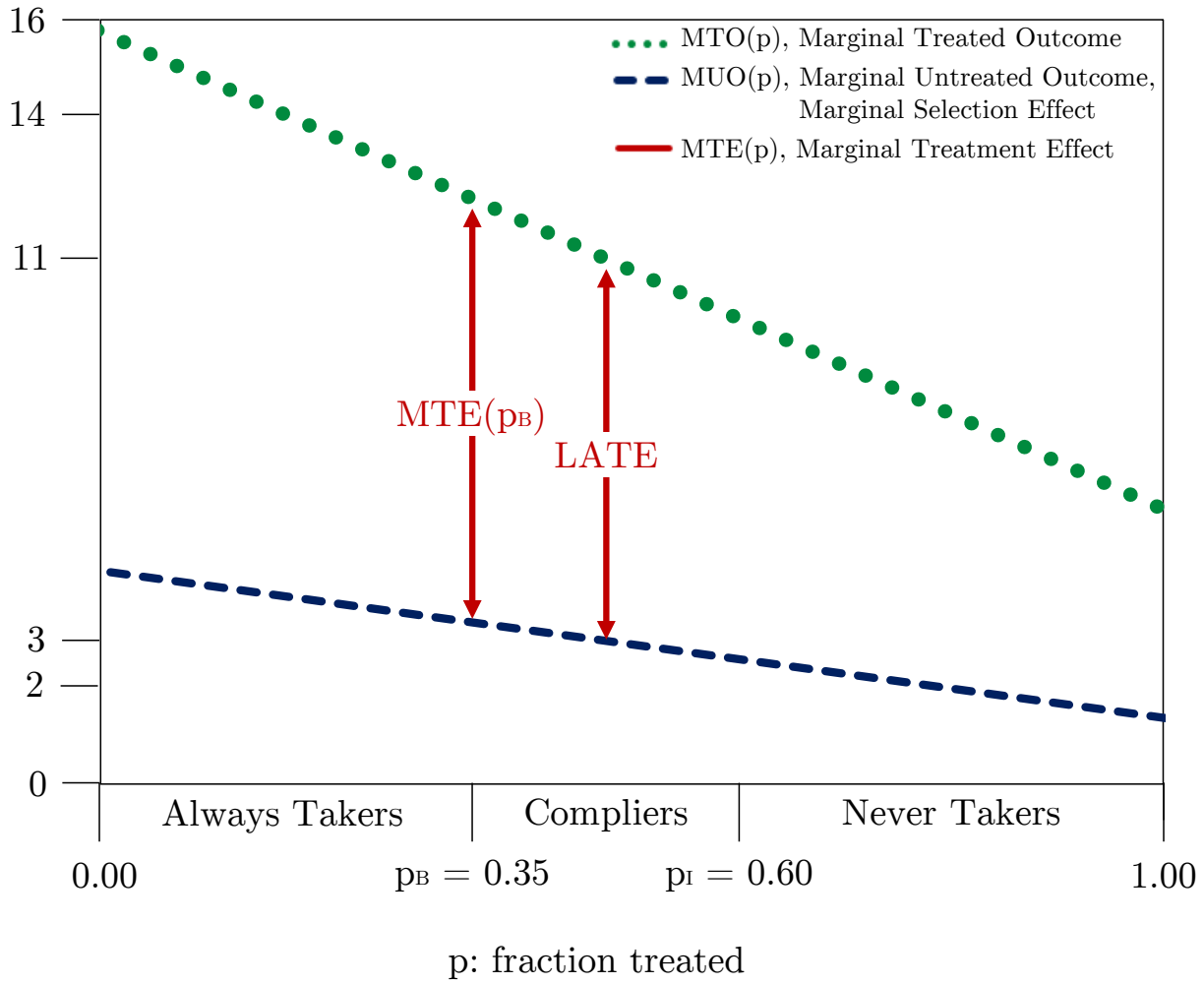
Researchers can derive the average characteristics of always takers, compliers, and never takers using the same approach used to derive their average outcomes. Some researchers already move slightly beyond the traditional analysis of experiments by comparing characteristics of the compliers to characteristics of the full sample. To the extent that the characteristics of the compliers look similar to the characteristics of the full sample, the researchers make stronger claims that their LATEs are externally valid. However, the comparison of the compliers to the full sample potentially obscures differences between the always takers, compliers, and never takers that would be observable in the disaggregated comparison of the three groups.

By comparing the characteristics of never takers and compliers, researchers can investigate which characteristics are related to selection and treatment effect heterogeneity. Suppose that the marginal untreated outcome test shows that never takers are negatively selected relative to the compliers. If the never takers and compliers differ on only one characteristic, researchers can investigate the relationship between the characteristic and selection by restricting the sample to individuals with the characteristic. If the marginal untreated outcome test applied to the restricted sample no longer shows selection, then the researchers have characterized selection. Researchers can use a similar approach to characterize treatment effect heterogeneity. In Kowalski [2016], I provide assumptions that allow for more detailed characterization of selection and treatment effect heterogeneity using data on individual characteristics.

In the absence of data on individual characteristics, researchers can impose assumptions stronger than weak monotonicity to recover estimates of treatment effect heterogeneity in lieu of bounds on treatment effect heterogeneity. Brinch et al. [forthcoming] impose linearity of the treated and untreated outcomes, which results in a linear marginal treated outcome function $MTO(p)$, a linear marginal untreated outcome or marginal selection effect function $MUO(p)$, and a linear marginal treatment effect function $MTE(p)$. Figure 6 depicts the

implications of the linearity assumptions on the same hypothetical data from Figure 5. These assumptions preserve the LATE while also yielding estimates of the treatment effect at every fraction treated p. Figure 6 labels $MTE(p_B)$, an estimate of the treatment effect when p_B of the sample is treated. As depicted, $MTE(p_B)$ is just under 8.9 which is larger than the LATE of 8.

Figure 6: Analysis of an Experiment with Overlooked Information and Linearity Assumptions



The marginal treatment effect function $MTE(p)$ can be used to recover many other treatment effects of interest using techniques developed by Björklund and Moffitt [1987], Heckman and Vytlacil [1999, 2005, 2007], Carneiro et al. [2011]. Until Brinch et al. [forthcoming], researchers did not have approaches to estimate marginal treatment effect functions in settings

with binary instruments. Therefore, researchers did not apply techniques requiring marginal treatment effect functions to the type of experiment that I consider.

For further extensions, I encourage the interested reader to refer to Kowalski (2016), in which I also advance the use of marginal treatment effect functions with experiments. In the process, I demonstrate that the ordinary least squares (OLS) difference in outcomes between all treated individuals and all untreated individuals in an experiment is not comparable to the LATE when the treatment effect is heterogeneous, and I introduce a new estimator that is. I also provide an empirical application to an important experiment for which the data are publicly-available. I can envision a large literature containing further extensions.

6 Implications for Experimental Design

My analysis of experiments yields a counterintuitive insight: researchers should consider designing experiments to allow for always and never takers to generate richer results. An experiment without always and never takers yields a single treatment effect that gives the impact of mandated treatment. If researchers are interested in the impact of a range of interventions that affect treatment, then they should consider running an experiment with always and never takers. Combining information from always takers, compliers, and never takers with assumptions, researchers can estimate the impact of a range of interventions, including mandated treatment. Researchers can also examine selection into treatment across the range of interventions.

Researchers should also consider designing experiments with a range of interventions instead of a single intervention. For example, researchers can offer several different randomized prices for a treatment. Ashraf et al. [2010], Berry et al. [2015], and Chassang et al. [2012] have some exciting recent work in this area. Experiments with a range of interventions instead of a single intervention can inform selection and treatment effect heterogeneity even if always and never takers are not possible. Therefore, if always takers are not possible because the treatment is new and only available to participants who win a lottery, then researchers can still learn about selection and treatment effect heterogeneity if they randomize a range of interventions. If the range of interventions induces a continuous fraction treated, then researchers can use experiments to identify selection and treatment effect heterogeneity nonparametrically, without further assumptions.

Finally, researchers should consider designing experiments so that they generate data that is comparable to other data of interest. Traditionally, researchers have attempted to assess external validity by comparing data from an experiment to other data. Approaches to assess external validity within an experiment are even more powerful when used in concert

with approaches to assess external validity across experiments.

In this essay, I advance the idea that analysis of external validity should begin within an experiment. This idea is powerful because it paves the way for researchers to learn more from existing experiments and to design new experiments that will be even more informative. Insights from my recent work in Kowalski (2016) show researchers how to analyze and design experiments to assess selection, treatment effect heterogeneity, and external validity.

References

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

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.

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.

Christian N Brinch, Magne Mogstad, and Matthew James Wiswall. Beyond late with a discrete instrument. heterogeneity in the quantity-quality interaction of children. *Journal of Political Economy*, forthcoming.

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.

James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER, 1976.

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. 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.

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

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.

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`.

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