University of Calgary

PRISM Repository **https://prism.ucalgary.ca**

The Vault Open Theses and Dissertations

2024-04-03

Addressing the Problem of Extrapolating Results from Randomized Controlled Trials

Zhou, Liyan

Zhou, L. (2024). Addressing the problem of extrapolating results from randomized controlled trials (Master's thesis, University of Calgary, Calgary, Canada). Retrieved from https://prism.ucalgary.ca. https://hdl.handle.net/1880/118374

Downloaded from PRISM Repository, University of Calgary

UNIVERSITY OF CALGARY

Addressing the Problem of Extrapolating Results from Randomized Controlled Trials

by

Liyan Zhou

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN PHILOSOPHY

CALGARY, ALBERTA

APRIL, 2024

© Liyan Zhou 2024

Abstract

Randomized controlled trials (RCTs) are experiments conducted in specific experimental settings by researchers. In educational research, RCT results obtained in experimental settings are often extrapolated/applied to other settings by educational practitioners. RCTs are recognized as a "gold standard" for obtaining reliable causal claims in education. However, RCTs face a serious problem of extrapolating RCT results, i.e., when practitioners extrapolate the results from RCT studies to their local settings, the causal claims sometimes do not hold in the new settings. This problem, according to Nancy Cartwright and her colleagues, can be explained by the input-output logic of RCT studies. This means that RCT studies work in a black box with no explanation of why and how the results are established, and therefore practitioners may blindly apply the results to local contexts. To help address this problem, Cartwright and her colleagues suggest scientists turn their attention beyond RCT studies. They argue that local practitioners can use ethnography to gather local information and use the information to localize the implementation of RCT results to different contexts. I agree that RCT results can

be applied more effectively if practitioners apply the results based on their investigations into local environments. To build upon their proposals, I argue that attention should not be moved away immediately from RCT studies because researchers can also help address the problem by recording data on different influencing factors and organizing them into useful causal information. In this way, researchers can reduce the burden on practitioners and provide practitioners with more resources for extrapolating the results.

Table of Contents

List of Figures

Introduction

Randomized controlled trials (RCTs) are a popular research method of designing scientific tests that investigate whether hypotheses are true or false. RCTs are recognized as a "gold standard" for obtaining reliable causal claims, and thus, occupy a special role in medical and social science research that other popular research methods do not have. However, RCTs are criticized for facing a serious problem of extrapolation, i.e., some RCT results cannot be extrapolated/applied to contexts outside of the original study setting (Cartwright and Hardie 2012; Connolly et al. 2017; Deaton and Cartwright 2018; Fuller 2021; Joyce and Cartwright 2018, 2020; Joyce 2019; Kvernbekk 2016; Morrison 2019). In order for RCT studies to be more useful in applications, scientists hope to explain why extrapolating RCT results may fail in contexts other than the original ones and to address this problem of extrapolation.

With a focus on educational research, this thesis discusses three main limitations of RCT studies that may result in the problem of extrapolation in education. These limitations concern the input-output logic of RCTs, the process of

randomization, and the applicability of RCT studies in value-laden contexts (Cartwright and Hardie 2012; Deaton and Cartwright 2018; Fuller 2021; Joyce and Cartwright 2020; Joyce 2019; Kvernbekk 2016; Morrison 2019). In response to the first limitation, the input-output logic of RCTs, Nancy Cartwright and her colleagues (Cartwright and Hardie 2012; Deaton and Cartwright 2018; Joyce and Cartwright 2018, 2020) turn their attention beyond RCT studies. They argue that reducing the possibility of failed extrapolations requires local practitioners to gather sufficient causal information about the study and local contexts. This thesis aims to build on Cartwright and her colleagues' approach to provide an interventionist alternative that builds on it. This alternative approach suggests that the attention should not be moved away from RCT studies immediately because RCT researchers can provide useful causal information by identifying potential factors that may affect causal relationships. The information can help practitioners understand the RCT results, thereby reducing their burden and facilitating extrapolations. Following Joyce and Cartwright (2020), this thesis uses the term "local" to describe contexts and settings outside of RCT studies, "practitioners" to refer to scientists who intend to use RCT results in their local contexts, and "researchers" to refer to scientists who conduct RCT experiments.

In this thesis, I employ the theory of interventionism à la James Woodward (Ross and Woodward 2016; Woodward 2003, 2010) to bolster my proposed

approach to helping address the problem of extrapolation using the concept of causal stability. Furthermore, I use interventionist-style causal graphs to visualize and clarify how causal information can be obtained. In Chapter 1, I will illustrate how RCT studies test causal hypotheses and establish causal relationships by introducing the elements, process, criteria, and methods of scientific testing in detail. In Chapter 2, I will introduce the problem of extrapolating RCT results and summarize three main reasons for this problem. I will also defend the special role of RCT studies against two alternative research methods, despite the problem of extrapolation. In Chapter 3, I will argue against Cartwright and her colleagues' approach to confronting the input-output logic of RCTs. I will also build on their approach to provide an interventionist account with useful causal graphs to help address the problem of extrapolation more effectively.

Chapter 1

Conceptual Background

Randomized controlled trials (RCTs) are a popular research method of designing scientific tests that investigate whether hypotheses are true or false. RCTs are recognized as a "gold standard" for obtaining reliable causal claims. This implies that RCTs test a special kind of hypotheses, causal hypotheses, and such tests are usually considered good tests. This chapter will introduce the conceptual background of RCTs as scientific tests. Scientific tests aim to investigate whether theoretical hypotheses are true or false, in which researchers use scientific methods, like experiments or simple observations, to figure out the extent to which a hypothesis is supported by evidence. The results of experiments or observations are compared to the predictions inferred from the hypotheses. Predictions are statements of the expected prospective results given a defined set of parameters provided by the initial conditions and auxiliary assumptions of the experiments or the observations if one assumes that the hypothesis is true. If the results agree with

the predictions, then it is probable that the hypotheses are true. If the results and the predictions are not consistent, then it is likely that the hypotheses are false.

To help understand how RCT studies establish reliable causal claims, this chapter will further explore the idea of scientific testing along the lines of Ronald Giere (1979, 1984). In Section 1.1, I will discuss the basic elements of scientific testing, which include theoretical hypotheses, predictions, initial conditions, and auxiliary assumptions. In Section 1.2, I will examine the criteria for good scientific tests, which include the requirements that the prediction be inferable[1], improbable, and verifiable. I will also use these criteria to analyze a famous experiment in education. In Section 1.3, I will discuss scientific tests related to causal hypotheses and introduce three main research methods of designing such tests, which include prospective designs, retrospective designs, and randomized experimental designs.

1.1 The Basic Elements of Scientific Testing

Theoretical Hypotheses

A theoretical hypothesis is a statement that aims to capture the regularity of relationships among factors in the system of interest to the researchers. For example,

[1] I use "inferable" instead of Giere's "deducible" in this thesis. "Deducible" is much more demanding than "inferable", which will be discussed later in Section 1.2.

the hypothesis “class size reduction has a positive impact on student achievement” describes the relationship between class sizes and student achievement in an educational system. Hypotheses can describe causal or non-causal relationships. This example is on causal hypotheses, which are the focus of scientific testing in educational research since researchers want to find out how to achieve the goal of improving educational results.

Predictions

A prediction predicts what will happen if the hypothesis is true. However, a hypothesis does not directly correspond to one specific prediction. Instead, one hypothesis can lead to multiple possible predictions. For example, if there is a sole hypothesis “class size reduction has a positive impact on student achievement in Class A”, it is hard to single out a prediction related to the real-world situation. The predictions can be “students in Class A outperform Class B, which has more students, on standardized tests”, or “students in Class A do better on standardized tests after Class A is reduced in size”, or “teachers rate the performance of students in Class A better than in Class B”, etc. The difficulty of determining one prediction results from the lack of important information, such as how to measure the impact on student achievement and how to determine if the impact is positive or negative. Therefore, to single out a prediction that can be used to test the hypothesis, it is first

necessary to identify some initial conditions that can determine the state of the relevant factors.

Initial Conditions

Initial conditions describe the occurrence of some states at the beginning of the experiment or the observation. They help determine the specific prediction by relating the prediction to the facts or reality being examined in experimental or observational designs. These conditions are “initial” since they involve all the conditions that researchers want to manipulate or control at the start of a scientific test. For example, in the class size reduction example, initial conditions should attempt to include all relevant factors that will influence the relationship between class sizes and student grades, such as course schedules, teacher quality, curriculum, and the aforementioned test criteria. The initial conditions might be that “course schedules and curriculum remain the same as the previous settings” and that “teacher quality are randomized by distributing teachers randomly”.

Auxiliary Assumptions

Ideally, all the influencing factors of a test will be included in the initial conditions. However, there are usually some residual factors left out of the initial conditions in practice. These factors are factors difficult to control in the tests or factors that

researchers are not aware of. To avoid overcomplicating the tests, researchers tend to make the auxiliary assumption that uncontrolled and unaware factors will not have a significant effect on the research. This auxiliary assumption is common in research and can potentially influence the results. Therefore, if a prediction is inconsistent with the results, researchers do not always conclude that the hypothesis is false. Instead, they may first examine the validity of the auxiliary assumption. For instance, they can take another test to check whether one uncontrolled factor can have an impact on the target factor or not. They can also try to find more factors that were previously unknown.

1.2 The Criteria for Good Scientific Tests

Three Criteria for Good Scientific Tests

As suggested by Giere (1984), there are three criteria for good scientific tests. They are the requirements that the prediction be deducible, improbable, and verifiable.

The first criterion is that the prediction is logically deducible from the hypothesis together with initial conditions and auxiliary assumptions. This means that the prediction does not contain any other information. If the hypothesis, initial conditions, and auxiliary assumptions were all true, there would be no possibility that the prediction is false. However, this condition is nearly impossible to achieve

in research in social sciences such as education. This is because initial conditions are difficult to exhaust all the conditions required for deductive reasoning since there are many influencing factors in social sciences. In this situation, the auxiliary assumption that these unlisted factors will not have an impact on the research may be too strong and not plausible/acceptable. For this reason, I lower the demand by using “inferable” in place of “deducible”. Here is an example in education that a prediction is inferable. Suppose that the hypothesis is “Reducing the class size has a positive impact on student achievement”. The initial conditions are “Class A is reduced in size compared to Class B in grades 1-3” and “both classes will use the same standardized tests to examine student achievement”. The prediction can be “students in Class A will have better grades than those in Class B”.

The second condition is that the prediction is improbable, which means that the probability of the prediction being true is only relevant to the hypothesis and its background knowledge. In raising this requirement, the purpose is to make sure that there are no current observable alternative hypotheses that could yield the same prediction. If multiple hypotheses can generate one prediction, then the result of the prediction cannot be used to prove the hypothesis definitively.

The third condition is that the prediction is verifiable. Researchers need to have reliable methods to verify the truth and falsity of the predictions. To be verifiable, the predictions should be set as observable and detectable. If a prediction is not

clear enough, for example, if it is vague and contains multiple possible results, then there is a good chance that the findings can be manipulated to interpret the experiment or observation results to make them consistent with the prediction. This means that the results cannot definitively falsify the predictions. Therefore, researchers need a clear and accurate criterion by which to determine whether the prediction agrees with the results.

A Good Scientific Test: the STAR Project

An influential class-size reduction project in Tennessee, the Student Teacher Achievement Ratio (STAR) (Word et al. 1990), is a good example of a good scientific test. I will describe this experiment and show how each parameter meets the criterion of a good study.

STAR was a four-year longitudinal study that aimed to analyze the relationship between student achievement and class size. Before the implementation of this program, a popular meta-analysis suggested that reducing class size to less than 20 students has a noticeable positive effect on student achievement (Glass et al. 1982). This hypothesis was considered problematic on the grounds that the study's sampling methods were biased, resulting in unrepresentative school and student types. Given the criticism, the Tennessee Legislature decided that a well-designed,

large-scale study of class sizes was required before investing in a costly class size reduction program. From this, the STAR project emerged.

The STAR project tracked the same group of students as they went from kindergarten (K) to grade 3, beginning in 1985-86 and ending in 1988-89. Approximately 6,400 students from 78 schools were randomly distributed into 128 small classes (13-17 students per teacher) and 200 regular classes (22-25 students per teacher). Once assigned to a class type, a student remained in the assigned class type for four years. The participating schools came from different regions and different types of communities (i.e., rural, urban, suburban, and inner city). Some important factors that the researchers were able to control, including curriculum, schedules, and expenditures, remained the same for each class type and school as they were before the program was initiated. Various factors that could have an impact on the relationship between class size and student achievement, such as teacher planning time and teacher experience, were not controlled, but were kept random instead.

The method of assessing the impact of class size reduction involved a systematic comparison of the grades of selected standardized tests among students in small and regular classes. Students were assigned individual identification numbers and the test score data for each student was collected at regular times for the duration of the study. Students were required to take standardized tests including

the Stanford Achievement Test, the Basic Skills Criterion Test in reading and math, and the Self-concept and Motivation Inventory.

The test result is encouraging. It shows that there is a strong, positive class-size effect in all K-3 grade levels. Such an effect showed that students in small classes get significantly better test scores than those in regular classes. The hypothesis of the test is considered to be true because of the success of the test. That is, the reduced student-teacher ratio has a positive effect on the achievement and development of students in grades K-3 in Tennessee's public elementary schools.

To verify the validity of the test result, the next step is to conclude the elements of this test and use the elements to check whether the test is designed as a good test (i.e., whether the prediction is inferable, improbable, and verifiable). The basic elements of this test included: (1) Hypothesis: Reducing class size to less than 20 students has a positive impact on student achievement in grades k-3 in Tennessee's public elementary schools. (2) Prediction: In this Tennessee experiment, the small class mean for grades k-3 is consistently higher than the regular mean for all school types on all standardized achievement measures. (3) Initial conditions: Students and teachers are randomly assigned; schools are spread over the state; all school types (i.e., rural, urban, suburban, and inner city) participate in the program evenly; class types are small and regular classes; criteria for measuring student achievement are

standardized test scores. (4) Auxiliary assumption: Nothing outside of the researchers' control or knowledge will interfere with the prediction.

These elements can then be used to verify whether STAR meets the criteria for a good test or not. First, the prediction is inferable. The initial conditions specify what kind of class-size reduction is conducted in the program and that the "impact" is measured by the standardized tests scores. In this case, the prediction can logically follow from the hypothesis together with initial conditions and the auxiliary assumption. Second, the prediction is improbable because the STAR program has tried to rule out other possibilities that would lead to the result that students in smaller classes have better test scores by controlling different influencing factors. Third, the prediction is verifiable because the STAR program uses standardized test scores to evaluate student achievement. The test scores are accurate criteria. Therefore, STAR is a good scientific test that established a positive experimental result.

1.3 Scientific Tests Related to Causal Hypotheses

Causal Hypotheses

Like the STAR test, many educational studies aim to identify whether there is a causal relationship between two factors in an educational system. What the

researchers are testing for, in these cases, is a special kind of theoretical hypothesis called causal hypotheses. A causal hypothesis invokes causal relationships between variables. For example, “smoking causes lung cancer”. Although determination of causality is always the object of these studies, the term “cause” is not always used when formulating causal hypotheses. Many words can reflect causal relationships between variables, such as “have an impact on”, “facilitate”, and “encourage”. Such words do not imply a relationship that is weaker than a causal relationship; they are synonyms of causality. In this sense, the hypothesis mentioned above that “class size reduction has a positive impact on student achievement” is a causal hypothesis.

One way to understand why these words denote causation is to refer to the counterfactual theory of causation. David Lewis (1979) famously developed the theory that event X is a cause of event Y if they can be connected by a counterfactual form such that “if X had not happened, then Y would not have happened either”. According to this account, for instance, class size reduction is the cause of the improvement in student achievement because if class size had not been reduced, then student achievement would not have improved, where the event X is “class size is reduced”, and the event Y is “student achievement is improved”. James Woodward (2003, 2010) suggests identifying conditions under which variables, instead of events, are manipulated or controlled. Using Woodward’s account to analyze the above example, X and Y are variables that stand for different states of

property, which means that variable X is “class size” and variable Y is “student achievement”. X can have different states as “class size” is “reduced” or “improved”, and Y can have different states as “student achievement” is “better” or “lower”.

Scientific Tests of Causal Hypotheses

There are three main methods of designing scientific tests of causal hypotheses (Giere 1984). They are prospective designs, retrospective designs, and randomized experimental designs. The first two methods are observational, or non-experimental, designs that do not impose any manipulation on the population or environment, but merely investigate relevant factors present in the observed population. In contrast, randomized designs are experimental designs that are conducted in a manipulable setting, where researchers expect to control all factors that may have an effect on the relationship between variables.

In prospective designs, researchers select a population suitable to be observed over time, observe it for the effects related to the research objectives, and record all potentially relevant factors. When researchers observe effects in any individuals in the population, they attempt to relate these effects to specific factors and consider these factors to be the cause of the effects. For example, the Nurses’ Health Study, established in 1976, is one of the largest prospective investigations of risk factors

for major chronic diseases in women (Nurses' Health Study, n.d.). Given that nurses have adequate health knowledge and are easily followed over time, they were selected as the study population. Every two years, researchers check to see if they have a chronic condition and give each member a follow-up questionnaire asking about health-related topics such as smoking, diet, and hormone use. In this way, the project can provide rich information about suspected causal factors of a number of chronic diseases in women. However, all these factors confound together, making it difficult to find a clear and robust causal relationship between various factors and chronic conditions. Further detailed studies are needed to find this causal relationship.

To reduce the effect of confounding factors, researchers may divide the population into two groups at the beginning of a prospective study. Two groups are alike in every feature, except that members of one group have the suspected causal factor hypothesized by the study, while the other does not. Except for the grouping, the design is the same as the prospective study described above. That is, all members do not exhibit the effect at the beginning. The researchers do not intervene in the study, but simply collect information and wait to see if there is an effect in the future. The difference that grouping makes is that researchers can calculate whether there is a statistically significant difference between the data from the two groups. If there is, then the hypothetical causal relationship between the suspected

causal factor and the target effect can be justified. An example is that in the 1950s, the American Cancer Society divided nearly 20,000 healthy men into groups of smokers and nonsmokers, recorded information on members' smoking for four years, and checked if any members died of lung cancer.[2] The conclusion of this study was that seven times more smokers died of lung cancer than nonsmokers. This significant difference provided a good justification of the causal hypothesis "smoking causes death from lung cancer".

In retrospective studies, researchers select a population that has already shown the specific effects related to the research objectives. They look backward, examine possible factors that may have contributed to the effects, and try to identify one or some of these factors to be the cause of the effects. The problem with prospective designs also applies to retrospective designs, where various factors are so confounding that researchers can hardly rely on simple retrospective designs to find clear causal relationships between variables. The solution to this problem is still to create a control group to be compared to the target group. Researchers attempt to find a group whose members do not show the target effect and examine whether the suspected factors have previously appeared in the members. If the suspected causal factors are rarely found in members of the control group and on a large scale in

[2] This example is from Giere 1984.

members of the target group, then this difference in the population can serve as good evidence for the hypothesis that the suspected causal factors are the causes of the target effects. In the 1960s, doctors in Britain began investigating an abnormal increase in the number of young women suffering from blood clots.[3] The doctors hypothesized that the possible cause was the use of birth control pills. They found that 45% of the women in the target group, who developed blood clots, had taken the pills in the month before they were admitted to hospital, while only 9% of the women in the control group, who were admitted for serious medical conditions other than blood clotting, were found to have taken the pills. This difference is significant enough to suggest that the use of birth control pills is the cause of the abnormal occurrence of blood clots.

Prospective and retrospective designs track the real lives of populations, which leads to two advantages and two disadvantages. The first advantage is that researchers do not need to impose manipulations in the studies, which helps to avoid many ethical, practical, and technological issues. The second advantage is that researchers can observe the natural history of causal relationships between variables, which allows them to obtain a great deal of information on potential factors that have an impact on the research objectives. The first disadvantage is that

[3] This example is from Giere 1984.

sometimes researchers may not have access to the data, or the data may not be of high enough quality, which can affect the validity of the research conclusions. In prospective studies, it is usual to lose track of some members and thus lose their information during the follow-up. In retrospective studies, researchers may not be able to find some information, or they seek information from secondary sources (records) and members' memories, which are less reliable and complete. The second disadvantage is that even with the inclusion of a control group, it is difficult to rule out the effects of confounding factors, especially those that co-occur with suspected causal factors. For example, smokers are more likely to live in an unsafe environment than nonsmokers. And the unsafe environment is one factor that was not excluded by the prospective design. In the blood clot example, a large percentage of women who use birth control pills have previously been pregnant. Having been pregnant was not excluded by the retrospective design either. This implies that what is observed in the prospective and retrospective studies may be correlations rather than causal relations.

These two disadvantages indicate that the causal relationships provided by the first two designs are not sufficiently robust. The third method, randomized experimental designs, or randomized controlled trials with the control group, can provide better justifications for causal relationships since this approach does not have these two disadvantages mentioned above.

In RCT studies, researchers select representative experimental participants and randomly assign them to two groups: the treatment group and the control group. The treatment group receives the manipulations related to the research objectives, while the control group does not receive the manipulations and remains in the normal condition. The manipulations will be considered effective and to be a cause of the research effect if there is a significant difference between the two groups. Researchers manipulate every factor and record data in RCTs, making the data relatively completer and more reliable than the other two designs. Furthermore, randomized grouping enables an experiment in which all members will have an equal chance of falling into either group. This can neutralize the impact of confounding factors in a way that makes the distribution of these factors equally likely in both groups. That is to say, researchers expect fewer factors other than the target factors to be present in the treatment group members but not in the control group members. This will reduce the probability of any confounding factors being the cause of the target effects in the experiments. The STAR project mentioned in Section 1.2 is a typical case of RCT studies. It randomly divided students and teachers into two types of classes, small and regular classes. All characteristics of the two types of classes were designed to remain the same except for class size. The researchers used standardized tests to examine student performance in both groups. The conclusion was that students in small classes got better grades on average than

those in regular classes. Given that there were no other factors that differed between the two groups, class size reduction was considered to be the cause of better student achievement.

Chapter 2

Randomized Controlled Trials in Education

Randomized controlled trials (RCTs) are recognized as a "gold standard" in education for obtaining reliable causal claims that can be replicated in study settings. However, RCTs are criticized for facing a serious problem of extrapolation, i.e., some RCT results cannot be extrapolated/applied to contexts outside of the original study setting (Cartwright and Hardie 2012; Connolly et al. 2017; Deaton and Cartwright 2018; Fuller 2021; Joyce and Cartwright 2018, 2020; Joyce 2019; Kvernbekk 2016; Morrison 2019). Three limitations of RCT studies are proposed to explain the problem of extrapolation. These limitations concern the input-output logic of RCTs, the process of randomization, and the applicability of RCT studies in value-laden contexts (Cartwright and Hardie 2012; Deaton and Cartwright 2018; Fuller 2021; Joyce and Cartwright 2020; Joyce 2019; Kvernbekk 2016; Morrison 2019). Some educational researchers believe that these limitations lead so severe

the extrapolation problem that RCTs are not a reliable research method, and therefore researchers should use other research methods such as surveys and ethnography instead of RCTs (Connolly et al. 2017; Joyce and Cartwright 2020; Kvernbekk 2016).

In this chapter, I will first in Section 2.1 introduce how RCT studies can establish reliable causal claims and are recognized as valuable research method in education in the light of the theory of interventionism. In Section 2.2, I will illustrate the problem of extrapolation with an example in California. In Section 2.3, I will summarize three main limitations of RCT studies that may lead to the failure of extrapolations. In Section 2.4, I will introduce two main alternative research methods in education, compare them with RCTs, and conclude that RCTs are a method that educators should adhere to as they play an indispensable role in educational research.

2.1 RCTs as a "Gold Standard"

As briefly introduced in Chapter 1, researchers randomly assign participants into two groups: the treatment/experimental group and the control group in RCT studies. Treatment groups receive experimental manipulations, while control groups do not and remain on usual treatments. If a significant difference between the two groups

is observed during the experiments, such an effect will be considered to be caused by the manipulations. The reason is that, ideally, RCT studies create environments where manipulations are the only difference between two groups that can be responsible for the variance in their outcomes. The difficulty lies in how to create such ideal environments. Researchers expect that every factor in experiments that has a potential impact on the observed difference, except the manipulated one, would be controlled in a stable status. It is believed that RCT studies can approximate this expectation through randomization. The randomized grouping enables all participants to have an equal chance of falling into either group. Factors difficult to be controlled directly can thereby be randomized in experiments, which makes the distribution of these factors equally likely in both groups. Thus, participants of either group are less probable to possess values of factors that vary significantly from those of the other group. In this sense, RCTs can provide clear evidence of manipulations being the sole cause of experimental results and are the "gold standard" for obtaining reliable causal claims.

The theory of interventionism is a useful tool to help understand the causal relationships established by RCT studies. The essential idea of interventionism, as Ross and Woodward (2016) suggest, is that "if you can intervene on [X] in such a way that changes in [X] are reliably associated with changes in [Y], then [X] causes [Y]" (p. 38). Interventionism examines causal relationships from the perspective of

manipulation: interventions are ideal manipulations that have ruled out the influence of all potential confounders to leave the change in Y fully derived from the intervention on X. A confounder is a factor C that is "responsible for the presence of an association between [X] and [Y] and that makes it look as though [X] causes [Y], even though it does not" (p. 38). This is to say, if the impact of C is not controlled in the experiment, researchers cannot make the claim that X independently causes Y because C may contribute to the change in Y. In this case, the manipulation of X is probably not the cause or is merely a part of the cause of the effect Y, and thus, does not serve as an intervention.

For example, imagine a case that X is class size and Y is student performance. Teacher quality can influence student performance and will be a confounder C if it is not controlled in the experiment. Suppose that all small classes happen to have teachers of higher quality, which obviously can increase student performance, the reduction of class size cannot be considered as the cause of better student performance. This does not mean that class size reduction has no contribution to the effect, but that it does not independently cause the effect. The result of this RCT study, "class size→student performance", is thereby not a reliable causal claim and may not maintain in replicated studies. Therefore, successful interventions require all potential confounders in control and not to have unexpected impacts on the relationships between X and Y. Figure 1 presents an intervention on X that causes

the change in Y perfectly as there is no uncontrolled confounder influencing the causal process. In contrast, figure 2 presents a manipulation of X together with the influence of a confounder C leading to the change in Y. Two question marks in the figure mean that it is underdetermined the change of which factor, X or C, causes the change in Y.

Figure 1: An ideal intervention. Directed acyclic graph is used to depict the causal relationships among variables. This graph is composed of variables X and Y and the arrow. X and Y represent different properties, and the arrow represents a causal relationship from X to Y. Accordingly, X represents the causal property, and Y represents the effect property. This means that if property X changes, then there will be a change in property Y. The graph presents an ideal intervention under the condition that the manipulation of the value of X is the sole cause of the change in the value of Y as no confounder influences the causal process from X to Y.

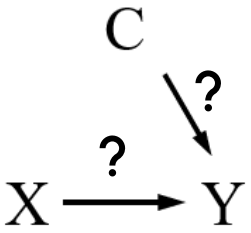

Figure 2: A manipulation with confounders. C represents a possibly confounding property, and the question marks represent the underdetermined status of the relationships. The graph presents a manipulation where confounders are not well controlled. It is not an ideal intervention because researchers do not have enough information to determine whether the change in Y is resulted from the manipulation of X or the influence of confounders.

This idea of interventionism highlights the relationship between X and Y. A causal explanation of the relationship, according to Campbell (2013), is too complicated and of no need at least in the contexts of social sciences.[4] First, the processes of how X leads to Y may involve enormous influencing factors and require different levels of explanation, which are beyond the current scientific understanding of humans. Second, as mentioned above, how the causal processes work is not important in determining the relationships between X and Y once other factors are controlled for. Thus, experiments that employ the theory of interventionism can establish reliable causal relationships although they omit the explanations of the causal processes and are therefore ideal for use in social sciences. RCT studies are experiments that hope to approximate the idea of interventionism. As mentioned, researchers assign invariable values to variables or keep them randomized to control different factors. These factors will not be confounders to mask the impact of X on Y.

The STAR project described in Chapter 1(Word et al. 1990) can show how RCT studies as a gold standard establish reliable claims in practice. In the STAR project, students were randomly assigned to small or regular classes. Several important factors that the researchers were able to control, including tests,

[4] Campbell uses experiments in psychiatric research to explain this idea.

curriculums, schedules, and expenditures, remained approximately the same for both class type as these factors were before the program was initiated. Various factors difficult to control, such as teacher planning time and teacher experience, were kept randomized by distributing teachers randomly to the classes to minimize their impact. In these two ways, researchers tried to rule out other possibilities that would lead to the difference between the two groups. There was still a difference between the groups, students in the small classes had better test scores. The only explanation remaining was the smaller class size. Class size reduction was thus considered to cause better student performance in the experiment independently.

2.2 The Problem of Extrapolating Results from RCTs

The STAR project shows how RCTs establish reliable causal claims by controlling influencing factors in experiments. It hypothesized and predicted that class size reduction will have a positive impact on student achievement. And the test result was that the small class mean is consistently higher than the regular class mean on all measures in this experiment, which strongly proved the hypothesis of the program. After seeing the success in Tennessee, the California government was eager to replicate this effect. Disappointingly, however, as shown in the final report

(Bohrnstedt and Stecher 2002), practitioners met a great failure when they tried to extend the STAR result to schools in California. They undertook a massive instructional reform that reduced the size of a large number of classes to less than 20 students. The result was that students in small classes did not significantly outperform students in regular classes in California.

In response to this result, Cartwright and Jeremy Hardie (2012) identified two salient influencing factors that contributed to the failure to apply the STAR result: instructional space and teacher quality. California, at the time, did not have enough spare classrooms and qualified teachers to cope with the number of extra classes created by downsizing. The classrooms assigned to small classes were not as good as the original ones and could not accommodate a variety of learning activities. In addition, more than 10,000 additional teachers are hired for this policy, but many of them did not have sufficient teaching experience and skills. As a result, the teacher quality in small classes was worse than the average of regular classes.

In this case, the auxiliary assumption in the experiment is invalid. Researchers, both in Tennessee and California assumed that nothing outside of the researchers' control or knowledge would interfere with the prediction. These researchers actually noticed these two factors and set corresponding initial conditions in experimental designs by randomly assigning all classes to different spaces and all teachers to different classes. However, researchers did not expect that reducing

class sizes would expand the number of classes to a level they could not handle. This led to the problem that factors of instructional space and teacher quality were beyond the control of initial conditions. When the manipulations are repeated in a different setting, factors that were controlled through randomization are not controlled with respect to the before-manipulation group (California schools before the class size changes) and the after-manipulation group (California schools after the class size changes). The extrapolation is therefore failed. This failure in California is regarded as typical evidence for RCTs having the problem of extrapolation: the reliable causal claims given by RCT studies cannot be stably applied to other contexts.

2.3 Three Limitations of RCT Studies

Three limitations of RCT studies are recognized as the main reasons of the problem of extrapolation.[5] First, RCT studies run in an input-output logic and provide no explanations of why and how the results are generated. Second, the randomization

[5] Besides these limitations, scholars also criticize RCTs for other problems, including the difficulty to measure the effect of RCT results and ethical issues of conducting RCTs in real educational system. These problems are beyond the scope of this thesis. See Social Science & Medicine 210 for a paper from Angus Deaton and Cartwright and 18 commentaries on the paper for the reference.

in RCT studies may fail in a dynamic environment such as education. Third, manipulations in social contexts like education are extremely value-laden. In this section, I will introduce these three limitations in detail.

Input-output Logic[6]

The first limitation of RCT studies is that they are not able to provide any relevant causal information given that the studies run in an input-output logic (Cartwright and Hardie 2012; Fuller 2021; Kvernbekk 2016; Morrison 2019). The input-output logic means that RCT studies input an X-value and get an output of a Y-value, presenting experimental outcomes in the form of "X→Y"[7] without any causal explanation of the outcomes. The form "X→Y" illustrates that the experimental manipulation of factor X can cause the expected change in factor Y, but it does not explain why and how X causes Y. For instance, the result of the STAR project was presented as "class size→student performance", indicating that a specific manipulation (reduction) of class size causes the observed change in student performance. However, the report did not provide a narrative of how these two factors interact to achieve such a causal relationship. As explained in Section 2.1,

[6] This limitation will be recalled and replied in Chapter 3.

[7] In this thesis, X represents the cause/manipulation, Y represents the effect, and the arrow represents a causal relationship from X to Y.

under the condition that all other factors are controlled, when the value of Y changes when researchers manipulate the value of X, this change is considered to be caused by the manipulation. In RCT studies, according to these scholars, researchers can merely observe the values of X and Y, but not the causal process of X to Y.

Educators usually meet problems in applying the results to specific situations when they do not know the answers to some questions such as "is there a special reason to choose X to conduct the manipulation?"; "how is X related to Y and leads to a change in Y?"; "are there any other factors that play a role in the causal process?" If they blindly follow the input-output logic of RCTs, which means emphasizing only X and Y and being unaware of any potential related factors, they risk missing or mistaking important conditions of the relationship and developing partial causal claims. Causal claims can then only hold in environmental settings that are the same as or at least similar to the original experimental settings. Once the settings change, educators have no way of knowing how to control the influence of other factors on the relationship between X and Y. Extrapolation is thus likely to fail.

Randomization and Post-randomization

The second limitation is related to a significant feature of RCT studies: randomization. Recall the discussion in Section 2.1, randomization is expected to balance differences across populations and neutralize the impact of confounding

factors by selecting and distributing each participant of an experiment randomly. For example, to reduce the influence of different quality of teachers in an experiment, researchers may randomly select 100 teachers among 200 and randomly assign them to different experimental groups. However, the expectation of conducting randomization is hard to be met for two reasons (Deaton and Cartwright 2018; Fuller 2021; Joyce 2019; Morrison 2019).

First, the balancing cannot be perfect. Ideally, Randomization would prevent a group from having distinctive features than others. This ideal situation relies on idealized assumptions of probability, which are not always the reality. For instance, in an environment where the population of males is roughly equal to females, a randomly selected group should, probably, have a close number of male students as female students. However, disproportionality may appear in some cases. Second, even assuming that the balance is perfect, changes may occur after randomization. In a dynamic and evolving environment such as education, it is impossible to hold variables constant in the manner required by RCTs. The real environment continues to change after RCT studies started, while the manipulation on the factor X has an impact on the whole environment settings instead of merely the factor Y. Therefore, a new system emerges after randomization, which can affect the features groups under study. If researchers conceive that the causal claims provided by RCT studies are generated under invariable pre-designed initial conditions, while the change of

conditions have considerable influence in the causal relationship between X and Y, then they may fail to apply these causal claims to other situations.

The randomization problem is instantiated in two relationships, one between the sample group and the target group and the other between the treatment group and the control group. As for the first pair, the problem is that the sample group may not be sufficiently representative of the target group. A sample of a population refers to specifically selected individuals based on their diversities and best represents the target population that researchers want to study. A sample genuinely representing the target should be the same or at least similar enough with respect to all factors that may affect the outcome. Nevertheless, the sample and the target groups only match in expectation. For instance, in the STAR project, the sample group was drawn from four different areas: rural, urban, suburban, and inner city. The project wanted to mirror the diversity and characteristics of the entire population in Tennessee. It considered the differences between schools in the four areas. However, the chosen rural schools may still differ slightly in some perspectives from other rural schools in Tennessee.

Also, problems may arise between the treatment and the control groups. Members in both groups are supposed to be similar in various factors other than the manipulated one in the experiments. Researchers tend to control possible causal factors that are easy to control, such as school schedules, rules, and expenditures,

while randomizing all other factors that are difficult to control. The mentioned drawbacks of the randomization method can be shown here as well. For example, the treatment and the control groups may have teachers of similar quality at the beginning of an experiment. However, teachers in the control group may realize that their students are expected to have lower grades and become unwilling to passionately educate their students. Teacher quality can change in this case post randomization.

Value-laden Contexts

The third limitation is that applications of RCTs in value-laden contexts are vulnerable to research biases (Becker and Luthar 2002; Joyce and Cartwright 2018). RCT is a widely used and highly regarded research method in medical research. Medical practitioners often use RCTs to test the effectiveness of new medicines. The selected treatment group will try new medicines, while the control group only takes placebos or receives existing standard medicines. If participants in the treatment group are in a better situation than those in the control group, and there is a significant difference between two groups' situations, then researchers prefer to conclude that the new medicines have a positive effect. In an ideal experiment, neither the participants nor the researchers should know which group is the treatment group and which is the control group. They would not be able to tell the

new medicines from the old ones or the placebos. This method, known as a double-blind study, reduces possible bias from participants and researchers, allowing the RCT results to be more independent of personal factors.

Such a method is tailored for laboratory environments but hard to be conducted in more value-laden contexts like education. Blinding is not easy to implement in educational experiments. Unlike in medicine, educational manipulations are often too obvious. In medicine, new medicines and placebos are similar and can be difficult distinguished, while the difference between small and regular classes is realized immediately. Even if teachers and students in the experiments do not know the purpose of the distinction, they will know that there is a difference lies in the size of the class. Further, educational manipulations usually involve more human behaviors and interactions, which are prone to personal views. It is possible that scientists may unintentionally divulge information about the experiments, such as purposes or expected results of the experiments during the pre-experiment training for teachers. Teachers, in turn, may more or less provide information related to the experiments to their students during the teaching process. For example, if teachers know which group students are assigned to, i.e., small classes are the treatment group and regular classes are the control group, then they may have a biased expectation of better performance from the students in small classes. Such an expectation can motivate students to perform better. Similarly, if students know

which group they are assigned to, they are likely to change their behavior in order to cater to the expectations of others. It is also probable that they will become more confident and thus perform more prominently.

2.4 Alternative Research Methods in Education

As shown in Section 2.3, education consists of a dynamic and complicated system, while RCT studies usually work in a relatively stable and simplified environment deliberately designed by researchers. RCT studies play a limited role in education because of this gap. In this context, many educators strongly promote the use of other research methods, such as surveys and ethnography (see Mertens 2023).[8] In this section, I will present these methods and compare them to RCT studies.

Surveys are conducted by asking participants a set of questions, which can be close-ended or open-ended. Close-ended questions have a pre-defined set of answers, asking responders to choose among these answers, while open-ended questions require responders to present their own views with words. By asking close-ended questions, scientists can code each answer as data. For example,

[8] Other research methods like grounded theory, phenomenology, and case study are not similar to RCT studies in terms of the applied ranges, and thus are not often compared to RCT studies. See Mertens' 2023 book for the reference.

answer “yes” is coded as “0” and answer “no” as “1”. And thus, the answers can be analyzed and evaluated statistically. On the contrary, answers gathered from open-ended questions cannot be analyzed statistically. They are expected to provide a number of opinions about the questions from the participants. This method is easy and cheap to use. The questionnaire can be distributed and collected without investing more time and money than other methods once its content is organized.

Ethnography requires scientists to immerse themselves in the social environments they are studying in order to gain a holistic and contextual understanding of their research. They will learn local languages and cultures, access to past archives, collect local objects related to the research, interview local people, and observe every detail, with interviewing and observation being the two main segmented research methods in ethnography.

Interviews can be formal or informal. In formal interviews, scientists meet the respondents at a designated time and ask structured questions about the studies. In informal interviews, scientists have the flexibility to start conversations with people relevant to their studies. Answers recorded in interviews may be less honest and more biased than in surveys due to the lack of anonymity. This limitation prompts scientists to conduct careful observations during the interview processes to obtain additional information to contrast and complement the interview research.

In observational research, scientists observe both people's behaviors and physical objects. For example, teachers' teaching styles and class interactions are valuable information to observe. Researchers will participate in classes to make notes or conduct video recording. Physical objects like school maps, regulations, and buildings also deserve observation. Meaningful information may be implicit in unobtrusive traces. The condition of the books, both old and new, shows how much students use them; the number of papers in the office trash can shows teachers' workload; and the presence of classroom stickers shows the teachers' focus on students' creativity. If a teacher claims in an interview that she values the creativity of her students, but there are no student creatives in the classroom, then the credibility of the answer goes down.

The obvious advantage of surveys and ethnography compared to RCT studies is that scientists can gather information about how something occurs in the natural environment as No manipulation will be performed on the research subjects. Scientists can gain a deep and holistic understanding of the research topic in an ethical manner. However, information provided by ethnography is difficult to be standardized or generalized to broader populations because it is limited to the specific place where the information originates. Ethnography also has the potential to increase the risk of research bias since the perspectives of both the observer and the observed are mixed into the information.

Furthermore, the avoidance of manipulations can adversely affect surveys and ethnographic studies. While scientists actually physically manipulate the factors in experiments like RCT studies, in surveys and ethnographic studies that are purely observational, as suggested by Ross and Woodward (2016), scientists merely observe patterns of association of factors in nature. The limitation of such evidence is that the observations that X is always found presented with the presence of Y cannot exclude the possibility that some other factors that co-occur with X have an impact on Y. However, RCT studies can respond to the objection that some other factors are the cause of Y because they have already controlled the influence of these factors by manipulations. RCT studies, therefore, have an epistemically privileged role in providing evidence for causal relationships. They can play a unique role in research compared to other research methods and are worthwhile for scientists to keep using them in educational studies.

Chapter 3

How Philosophers Can Help Address the Problem of Extrapolation in Education

RCT studies, as described in Chapter 2, try to eliminate the influence of possible confounders and establish causal relationships between two research variables. Other popular research methods in education such as survey and ethnography provide only correlations between variables because a great number of confounders cannot be ruled out by these methods. Therefore, RCT studies occupy a special role and deserve to be used in educational research. In order for RCT studies to be more useful in applications, scientists hope to address the problem of extrapolation that some RCT results may fail to be applied in contexts other than the original experimental contexts. This chapter aims to critically introduce one way that Cartwright and her colleagues help address the problem of extrapolation in

education and to build on Cartwright and her colleagues' account to provide an alternative way to help address the problem from the interventionist perspective.

This chapter focuses on how to confront the first limitation of RCT studies, namely the input-output logic of RCTs. Recall the brief introduction in Chapter 2, the input-output logic means that RCT studies provide the sole result as "X→Y" without explanations of why and how X leads to Y. This, according to Cartwright and her colleagues (Cartwright and Hardie 2012; Deaton and Cartwright 2018; Joyce and Cartwright 2018, 2020), is an important reason why extrapolations of RCT results may fail. Without causal explanations, local practitioners who apply the results to their contexts may blindly follow the input-output instructions and cannot fill in the gap between the experimental and local contexts. Due to this limitation, Cartwright and her colleagues turn their attention beyond RCT studies. They suggest that local practitioners can use other research methods like ethnography to gather sufficient causal information about the experimental and local contexts and then fill in the gap between two contexts.

I agree with this account that a lack of causal explanation can lead to blind and, likely, failed applications. However, I consider that the attention should not be moved away from RCT studies immediately. In many cases, RCT researchers can nevertheless provide useful causal information by identifying potential factors that may affect causal relationships. This can help practitioners understand the RCT

results, thereby reducing their burden, helping fill in the gap between experimental and local contexts, and facilitating extrapolations.

In Section 3.1, I will present Cartwright and her colleagues' view on how to confront the input-output logic of RCTs and the problem of extrapolation. In Section 3.2, I will argue against this account and provide an account based on it in the light of the theory of interventionism à la Woodward (Ross and Woodward 2016; Woodward 2003, 2010). The main idea is that, according to the concept of causal stability, RCT researchers can provide useful causal information and then reduce the burden on practitioners. Finally, in Section 3.3, I will explain in detail how scientists can employ the interventionist account in practice by providing and interpreting two sets of interventionist-style causal graphs. These graphs can visualize and clarify the factors that potentially influence the causal relationships in RCT studies. The hope is that these factors can then be used by researchers to gather useful information and arrive at more stable and applicable experimental results. In this sense, the interventionist alternative reduces practical burden and helps address the problem of extrapolation more effectively than Cartwright and her colleagues' approach.

3.1 Cartwright and Her Colleagues on the Problem of Extrapolating Results from RCTs

Input-output Logic Revisited

RCT studies manipulate the value of one factor X and observe the change in factor Y's value without mentioning in experimental reports why and how the manipulation of X may lead to the change in Y. This input-output logic is considered by Cartwright and Hardie (2012) a limitation of RCT studies and an important reason why extrapolating RCT results may fail in local contexts. On page 11, they say, "[we] do not think [using only RCT-approved policies] is a good rule, because we think that to decide about effectiveness requires an open-ended process of thinking that is inevitably contextual and cannot be reduced to rules." This is what they consider as the fundamental problem with the input-output logic of RCT studies. In other words, as Joyce and Cartwright (2020, p. 1045) put it, "[these] can support causal ascriptions (''It worked'') but provide little basis for local effectiveness predictions (''It will work here''), which are what matter for practice."

The idea behind these two claims is that RCT studies are merely able to provide claims about what works in experimental settings instead of what will work in local settings. Maintaining causal claims in various contexts require a strong premise mentioned in Section 2.1: every factor other than the manipulated one is controlled

in experiments. However, the premise does not hold widely given that influencing factors in local settings are enormous and different from those in experimental settings. The enormous factors make it impractical for practitioners to control all factors in local contexts. Practitioners may then choose to control only the factors reported by RCT studies as needing control. This blind application may not be effective because influencing factors in two settings are different. Different values of factors or new factors in local settings may prevent manipulations from producing the anticipated effects. As a result, causal claims established by RCT studies may fail to maintain in contexts other than the experimental ones.

This explanation can lead to a strong and counter-intuitive conclusion that causal claims cannot maintain in local contexts, namely, that causal claims cannot be extrapolated at all because experimental and local settings may always differ even slightly. Scholars are reluctant to accept this conclusion because it is inconsistent with the fact that many RCT results have been successfully applied. They then attempt to find a way to account for the successful extrapolating cases and to extend such success. I will present one way from Cartwright and her colleagues to address the problem of extrapolating results from RCTs.

Causal Principles

Cartwright and Hardie (2012) explain successful extrapolations by introducing the idea of causal principles. They argue that different settings have different causal principles for the production of specific effects.[9] These principles characterize the conditions, as specified by a variety of factors, under which the manipulations can lead to the effects. Causal relationships identified by RCT studies fit the causal principles for experimental settings and can be repeated in the same settings. Such relationships may also sometimes fit local causal principles in unknown ways, resulting in successful extrapolations even though practitioners had blindly extended RCT results to their local contexts.

Cartwright and Hardie (2012) suggest that the problem of extrapolation can be solved once practitioners "get a good argument from 'it works somewhere' to 'it will work here' [by identifying] facts about causal principles here and there" (p. 23). With information about causal principles for both experimental and local contexts, practitioners are in a better position to understand how manipulations cause the observed effects in experiments and how to localize the causal relationships in their new settings. Joyce and Cartwright (2020) summarize three kinds of useful

[9] They further provide a technical description of causal principles for the production of an outcome y in the form of $y(i)c=a_1+a_2y_0(i)+a_3b(i)x(i)+a_4z(i)$ (p. 26). See the illustration of this formula in their 2012 book.

information that practitioners should gather for extrapolations: structures, support factors, and derailers. Structures are social/economic/cultural/material systems that underlie causal processes. Support factors help manipulations achieve the intended effects. Derailers undermine the manipulations. Joyce and Cartwright expect practitioners use methods like survey or ethnography to assess whether local structures can afford the operation of the given causal relationships, whether support factors are in place, and whether derailers are eliminated.

3.2 An Interventionist Response to Cartwright and Her Colleagues' Approach

As discussed in Section 3.1, Cartwright and her colleagues argue that differences between the study and local contexts require practitioners to gather information about causal principles to increase the probability of successful extrapolations. Some practitioners who do not realize this may blindly follow the input-output instruction of RCT results as the studies cannot provide useful causal information about why and how X leads to Y, not to mention information about the structures, support factors, and derailers of their results. Practitioners in this case manipulate X and expect Y to change by default. Once local settings are different from the original experimental ones, practitioners have no way of knowing how to control

other factors and lead the causal process from X to Y. Therefore, such applications are vulnerable to failure, while their successes rely on cases where RCT results happen to fit local causal principles.

In this section, I will reply to this view by arguing that while RCT researchers cannot provide causal explanations, they have the potential to provide useful causal information during experiments. I will use the theory of interventionism, especially its concept of causal stability, to suggest that researchers can at least partially identify possible factors that may influence the resulting causal relationships before extrapolations. As a result, researchers can establish more causally stable RCT results and reduce the burden on local practitioners.

The Potential to Provide Useful Causal Information

The critique that RCT studies cannot provide any useful causal information about why and how X causes Y is understandable. This is because RCTs as a research method is supposed to provide the sole statistical analysis of changes in the value of X and Y. Such analysis skips the causal process from X to Y completely and researchers have no way to know this process according to the statistics. One way to respond to the critique is to distinguish between RCTs as a method and concrete RCT studies. Although the RCT as a research method can only provide statistical correlations between the values of X and Y, RCT studies, as concrete experimental

processes, include many other components from which researchers can gather important causal information. In fact, researchers have often recorded supplementary data during experiments for future reference, such as researchers in the STAR project recorded teacher salary, expenditures, student locations, etc. (see figure 3).

Figure 3: Partial table of contents of the STAR report (Word et al. 1990). This figure shows that researchers in the STAR project recorded data beyond the manipulated variable class size and the target variable student performance. For example, Table 1-3 included the data about teacher salaries and expenditures.

Cartwright and her colleagues might reply that this is an enormous amount of data which has no specific use, and researchers lack expertise to organize such data into applicable information. Thus, practitioners should be the ones who find the data they need from the reports and utilize them. Admittedly, researchers cannot include information about all factors, or even all the causally relevant factors. They may also not have a clear idea on when and how to use this data. However, this act of recording data suggests that there are other things that researchers can do in RCT

studies besides waiting and analyzing data that pertains only to X and Y. Once this is realized, the next step will follow, which is to find a feasible way to organize the information. I will then present one possible way in the light of the theory of interventionism, which is similar to the input-output logic of RCT studies, and thus, is a useful tool to help understand the causal relationships established by RCT studies.

Causal Stability and More Stable RCT Results

The theory of interventionism, as introduced in Section 2.1, aims to determine the causal relationship between X and Y under the condition that following the manipulation of X the value of Y changes and that no confounding factors influence the process. This theory has a clear similarity to RCT studies. Both of them concentrate on the fact that X and Y are correlated without providing information on how they are correlated. This fact, from an interventionist perspective, does not entail that one would remain completely oblivious of information related to "X→Y". The concept of causal stability suggested by Woodward (2010) is helpful to explain what information one can obtain. Causal stability is the degree to which the causal relationship would hold under various background conditions. According to the concept, researchers should pay attention to the background conditions of the relationships. Background conditions are a set of states of affairs related to the

causal claim "X→Y", but whose information is not explicitly mentioned in X or Y. For example, in the claim "class size→student performance", factors such as teacher quality and course schedule are not implied by the two variables in the claim, but they can influence the causal relationship and are the background conditions of this relationship. By collecting such factors, researchers can gain useful causal information about what may have an impact on the relationships established by RCT results.

The idea of background condition does not seem to add something special compared to the idea of causal principles proposed by Cartwright and her colleagues in that scientists are required to gather information about factors that have the potential to influence the given causal relationships in both situations. The difference is that the interventionist perspective expects RCT researchers to help local practitioners gather information, while Cartwright and her colleagues believe that researchers can only provide experimental outcomes in the form of "X→Y", leaving the task completely to practitioners. "X→Y", as suggested by Deaton and Cartwright (2018), is not a reliable claim since its extrapolations are vulnerable to failure. They argue that researchers cannot increase the claim's reliability as mentioned in Section 3.1. In this situation, the focus of research on how to use the claims effectively shifts from the results themselves to the local contexts in which the results are implemented. However, I argue that RCT results can be more useful

even though they are not reliable in the form of "X→Y". This is because the stability of causal claims can be adjusted by adding or removing some background conditions of the claims. There is a possibility that increased stability increases the usefulness of the causal claims. To understand this possibility, the concept of causal stability deserves further explanation. This concept indicates that causal claims do not typically hold under all background conditions. Some causal relationships hold in a very restricted set of circumstances and are relatively unstable, while some others hold under a large number of possible background conditions and are relatively stable. As a result, the more stable a causal claim is, the more backgrounds it can hold in, and the more applicable it is. In this sense, it is theoretically possible for researchers to gather information about background conditions and provide more causally stable and applicable RCT results to increase the possibility of successful extrapolations.

How the Interventionist Alternative Can Help

This interventionist alternative to Cartwright and her colleagues' approach, which assigns part of the work of gathering causal information to researchers, can help address the problem of extrapolation more effectively. Indeed, researchers work in experimental settings and cannot take away from practitioners the work of

identifying specific local influencing factors. They can, however, reduce the burden on practitioners by focusing practitioners' efforts on local contexts.

In Carwright and her colleagues' approach, practitioners have to collect information about different causal principles. A potential implication is that practitioners should have a complete understanding of both experimental and local contexts individually. They will first analyze the RCT reports thoroughly to find factors that may have a role in experiments. Second, they will delve into local environments to find specific factors that may influence the relationships. If, as the interventionist approach claims above, researchers can identify potential influencing causal factors in an organized way, then practitioners can put less effort into learning the original experimental contexts. It is effective since researchers are much more familiar with the original experimental processes and can provide relevant information without additional study. Further, if researchers can provide reports on the stability of RCT results by analyzing different background conditions, practitioners can recognize which kind of factors is essential to the background conditions of RCT results, and thus investigate local contexts purposefully easing the workload.

The concept of causal stability helps explain how researchers can organize the data collected during experiments into useful causal information. The general guide is to organize them into different background conditions under which causal claims

established in RCT studies will hold or fail. Researchers can try their best to find factors that influence relationships in experimental settings. However, they cannot predict any specific factors that may work in local settings. A noteworthy thing is that my interventionist approach does not expect researchers to find a complete list of specific influencing factors in both settings. Instead, it aims to highlight some kinds of factors that have special roles in the causal process. Realizing these causal roles can help researchers understand how specific factors, i.e., important background conditions, may affect the causal relationships.

A potential objection to my approach may be that the mentioned three kinds of information in Section 3.1: structure, support factors, and derailers, can also serve to organize researchers' data as each of them possesses a special causal role. In this case, the interventionist approach does not need to provide new kinds of causal information. However, these three kinds of information are more useful in the situation where practitioners have identified numerous local factors and want to analyze how these factors might play a role in extrapolations. Local factors would be categorized into three kinds and adjusted accordingly. On the contrary, my hope is that researchers can find factors that have potential influences based on different causal roles. This is to say, they would first understand potential causal roles that different factors may play, and second check if the established causal claims leave a space for these roles. If so, researchers can search for specific factors that play the

roles and alert practitioners about this. Information proposed by Cartwright and her colleagues provides general causal roles and cannot help researchers identify factors before extrapolations. In the next section, I will suggest, as an example, two kinds of factors that have more specific causal roles compared to those of Cartwright and her colleagues. The analysis of them can imply how they underlie, support, or impede the causal relationships, and thus provides a feasible guide for researchers to gather useful causal information.

3.3 An Interventionist Response to the Problem of Extrapolating Results from RCTs

In Section 3.2, I have argued that researchers have the potential to organize data collected in RCT studies according to their different causal roles. This can help identify important background conditions of the causal claims and help establish more stable claims. As a result, practitioners can focus on gathering information about local settings and apply the claims more effectively. This section aims to provide two kinds of factors that play various roles in the causal process as a

reference for researchers.[10] I will use two sets of interventionist-style graphs, to visualize and explain these factors. The first set of graphs identifies the confounding factors, and the second set identifies the intermediate factors.

Confounding Factors

The first kind of factor that may influence the causal process is the confounding factors. As mentioned in Section 2.1, confounders are factors that have an impact on the effect Y, and this impact may obscure the precise causal contribution of X to Y. If this kind of factor are noticed before experiments, it can be set as an object of control in experimental designs to minimize their impact. Figure 4 illustrates a scenario where the influence of confounder C_1 is controlled, and the experiment successfully observe an independent causal relationship between X and Y. However, although researchers attempt to rule out the influence of all confounders by experimental manipulations, it is possible for some confounders to be overlooked. Imagine a class size reduction project, in which X represents class size, Y is student performance, and the original C_1, teacher quality, is randomly assigned and thus will not influence the relationship between X and Y. However, as shown in figure

[10] I will only discuss the cases in which researchers have successfully established causal relationships between X and Y in experiments. This is to say, there must be an arrow from X to Y. RCT results are not obtained by coincidence or mistake.

5, a new C_2, online teaching materials, might have an impact on Y along with X. If researchers are not careful to control for this factor, X's effect on student performance might be exaggerated in the circumstance that students in small classes happen to have access to these materials and students in regular classes do not. In this case, the new factor C_2 will contaminate the independent causal connection "class size→student performance", weakening the claim that "class size reduction is the only cause of better student grades".

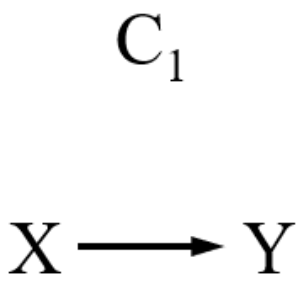

Figure 4: A controlled confounder in the manipulation. C_1 represents a possibly confounding property that is controlled by the experiment. The graph presents a situation where the relationship between X and Y is not influenced by C_1. Researchers find possibly confounding properties that they think are important, control these factors in the experiment, and assume that no other factors will have an impact on the relationship between X and Y. Researchers consider the manipulation of X is the sole cause of the change in Y under this condition.

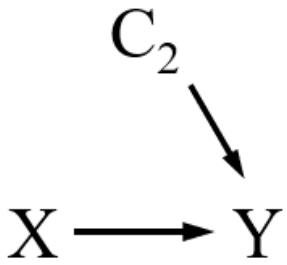

Figure 5: A confounder that is not controlled in the manipulation. C_2 represents a possibly confounding property that is not controlled by the experiment. The graph presents a situation where C_2 will have an impact on Y and then contaminate the causal connection between X and Y. This situation can occur if researchers fail to identify some important influencing factors and do not control these factors in the experiment.

This risk posed by new confounders can be mitigated. Researchers can observe carefully and record every suspected factor during or after experiments and add the factors to the list of factors needing control. Conducting surveys is a good way to check whether any important factors are missing. For example, the survey can simply ask students and teachers in small classes: "what do you think contributes to your grades?" Some students may answer "online teaching materials". A follow-up survey can then ask: "to what extent do you think online teaching materials contribute to your grades?" If a large number of positive answers are observed in this case, this factor deserves to be controlled in extrapolations.

However, a new risk that is much less obvious and harder to avoid may emerge during experiments. Figure 6 demonstrates the causal process in which this problem may arise. The experiment aims to manipulate X, but the implementation

unexpectedly changes the value of C* at the same time, which means that the manipulation M creates a new confounder that would not appear without the manipulation.[11] Imagine a case where X is class size and Y is student performance. The manipulation changes the value of C*, student expectation. Students may compare their class size with peers' and find that they are put in a different situation that their class size is smaller. They may consider themselves a part of a revolution with this observation, which will lead to a higher expectation regarding academic performance. They will think that they can gain better grades naturally in this situation. This mentality will positively influence their grades. Therefore, their better grades partly derive from their higher expectations. This additional factor may lead to a problem if manipulation does not produce higher expectations in different local contexts. For example, if a local institution usually implements bad policy which does not benefit the population, students may have a lower expectation of their grades due to the manipulation. In this case, student performance may not be as good as claimed by the previous RCT results, and the extrapolation fails accordingly.

[11] This idea is inspired by Deaton and Cartwright's 2018 paper, in which they propose that randomization can change the original settings and that new factors requiring randomization will appear in new settings.

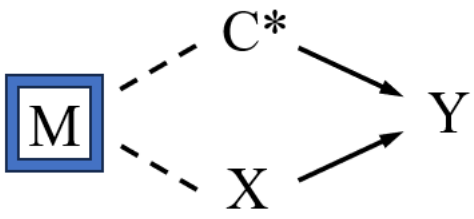

Figure 6: A new confounder created in the manipulation. M in the blue box represents the implementation of the manipulation, C* represents a confounding property, and the dotted lines represent the process of manipulating C* and Y. The graph presents a scenario where the implementation of the manipulation of X unexpectedly changes the value of C* and thus means that M is not an intervention on X with respect to Y.

Such extra confounders are hard to observe and control at the beginning of the experiments. Researchers may also conduct surveys and invite relevant experts to identify the confounders created by manipulations and recognize how they work for the reference of practitioners. For example, invited educators may inform researchers that the expectation is a common influencing factor in educational experiments. Based on this notification, researchers can label the expectation as an important factor to be observed and include in the survey a question such as “do you think you would do better in a small or regular class?”. Being able to recognize different confounders allows researchers to better understand RCT results and provides useful causal information for practitioners.

Intermediate Factors

The second kind of factor that plays a role in the causal relationship "X→Y" consist of intermediate factors. Figure 7 illustrates a scenario where an intermediate factor Z serves as an intermediate between X and Y, in which case the manipulation of X has an impact on Z first, and then the change in Z results in the change in Y.[12] This means that a causal process from X to Y can be realized only through the involvement of this intermediate factor Z. Intermediate factors are difficult to notice during experiments since they may happen to remain involved in the research process under the condition that experimental settings seldom change. Researchers prefer to believe that no unknown factors would influence causal relationships in experiments since they succeed every time. However, in different background settings, intermediate factors may not work in the same way. If X cannot make a difference in Z in a similar manner, or it does but Z fails to influence Y similarly, the causal claim "X→Y" would fail to hold.

Figure 7: An intermediate factor. Z represents the property of being intermediate between X and Y. This graph presents the case that the manipulation of X has an impact on Z first, and then the change in Z results in the change in Y.

[12] This is a causal structure discussed in Woodward's 2003 book.

For example, the government of Hangzhou, China, delayed primary school start time by half an hour to 8:30 a.m. to improve student sleeping hours (Qianjiang Evening News, 2011). However, this reform did not achieve the expected result given that most parents must send their children to school by 8:00 a.m. in order to arrive at their workplace on time. In this example, the causal claim that delaying X "school start time" influences Y "student sleeping hours" can only maintain with the presence of an intermediate factor Z "student arrival time". Since students still arrived at school early before 8:00, instead of 8:30 as expected, their sleeping hours did not change even upon the manipulation of school start time. Figure 8 shows the case where X fails to affect Y because it fails to change the value of Z.

Figure 8: A manipulation without the intermediate factor. This graph does not have the arrow from X to Z. This presents a situation where the manipulation of X fails to change the value of Z resulting in the failure of affecting Y. This situation can occur when a causal process from X to Y can be realized only through the involvement of the intermediate factor Z.

Intermediate factors can affect causal process in a more complicated way. Figure 9 presents the situation where two intermediate factors Z_1 and Z_2 co-influence the causal process between X and Y. The change in Z_1 produces a positive change in Y, while the change in Z_2 produces a negative change in Y. Woodward

(2003) calls this situation as "combined interventions". Imagine a special situation that may arise in the class size reduction project, in which class size reduction respectively affects the performance of low-performing and high-performing students.[13] Low-performing students get better grades in small classes, while high-performing students get lower grades. These two impacts are combined to influence the achievement of all students. If the improvement in the performance of low performers is large enough, then the project's overall effect will be positive, even if some students are harmed instead of benefited. In this case, there is indeed a causal relationship between reducing X "class size" and improving Y "student performance" derived from a combined effect of Z_1's positive effect and Z_2's negative effect. The successful establishment of this causal relationship depends on a specific proportion of low-performing students' performance to high-performing ones and relative numbers of these two groups of students. In applications, this proportion may be different so that the overall effect turns out to be negative and the extrapolation fails.

[13] This case is transformed from an example from Joyce and Cartwright's 2018 paper.

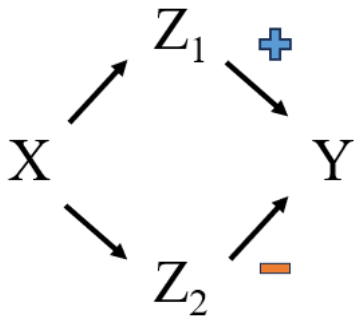

Figure 9: A combined manipulation. Z_1 and Z_2 represent the property of being two specific intermediate factors, the plus sign represents that an intervention on Z_1 with respect to Y that increases the value of Z_1 will increase the value of Y, and the minus sign represents that an intervention on Z_2 with respect to Y that increases the value of Z_2 will decrease the value of Y. This graph presents the situation where there are two intermediates between X and Y. The impact of the manipulation of X on Y combines the two effects of the two intermediates.

The negative influence of intermediate factors can also be avoided by conducting surveys or referring to experts. To observe intermediate factors, researchers have to know what is changed by manipulations. Researchers, along with the invited experts, should be keenly aware of various changes that occur during experiments. For example, experts who are sufficiently sensitive may notice that an increase in student sleep time requires a change in student arrival time. Researchers can design better experiment based on this observation, allowing the experimental manipulation to change student arrival time under more background conditions. Similarly, researchers can have better experimental designs if experts notice that reducing class size has a different effect on the performance of low-

performing and high-performing students under some specific background conditions.

One may object to this idea as expecting too much from researchers and experts, who often have difficulty finding intermediate factors before the extrapolations fail. Admittedly, it is very challenging to find factors before they have a negative impact. My approach is trying to raise the sensitivity of researchers and experts in experiments by bringing their attention to factors that may be playing a causal role. Further, although researchers may not be able to identify the specific factors during experiments, post-experimental surveys can help them gather relevant information before extrapolations. With experts help to provide optional answers, researchers can give a multiple-choice question such as "what do you think [the manipulation] has changed?" An open-ended section in the question that allows participants to offer their own answers is also useful. For instance, in response to the question, a large number of the answer "better grades" may be missing from high-performing students. Researchers would be aware of the effect of combined interventions in experiments. In this way, useful causal information can be provided to practitioners.

Conclusion

The problem of extrapolating results from RCTs is that RCT results cannot be applied from the study context to local contexts. This problem can be explained by three limitations of RCT studies. They are the input-output logic, imperfect randomization, and the vulnerability to value-laden contexts. Despite these limitations and the problem of extrapolation, RCTs are still recognized as a "gold standard" for obtaining reliable causal claims, which put RCTs in a special position in educational research. In order to better use RCT results, Cartwright and her colleagues suggest having practitioners gather information about causal principles to fill the gap between experimental and local contexts. In particular, they argue that RCT studies cannot provide helpful information about causal principles for extrapolations because RCTs run in an input-output logic and cannot explain their results.

However, once RCT researchers realize that causal claims can be evaluated at the level of stability, they are able to give useful causal information about the claims, i.e., the background conditions under which the claims may hold or fail. In this case,

the burden on local practitioners can be reduced as they can purposefully focus on collecting information in local contexts. The causal graphs in this thesis, which clarify the causal roles of confounding factors and intermediate factors, provide one way to investigate possible background conditions for RCT results. These graphs cannot be directly used to illustrate how RCT studies could establish more applicable results, but they are intended to point to the possibility that RCT researchers have the potential to provide causal information more than simple causal claims. Aware of this possibility, researchers may find other methods that are more effective at generating stable causal claims than my approach.

References

1. Becker, Bronwyn E., and Suniya S. Luthar. 2002. “Social-emotional factors affecting achievement outcomes among disadvantaged students: Closing the achievement gap.” Educational psychologist 37.4: 197–214.
2. Bohrnstedt, George W., and Brian M. Stecher. 2002. “What We Have Learned about Class Size Reduction in California.” California Department of Education.
3. Campbell, John. 2013. “Causation and Mechanisms in Psychiatry.” In *Oxford Handbook of Philosophy and Psychiatry*, ed. Kenneth William Musgrave Fulford et al., 935–949.
4. Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-based policy: A practical guide to doing it better*. Oxford.
5. Connolly, Paul, et al. 2017. Using Randomised Controlled Trials in Education. Sage.
6. Deaton, Angus, and Nancy Cartwright. 2018. “Understanding and misunderstanding randomized controlled trials.” *Social science & medicine* 210: 2–21.

7. Fuller, Jonathan. 2021. “The myth and fallacy of simple extrapolation in medicine.” *Synthese* 198: 2919–2939.

8. Giere, Ronald N. 1979. *Understanding Scientific Reasoning (1st Edition)*. Holt, Rinehart, and Winston.

9. Giere, Ronald N. 1979. *Understanding Scientific Reasoning (2nd Edition)*. Holt, Rinehart, and Winston.

10. Glass, Gene V., et al. 1982. “School class size: Research and policy.” Sage.

11. Joyce, Kathryn E., and Nancy Cartwright. 2018. “Meeting our standards for educational justice: Doing our best with the evidence.” *Theory and Research in Education* 16.1: 3–22.

12. Joyce, Kathryn E. 2019. “The key role of representativeness in evidence-based education.” Educational Research and Evaluation 25.1–2: 43–62.

13. Joyce, Kathryn E., and Nancy Cartwright. 2020. “Bridging the Gap Between Research and Practice: Predicting What Will Work Locally.” *American Educational Research Journal* 57.3: 1045–1082.

14. Kvernbekk, Tone. 2016. *Evidence-based practice in education: Functions of evidence and causal presuppositions*. Routledge.

15. Lewis, David. 1973. “Causation.” *Journal of Philosophy* 70: 556–67.

16. Mertens, Donna M. 2023. *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Sage.

17. Morrison, Keith. 2019. “Realizing the promises of replication studies in education.” *Educational Research and Evaluation* 25.7-8: 412–441.

18. Nurses’ Health Study. “History.” Nurses’ Health Study. Accessed June 13, 2023. https://nurseshealthstudy.org/about-nhs/history.

19. Qianjiang Evening News. “Hangzhou Proposed to postpone the elementary school start time.” China Daily. August 26, 2011. https://global.chinadaily.com.cn/dfpd/zj/2011-08/26/content_13199575.html

20. Ross, Lauren N., and James F. Woodward. 2016. “Koch’s postulates: an interventionist perspective.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 59: 35–46.

21. Woodward, James. 2003. *Making things happen*. Oxford.

22. Woodward, James. 2010. “Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation.” *Biology and Philosophy* 25.3: 237–318.

23. Word, Elizabeth., et al. 1990. “The State of Tennessee’s student/teacher achievement ratio (STAR) Project: Technical Report (1985-1990).” Tennessee State Department of Education.