Laboratorio Interdisciplinario de Estadística Social

Discussion Paper Nº2017|02

# On the Implementation of Integrated Leadership Model as Public Policy: A Critical Analysis

**Trinidad González Larrondo and Ernesto San Martín**

**Facultad de Matemáticas**
**Pontificia Universidad Católica de Chile**
Av. Vicuña Mackenna 4860, Macul
**lies.mat.uc.cl/**

Laboratorio
Interdisciplinario
de Estadística Social

**To cite this paper**:
González, T., and San Martín, T. (2017). On the Implementation of Integrated Leadership Model as Public Policy: A Critical Analysis. *LIES Discussion Paper* Nº2017|02.

# On the Implementation of Integrated Leadership Model as Public Policy: A Critical Analysis

TRINIDAD GONZÁLEZ[a] AND ERNESTO SAN MARTÍN[b,c]

[a]*School of Psychology, Pontificia Universidad Católica de Chile, Chile*

[b]*Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile*

[c]*The Economics School of Louvain, Universit Catholique de Louvain, Belgium*

May 11, 2017

**Abstract**

*In the context of the of the implementation of the Chilean National System of Quality Assurance of Education, the Agency of Quality of Education applied a national survey to collect information about opinions and perceptions principals and teachers have on this system. The Agency of Quality particularly inquired if intra-schools decisions leading to implement pedagogical and curricular changes are based on the analysis of the information on the school performance provided by the Agency of Quality to each Chilean school. The Agency of Quality promotes the use of such an information in line with the integrated leadership model. The survey showed that when such leadership practices are implemented, 71% of public Chilean schools report curricular and pedagogical changes, whereas the 28% of public schools where those leadership practices are absent report those changes. Similar results are found for subsidized schools. Based on these results, it seems coherent to promote as a public policy that principals and teachers adequate their intra-school practices to the integrated leadership model. The relevant question is, What would be the percentage of public and subsidized schools performing curricular and pedagogical changes if both principals and teachers were to modify their intra-school practices in line with integrated leadership? This question cannot be answered using the empirical evidence alone because of the fundamental problem of causal inference. However, it is possible to offer an impact evaluation if non empirical hypotheses are assumed. In this paper, we argue that such evaluation should include two dimensions: first, the evaluation should be done with respect to different scenarios, which are characterized by non-empirical assumptions leading to handle at some extent the fundamental problem of causal inference. Second, the impact*

*evaluation should provide some action lines coherent with the non-empirical assumptions underlying the evaluation itself. We show that opposite policy recommendations are obtained, one suggesting the implementation of integrated leadership practices, and the other not. Both recommendations are related to different political perspective about the relationships between schools and the State: the scientific findings do not support such perspectives, but only provide information enough to justify different behaviors of the policy maker.*

Key words: *Inductive Behavior; Partial Identification; Collaborative Instructional Leadership; Transformational Leadership; School Improvement.*

# 1 Introduction

Most of the public educational policies are motivated by what it is typically called *empirical evidence*. Broadly speaking, the empirical evidence consists in reporting the effect of a treatment on a specific outcome, which in turn is compared to the effect of an alternative treatment on the same outcome, or even the effect of the absence of the treatment. In order to fix ideas, let us consider the case study analyzed in this paper: Since 2011, Chilean schools are monitored by the National Agency of Quality of Education; as a result of the continuous monitoring, schools receive information regarding the performance of both students and the school itself. As a reaction to such an information, it is expected that schools implement (if necessary) changes in order to improve the quality of education. Due to the internal organization of schools, such changes should be leaded by their principals in collaboration with teachers. It is, therefore, expected by the National Agency of Quality of Education that such leadership is in line with both collaborative instructional leadership and transformational leadership. When such leadership practices are implemented, 71% of public Chilean schools report curricular and pedagogical changes –changes that are consistent with school improvement, whereas the 28% of public schools where those leadership practices are absent report curricular and pedagogical changes (DESUC-CEPPE, 2015). Based on this empirical evidence, a possible educational public policy is the implementation of collaborative instructional and transformational leadership models in schools.

If this were the case, a relevant question from a policy perspective is the following: What would be the percentage of public Chilean schools performing curricular and pedagogical changes if both principals and teachers were to modify their practices in line with collaborative instructional and transformational leadership? This question call for extrapolations and therefore cannot be answered using the empirical evidence alone. The empirical evidence is not enough to take the social decision of using the treatment

(in our case, collaborative instructional and transformational leadership practices) or not. This is due to the fundamental problem of causal inference, namely that it is impossible to observe the outcomes of a same school under the treatment and under its absence (Holland, 1986). However, in order to ensure the welfare of the society, it is necessary to compare both "fictitious worlds" and to choose the one in which the welfare is maximized (Manski, 1996; Angrist & Pischke, 2008). This is precisely the criterion leading to chose between the implementation of a policy (the "universal" application of the treatment), or not.

It is well known that if the empirical evidence is combined to a particular distributional assumption, then the fundamental problem of causal inference can be avoided. Consequently, In this case, it is possible to perform the evaluation of the impact of a public policy. This distributional assumption, known as *exogenous switching condition* (Maddala, 1983) or *strong ignorability condition* (Rosenbaum & Rubin, 1983), asserts statistical independence of the realized treatments and the outcomes, conditional on a set of covariates. This condition mimics the definition of a randomized experiment (Fisher, 1935; Cochran & Chambers, 1965; Angrist & Pischke, 2008), but it is not empirically refutable (Manski, 1995, 2007): it can only be justified in specific applications, the justification essentially depending on specific covariates. Now, suppose there is a concrete policy maker that could perform policy actions in a legal context, in particular the implementation of the treatment under discussion. Suppose moreover that an impact evaluation based on a strong ignorability condition is available to the policy maker. The question we want to emphasize is the following: to what extent the ignorability condition determines the policy maker behavior with respect to the implementation of the treatment?

We argue that the (ex-ante) impact evaluation of a public policy is a combination of empirical evidence and non-empirical hypotheses. These hypotheses determine the behavior of the policy maker during the eventual implementation of the treatment. Thus, the first dimension of an impact evaluation consists in providing evaluations under different scenarios, which are characterized by non-empirical hypotheses leading to handle at some extent the fundamental problem of causal inference. Specifically, two scenarios will be explored in this paper: the first scenario is summarized by saying that "to apply the treatment is always better than don't apply it" (technically speaking, *ordered outcomes*); the second scenario is summarized by saying that "the realized treatment reflects the actual relationships between principals and teachers" (technically speaking, *selection of the treatment with the larger/smaller outcome*). The second dimension of an impact evaluation should provide to the policy maker some action lines coherent with those non-empirical hypotheses. We argue that such a "behavior" corresponds to the concept of *inductive behavior* as introduced in statistics by Neyman (1957).

We develop these ideas in the concrete case of the implementation of the monitoring system of Chilean schools by a specific policy maker, The National Agency of Quality of Education. Both schools and the National Agency of Quality of Education regulate their relationships and decisions by the Law 20.529 *National System of Quality Assurance of Education* (*Sistema Nacional de Aseguramiento de la Calidad*).

Due to the high stakes consequences of the new educational system on the schools, as well as to the complexity of the new regulatory framework, the Chilean Government decided to start in 2015 the implementation of the continuous monitoring system as a trial period. Three were the main objectives of such a period:

1. To collect information about the degree of knowledge that schools have of both the National System of Quality Assurance of Education and the functions of the National Agency of Quality of Education.

2. To collect information about the perceptions and opinions that school communities have about the monitoring system of the schools, in particular on the usability of the information regarding the performance level obtained by the schools.

3. Improve several aspects of the monitoring system, as for instance the information system that the National Agency of Quality of Education generates, or statistical issues regarding the methodology of school classification.

This type of information was collected through a *National Survey on the Pilot Implementation of the Schools Monitoring System* applied in 2015 by the National Agency of Quality of Education (in what follows, the Agency of Quality). A relevant aspect the Agency of Quality promotes is how school leadership determines school changes in order to improve students learning. The National Survey on the Pilot Implementation of the Schools Monitoring System accordingly includes questions leading to understand to what extent the information produced by the Agency of Quality promotes both school leadership and pedagogical and curricular changes. The empirical evidence based on these questions suggests that the combination between the information produced by the Agency of Quality and the models of collaborative instructional and transformational leaderships, entail pedagogical and curricular changes in a relatively high proportion of schools. This leads to propose as a public policy that principals and teachers adequate their intra-schools actions to those school leadership models. In this paper, we evaluate the impact of this public policy considering the two scenarios mentioned above, making explicit the corresponding inductive behaviors.

This paper is organized as follows: Section 2 describes the Chilean National System of Quality Assurance of Education, focusing its attention on three axes: the Chilean monitoring system; the leadership school standards and their relationships to a combination of two leadership models, namely collaborative instructional and transformational (which is called *integrated leadership*); and the structure of the information the Agency of Quality provides to schools. The National Survey on the Pilot Implementation of the Schools Monitoring System, along with the corresponding empirical evidence, is discussed in Section 3; here it is emphasized the role of a theoretical framework in the statistical interpretation of the empirical evidence. Section 4 performs the impact evaluation under two scenarios. The paper ends by a general discussion.

# 2 The Chilean National System of Quality Assurance of Education

After five years of parliamentary discussion, the Chilean government approved in 2011 the Law 20.529 on the *National System of Quality Assurance of Education*. This law regulates the Chilean educational system by means of a continuous monitoring system. Its main objective is to achieve a continuous improvement of students learning. A basic condition is required, namely to increase the educational capabilities of principals and teachers (Law 20.529, Art. 2). It is accordingly needed to evaluate the achievement of the learning standards as well as to promote standards on the performance of both teachers and principals. This in turn requires not only to provide public information to schools on their performance at different dimensions of quality, but also support them in their functions related to the continuous improvement (Ley 20.529, Art. 3). Each of these axes –measuring the quality of a school, promoting school leadership standards, and providing to schools information and guide– is described in some detail below.

## 2.1 The multidimensional character of quality of education

According to the Law 20.529, the role of the Agency of Quality is to conduct such a monitoring system (Art. 38). The most essential part of this system is the school performance categorization. The School Performance Categorization Methodology leads to catalogues schools in four performance categories (high, medium, medium-low and insufficient). It is based on four dimensions of quality (Law 20.529, Art. 18): the first of these comprises learning standards (67%), which indicate the level of learning achieved by students. The second includes progress measures (3.3%), which indicate an improvement in

results over time. The third area is SIMCE scores (3.3%), where the SIMCE is a national standardized test in Mathematics and Language, plus other disciplines when correspond (for details on the SIMCE test, see Meckes & Carrasco, 2010; Manzi & Preiss, 2013). The final dimension is related to Indicators of Personal and Social Development (26.4%). More specifically,

1. The learning standards comprise three levels of learning according to performance on the SIMCE tests: sufficient level of learning, elementary level of learning and insufficient level of learning. The learning standards, along with their cut-off scores, are set by the Ministry of Education using the SIMCE scores obtained by the students and a bookmark standard setting procedure (ACE, 2016a).

2. Progress measures refers to the SIMCE trend indicator that identifies changes in the school's SIMCE test results, defining whether they have increased, decreased or remained constant over recent years.

3. The SIMCE score refers to the average scores obtained by schools in each subject (Language, Mathematics and other disciplines when correspond) and at every grade level that is assessed using SIMCE tests.

4. Indicators of Personal and Social Development are constructed from the information gathered using the questionnaires applied by the Agency of Quality to students, teachers and parents/guardians. These indicators are the following:

   (a) The indicator for academic self-esteem and school motivation takes into consideration self-perception and self-assessment by students regarding their ability to learn. It also takes into consideration student perceptions and attitudes towards learning and academic achievement.

   (b) The indicator for school environment takes into consideration the views and attitudes of the students, teachers and parents/guardians regarding the presence of an environment of respect, organization and safety within the school.

   (c) The indicator for participation and civic education takes into account student attitudes regarding their school; student and parent/guardian perceptions regarding the degree to which the school encourages participation and commitment from members of the educational community; and student perceptions regarding the degree to which democratic life is encouraged.

(d) The indicator for healthy lifestyle habits assesses self-declared student attitudes and behavior regarding a healthy lifestyle, as well as their perceptions regarding the degree to which the school encourages habits that are beneficial to the students health.

(e) The school attendance indicator takes into consideration the distribution of students according to four categories that are developed based on the number of days that a student attends school, in relation to the official number of school days in a year.

(f) The school retention indicator takes into consideration the schools ability to ensure that its students remain within the formal education system. It is calculated by taking into account all of the students from a school, and not just those that sit SIMCE tests, which is the case for the other indicators.

(g) The gender equality indicator assesses the equitable achievement of learning results obtained by girls and boys in co-ed schools

Once an unadjusted index is constructed based on the previous indicators, the index is adjusted by socio-economic characteristics of the students. This index is used to classify the schools in the four performance levels mentioned above; for details, see MINEDUC (2014). It should be mentioned that if a school classified as insufficient does not improve its performance level after four years, it risks to be closed (Law 20.529 Art. 11).

## 2.2 Integrated leadership model in the Chilean policy context

As it was mentioned before, the improvement of students learning requires to increase the educational capabilities of principals and teachers. From an educational theoretical perspective, this requires to disseminate and apply a series of initiatives to improve the school leadership. In Chile, these initiatives started in 2004 with the Law 19.979 that established that the main function of the principals of schools is to "direct and lead the school mission and goals" (Art. 5). One year after, the Ministry of Education published a set of criteria leading to evaluate the professional development and performance of both principals and teachers (MINEDUC, 2005). On these initiatives, the Ministry of Education of the Chilean government defined in 2015 standards on school leadership (MINEDUC, 2015), which constitute the general framework used by the Agency of Quality to improve the capabilities of the members of the school community.

The relevance of those standards to educational policies was explicitly justified by the Chilean edu-

cational authorities on the well-known impact that school leadership has in terms of effectiveness and improvement, as well as in terms of alignment with the tendency to decentralize, the attainment of high levels of self-sufficiency and accountability in the school system. Following the literature of school leadership, it was argued that leadership entails influencing ones colleagues to act in ways likely to help accomplish the short-term goals and long-term directions considered desirable for the school. Although its effects on pupils are mostly indirect, it was retained the fact that such effects are mediated by teachers and, consequently, are relevant in the context of the National System of Quality Assurance of Education. For details, see, among many others, Bush and Glover (2003), Waters, Marzano, and McNulty (2003), Leithwood, Day, Sammons, Harris, and Hopkins (2006), Barber and Mourshed (2007), Pont, Nusche, and Moorman (2008), Seashore, Leithwood, Wahlstrom, and Anderson (2010), Horn and Marfán (2010) and Weinstein and Hernández (2014).

The Chilean standards on school leadership are organized in five dimensions (see MINEDUC, 2015, Section 4.1):

1. Develop and implement a shared vision on the mission and goals of the school.
2. To improve and empowered teachers and other staff members abilities and leadership.
3. Coordinating, monitoring, and evaluating curriculum, instruction and assessment.
4. Promote inclusive and collaborative organizational culture, stimulating community relationships among school members.
5. Manage an effective organizational performance ensuring school resources, disseminating results and defining roles of the school members.

Dimensions 1, 3 and 5 are mostly related to collaborative instructional leadership, whereas dimensions 2 and 4 are more related to the transformational leadership model. As a matter of fact, in the collaborative instructional model the principal is oriented to educational development looking for conditions that directly impact on the quality of curriculum and instruction, considering the opinion, perspective and work of teachers. This requires to develop and implement practices leading to coordinate the members of an organization (the school). By so doing, it is expected to mobilize them to the search of a set of common objectives (Robinson, Lloyd, & Rowe, 2008; Hallinger, 2003).

Following Hallinger (2003), the specific practices of collaborative instructional model can be organized in three dimension: 1) Defining the school's mission, which involves two functions: framing the schools

goals and communicating the schools goals. These functions concern the principals role in working with staff to ensure that the school has clear, measurable goals that are focused on the academic progress of its students. It is the principal's responsibility to ensure that these goals are widely known and supported throughout the school community. This dimension assumes that the principals responsibility is to ensure that the school has a clear academic mission and to communicate it to staff. 2) Managing the instructional program; it means to focus on the coordination and control of instruction and curriculum. This dimension incorporates three leadership functions: supervising and evaluating instruction, coordinating the curriculum, monitoring student progress. These functions require the leader to be deeply engaged in the schools instructional development. This framework assumes that development of the academic core of the school is a key leadership responsibility of the principal. 3) Promoting a positive school learning climate, includes several functions: protecting instructional time, promoting professional development, maintaining high visibility, providing incentives for teachers, providing incentives for learning. This dimension conforms to the notion that effective schools create an academic press through the development of high standards and expectations and a culture of continuous improvement. It is the responsibility of the instructional leadership to align the schools standards and practices with its mission and to create a climate that supports teaching and learning.

The transformational leadership model provides intellectual direction and aims at innovating within the organization, while empowering and supporting teachers as partners in decision making (Marks & Printy, 2003). Transformational leaders motivate followers by raising their consciousness about the importance of organizational goals and by inspiring them to transcend their own self-interest for the sake of the organization. In their relationships with followers, transformational leaders exhibit at least one of these leadership factors: idealized influence, inspirational motivation, intellectual stimulation, and individualized consideration. In the school context, the effective functions of transformational leadership that has been identified are the following: (a) mission centered (developing a widely shared vision for the school, building consensus about school goals and priorities), (b) performance centered (holding high performance expectations, providing individualized support, supplying intellectual stimulation), and (c) culture centered (modeling organizational values, strengthening productive school culture, building collaborative cultures, and creating structures for participation in school decisions). For details, see Leithwood et al. (2006) and Leithwood and Jantzi (2006).

Summarizing, in the Chilean policy context, both collaborative instructional leadership and transformational leadership operate "in tandem" which, according to Marks and Printy (2003)'s terminology, is called *integrated leadership*.

### 2.3 Information and guidance: the inputs provided to schools by the Agency of Quality

The Agency of Quality promotes the school leadership standards by means of a report of results which communicates to each school its performance category level obtained according to the School Performance Categorization Methodology (ACE, 2016b). The report is organized in such a way that is useful to facilitate and orientate principals' and teachers' actions regarding school improvement. As a matter of fact, the performance category level obtained by a school is explained in terms of the expected achievements attained by the students in each of the 10 indicators used to construct such a category level (for the construction, see Section 2.1). Associated to this information, the Agency of Quality offers three types of supports: (a) online resources that orientate the analysis of the results gathered in the report in such a way that the resulting reflections help to take intra-school decisions related to school improvement, particularly curricular and pedagogical changes; (b) progressive evaluation, a new voluntary and self-applied evaluation that intends to provide information on students progress in Language at second grade of primary school (this is the only grade at which the test is available); (c) learning visits that, thanks to the active participation of the school community, allow to identify practices significant for the development of students as well as to the institutional improvement.

These supports are aligned with the integrated leadership model. It should be mentioned that this alignment is emphasized by reporting the positive marginal effects of leadership and teacher feedback on the learning results in Language and Mathematics as measured by the SIMCE test. Among the orientations proposed by the Agency of Quality, three are relevant in this respect: (a) teachers are invited to promote a school climate of respect, and provide feedback to students; (b) principals are invited to base curricular and pedagogical decisions on the results obtained by the school; (c) to involve parents/guardians in the design of the mission of the school.

## 3 The National Survey on the Pilot Implementation of the Schools Monitoring System

As it was mentioned at the introductory section, due to the high stakes consequences of the National System of Quality Assurance of Education, the Chilean Government decided to start in 2015 the implementation of the continuous monitoring system as a trial period. During such a period, the Agency of Quality was concerned in collecting information about five dimensions: (a) general conceptions of the school communities on the concept of quality of education, on the National System of Quality Assurance

of Education, and on the Agency of Quality; (b) general knowledge of the school communities on the four categories of school performance levels, and the access to related information; (c) perceptions and general evaluation of the school communities about the School Classification Methodology; (d) perceptions of the school communities about the relationships between the school performance level and both the learning visits and the eventual close of a school; (e) changes done by the school communities as a reaction to the performance level obtained by a school (DESUC-CEPPE, 2015).

In this section, we focus our attention on two questions from this national survey that are related to the fifth dimension mentioned above, in particular to the relationship between school leadership and pedagogical and curricular changes. More specifically, in Section 3.1 we describe the questions and their relationship to integrated leadership. Thereafter, in Section 3.2, we provide some details about the sampling framework used by the survey. We discuss in Section 3.3 the role of the leadership theoretical framework to interpret the results of the survey. Finally, in Section 3.4, we summarize the empirical evidence, explaining why it suggests to implement as an educational public policy that the intra-school practices of principals and teacher be coherent with integrated leadership.

### 3.1 Measuring the impact of the school performance level on the schools

As it was described in Section 2.3, the Agency of Quality promotes that the decisions concerning school improvement be based on the information related to the school performance level. In the National Survey on the Pilot Implementation of the Schools Monitoring System, this was explicitly inquired through questions P34 and P39 (following the original labels):

**P34.** Which of the following activities was or will be implemented in your school as a consequence of the school performance level which was informed during the pilot implementation of the school monitoring system?

- **A.** Teachers receive in a writing format the information regarding the obtained school performance level.
- **B.** Meetings with the teachers to explain the obtained performance level.
- **C.** Principals and teacher discuss, analyze or evaluate the results obtained by the school.

**P39.** According to the information you manage, which of the following actions was or will be implemented in your school as a consequence of the school performance level which was informed during the pilot implementation of the school monitoring system?

**A.** Perform pedagogical and curricular changes.

**B.** Introduce adjustments to the Plan of School Improvement (this plan is agreed between each school and the Ministry of Education).

**C.** Propose changes to the institutional educative project.

Question P34 supposes that the principals share with teachers the information provided by the Agency of Quality. The alternatives propose different levels at which the information is shared: from the mere reception of a writing communication to a joint discussion and analysis of the information. Question P39 describes different intra-school changes. The pedagogical and curricular changes are more immediate for students as well as directly related to the school improvement. The changes described in alternatives **B** and **C** have impact at the school level and at the official relationship between the school and the educational authorities.

### 3.2 Sampling framework

The sampling framework corresponds to a stratified-multistage sampling procedure. This procedure was applied to each region of the country (Chile is politically divided in 15 regions). For each region, it was available the official percentages of type of schools: public schools (fully financed by the respective county), subsidized schools (financed by both the county and parents/guardians) and private schools (financed by parents/guardians); for details, see Bellei, Vanni, Valenzuela, and Contreras (2016). At the first stage of the sampling, schools were chosen with a probability proportional to the percentage of type of schools; the choice of a school implied the choice of the principal of the school, plus the head of the technical-pedagogical unit. At the second stage, from each school at most four teachers were randomly chosen. To the final sample, specific weights were applied in such a way that the sampling distribution of schools were similar to the national distribution of schools; the weights were constructed using the number of schools by region and, conditionally to a region, the number of schools by type of school; for details, see Kalton and Flores-Cervantes (2003).

The final sample includes 385 schools, 385 principals, 385 heads of the technical-pedagogical units, and 1218 teachers; at a confidence level of .95, the error of the sample is +/-.2; for details, see DESUC-CEPPE (2015).

### 3.3 The role of the theoretical framework in the interpretation of the empirical evidence

Although the empirical evidence is not enough to perform an impact evaluation of a public policy that in turn was motivated by it, such an evidence needs a theoretical framework that organizes it in terms of outcome (denoted as $Y$), treatment (denoted as $Z$) and covariates (denoted as $X$, which can be a vector). From a modeling perspective, this means to specify a structural model enabling the interpretation of exogeneous variables as causes (in our case, the treatment $Z$) of the phenomenon to be explained (in our case, the outcome $Y$). A structural model (or causal model) conveys the idea of a representation of the world that is stable, or invariant, under a large class of interventions or of modifications of the environment; the covariates $X$ captures this aspect rather than to be explanatory factors. In this perspective, the concept of causality is internal to a model and, therefore, justified by theoretical arguments only. A structural model aims, therefore, at capturing an underlying structure; modeling this underlying structure requires taking into account the contextual knowledge of the field of application; for details, see Manzo (2010); Mouchart, Russo, and Wunsch (2010); Illari and Williamson (2012); Wunsch, Mouchart, and Russo (2014).

The empirical evidence is consequently described by the conditional distribution of $Y$ given $Z$ and $X$, which is denoted as $P(Y \mid Z, X)$. This conditional model covers the main features of a structural model:

1. Under the different contexts characterized by $X$, it is expected to observe the probabilistic relationship between the treatment $Z$ and the outcome $Y$. This corresponds to the stability character of the model.

2. For all realizations of $X$, the treatment $Z$ is an agent "exogenous to the system" that produces in probabilistic terms the outcome $Y$. This corresponds to the exogeneity of $Z$ with respect to $Y$.

It should be remarked that the causal relationship between $Z$ and $Y$ in the context characterized by $X$ is probabilistic in nature, not deterministic.

**Remark 1** Most of the empirical researchers focus their attention on the conditional expectation $E(Y \mid Z, X)$ rather than the conditional distribution $P(Y \mid X, Z)$. It is furthermore assumed that such a conditional expectation is linear in $Z$ and $X$. However, $E(Y \mid Z, X)$ is a specific characteristic of $P(Y \mid Z, X)$ and, therefore, provides a limited information.

The empirical evidence corresponding to the case study under analysis will be described in terms of the conditional distribution $P(Y \mid X, Z)$. Before doing that, let us show how the school leadership

theoretical framework allows us to relate questions P34 and P39 in causal terms. As a matter of fact, the alternative C of question P34 represents actions undertaken by principals in order to share with teachers information relevant to take the better joint decisions on school improvement. This is precisely related to both collaborative instructional leadership and transformational leadership –that is, integrated leadership.

Question P39 includes aspects related to the effects of integrated leadership. As it was noticed in Section 3.1, the alternative A measures a more immediate intra-school change than alternatives B and C, and therefore we focus our analysis on alternative A because of the temporal horizon of the National Survey. According to Marks and Printy (2003), P39A can be considered as an effect of P34C which in turn represents a combination of both leadership models. As a matter of fact, "when the principal elicits high levels of commitment and professionalism from teachers and works interactively with teachers in a shared instructional leadership capacity, schools have the benefit of integrated leadership; they are organizations that learn and perform at high levels" (p. 393). Moreover, question P39A is precisely what the Agency of Quality expects as an effect of the joint analysis of the information regarding the school performance level obtained by a school.

Summarizing, the question P34C is interpreted as the treatment policy $Z$ that the Agency of Quality searches to implement in the Chilean schools. The question P39A corresponds to the outcome $Y$ which is generated by the treatment policy: this causal relationship, illustrated in Figure 1, has been justified by the theoretical framework corresponding to integrated leadership.

### 3.4 Summary of the empirical evidence

As it was explained in Section 3.2, the survey was applied to the principal of the school, the head of the technical pedagogical unit and at most four teachers of school. From their responses to the questions P34C and P39A, we should operationalize both the variable representing the outcome $Y$ and the variable assigning the treatment policy $Z$. Two considerations should be taken into account in order to propose an operationalization of $Y$ and $Z$:

1. The Agency of Quality applied the survey as a means to inquire if intra-school changes at the pedagogical and curricular level are decided on the basis of the information regarding the school performance level.

2. According to the theoretical framework exposed in Section 3.3, this type of changes is done by teachers as the effect of integrated leadership.

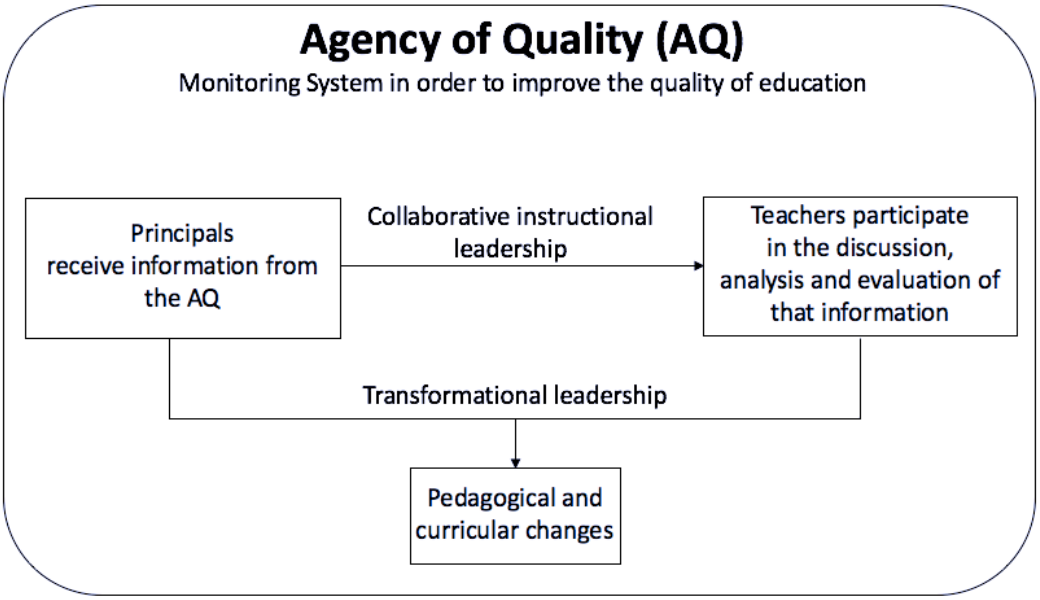


Figure 1: Causal relationship between treatment policy and outcome in the context of a explicit policy maker

Therefore, both $Y$ and $Z$ will be operationalized taken into account the information provided by the teachers. Given that the survey was applied to at most four teachers by school, three different binary codifications for the outcome can be considered:

$$Y_a = \begin{cases} 1, & \text{if at least one teacher of a school reports pedagogical and curricular changes;} \\ 0, & \text{if not;} \end{cases}$$

$$Y_b = \begin{cases} 1, & \text{if 50\% of teachers of a school reports pedagogical and curricular changes;} \\ 0, & \text{if not;} \end{cases}$$

$$Y_c = \begin{cases} 1, & \text{if 100\% of teachers of a school reports pedagogical and curricular changes;} \\ 0, & \text{if not.} \end{cases}$$

And three binary codifications for the treatment assignment variable:

$$Z_a = \begin{cases} 1, & \text{if at least one teacher of a school reports discussion and analysis of the information} \\ & \text{provided by the Agency of Quality;} \\ 0, & \text{if not;} \end{cases}$$

$$Z_b = \begin{cases} 1, & \text{if 50\% of teachers r of a school reports discussion and analysis of the information} \\ & \text{provided by the Agency of Quality;} \\ 0, & \text{if not;} \end{cases}$$

$$Z_c = \begin{cases} 1, & \text{if 100\% of teachers r of a school reports discussion and analysis of the information} \\ & \text{provided by the Agency of Quality;} \\ 0, & \text{if not.} \end{cases}$$

Using these variables, Tables 1 and 2 summarize the information collected in the survey:

Table 1: Results for Public Schools

| | $Z_a=0$ | $Z_a=1$ | $Z_b=0$ | $Z_b=1$ | $Z_c=0$ | $Z_c=1$ |
|---|---|---|---|---|---|---|
| $Y_a=0$ | 25 | 13 | 30 | 8 | 35 | 3 |
| $Y_a=1$ | 5 | 104 | 34 | 74 | 88 | 20 |
| $Y_b=0$ | 28 | 42 | 47 | 24 | 65 | 6 |
| $Y_b=1$ | 2 | 75 | 18 | 59 | 59 | 17 |
| $Y_c=0$ | 30 | 97 | 62 | 65 | 111 | 16 |
| $Y_c=1$ | 0 | 20 | 3 | 17 | 12 | 8 |

Table 2: Results for Subsidized Schools

| | $Z_a=0$ | $Z_a=1$ | $Z_b=0$ | $Z_b=1$ | $Z_c=0$ | $Z_c=1$ |
|---|---|---|---|---|---|---|
| $Y_a=0$ | 32 | 17 | 35 | 14 | 48 | 1 |
| $Y_a=1$ | 23 | 129 | 48 | 104 | 123 | 29 |
| $Y_b=0$ | 44 | 54 | 57 | 41 | 94 | 4 |
| $Y_b=1$ | 11 | 92 | 26 | 77 | 77 | 26 |
| $Y_c=0$ | 52 | 123 | 77 | 98 | 154 | 22 |
| $Y_c=1$ | 5 | 24 | 6 | 20 | 17 | 9 |

Different combinations of outcomes and treatment assignments can be used to analyze the empirical evidence. In order to show in which sense the results are stable, let us focus our attention on three of

Table 3: Estimation of the conditional probabilities for Public Schools

| **Case** | $P(Y = 1 \mid Z = 1, \text{Public})$ | $P(Y = 1 \mid Z = 0, \text{Public})$ | $P(Z = 1 \mid \text{Public})$ |
|---|---|---|---|
| $(Y_b, Z_b)$ | 0.70 | 0.28 | 0.56 |
| $(Y_c, Z_c)$ | 0.33 | 0.10 | 0.16 |
| $(Y_b, Z_c)$ | 0.74 | 0.48 | 0.16 |

Table 4: Estimation of the conditional probabilities for Subsidized Schools

| **Case** | $P(Y = 1 \mid Z = 1, \text{Subsidized})$ | $P(Y = 1 \mid Z = 0, \text{Subsidized})$ | $P(Z = 1 \mid \text{Subsidized})$ |
|---|---|---|---|
| $(Y_b, Z_b)$ | 0.65 | 0.31 | 0.59 |
| $(Y_c, Z_c)$ | 0.29 | 0.10 | 0.15 |
| $(Y_b, Z_c)$ | 0.87 | 0.45 | 0.15 |

them: $(Y_b, Z_b)$, $(Y_c, Z_c)$ and $(Y_b, Z_c)$. Tables 3 and 4 summarize the estimation of the corresponding conditional probabilities; for details on the estimation, see Appendix B. In the three cases, it is observed that the proportion of schools (public or subsidized) reporting pedagogical and curricula changes under the presence of transformational and instructional leadership is greater than the proportion of schools reporting such changes under the absence of the treatment. The only difference between these cases deals with the precision of the estimates of the conditional probabilities: for the case $(Y_b, Z_b)$, the precision is better because the number of observations in each cell is greater than in the other cases. Consequently, the empirical evidence suggests to promote that principals and teachers adequate their intra-school practices to the collaborative instructional and transformational leadership. From a public policy perspective, the relevant question is therefore the following:

> What would be the percentage of public and subsidized schools performing curricular and pedagogical changes if both principals and teachers were to modify their intra-school practices in line with collaborative instructional and transformational leadership?

This question is discussed in the next section.

# 4 Public Policy Evaluation

The previous question should be answered taking into account the specific Chilean educational context in which the Agency of Quality promotes integrated leadership in order to improve the capabilities of

principals and teacher. In particular, the Agency of Quality intends that the decision of intra-school changes be based on the analysis of the school performance level; it is expected that this analysis and the subsequent decisions are done jointly between principals and teachers. The intra-school changes, as for instance those at the curricular and pedagogical level, ensure a continuous improvement of quality of education, which in turn is of benefit for students.

### 4.1 The implementation of the treatment policy and the social welfare

To be more precise, let us consider the empirical evidence generated by $(Y_b, Z_b, X)$, where $X$ denotes the type of school, in our case public or subsidized. Let $Y_{b,1}$ be the outcome that is experienced under the treatment policy ($Z_b = 1$), and $Y_{b,0}$ be the outcome experienced under the absence of the treatment policy ($Z_b = 0$). The problem of interest is to learn about the distribution $P(Y_{b,1} = 1 \mid X = x)$ of outcomes that would be realized by schools of type $x$ if the treatment policy were in effect (that is, $Z_b = 1$), and also to learn about the distribution $P(Y_{b,0} = 1 \mid X = x)$ of outcomes that would be realized by schools of type $x$ if the treatment policy were not in effect (that is, $Z_b = 0$).

Taking into account the general purpose of the Agency of Quality, we approach this question from the perspective of a social planner required to choose between the two options, one under the treatment policy (program 1) and the other under its absence (program 0). The standard decision process of welfare economics calls for the planner to maximize a social welfare function $W(.)$, whose argument is the distribution of outcomes in a specified program. The planner should choose program 1 if $W\left[P(Y_{b,1} = 1 \mid X = x)\right] \geq W\left[P(Y_{b,0} = 1 \mid X = x)\right]$ for all $x$, and program 0 otherwise. Thus, a planner maximizing a conventional social welfare function wants to learn $P(Y_{b,1} \mid X = x)$ and $P(Y_{b,0} = 1 \mid X = x)$; for details, see Manski (1996).

### 4.2 The fundamental problem of causal inference

As it was pointed out at the introduction, the question at the end of Section 3.4 calls for extrapolations and therefore cannot be answered using the empirical evidence alone: the empirical evidence is not enough to take the social decision of implementing the treatment, that is, transformational and instructional leadership, or not. This is due to the fundamental problem of causal inference (Holland, 1986), namely that it is impossible to observe pedagogical and curricular changes in a same school under the treatment and under its absence. However, as we discuss above, in order to ensure the welfare of the society, it is necessary to compare both "fictitious worlds" and to choose the one in which the welfare is maximized.

More precisely, by the Law of Total Probability, the distributions of interest are decomposed as follows:

$$P(Y_{b,1} = 1 \mid X) \tag{4.1}$$
$$= P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) + P(Y_{b,1} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X)$$

$$P(Y_{b,0} = 1 \mid X) \tag{4.2}$$
$$= P(Y_{b,0} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) + P(Y_{b,0} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X)$$

where $P(Z_b = 1 \mid X = x)$ is the probability that a school of type $x$ is under the treatment policy ($Z_b = 1$), $P(Y_{b,1} = 1 \mid X, Z_b = 1)$ is the conditional probability that a school of type $x$ experiences pedagogical and curricular changes when actually is under the treatment policy ($Z_b = 1$), and $P(Y_{b,1} = 1 \mid X, Z_b = 0)$ is the conditional probability that a school of type $x$ would experience pedagogical and curricular changes under the treatment policy when actually is not under the treatment ($Z_b = 0$). Similar interpretations can be given for $P(Y_{b,0} = 1 \mid X, Z_b = 1)$ and $P(Y_{b,0} = 1 \mid X, Z_b = 0)$.

The empirical evidence reveals the treatment distribution $P(Z_b = 1 \mid X = x)$ and the conditional probabilities $P(Y_{b,1} = 1 \mid Z_b = 1, X = x)$ and $P(Y_{b,0} = 1 \mid Z_b = 0, X = x)$. However, the empirical evidence does not reveal the probabilities $P(Y_{b,1} = 1 \mid Z_b = 0, X = x)$ and $P(Y_{b,0} = 1 \mid Z_b = 1, X = x)$ and, consequently, the probabilities of interest, $P(Y_{b,1} = 1 \mid X = x)$ and $P(Y_{b,0} = 1 \mid X = x)$, are not identified: it is not possible to compare them in order to decide if the treatment policy should be implemented or not.

This problem can be solved by introducing hypotheses that are not empirically testable. We argue that the impact evaluation of a public policy should include two dimensions. First, the evaluation should be done with respect to different scenarios, which are characterized by non-empirical hypotheses leading to identify $P(Y_{b,1} = 1 \mid X = x)$ and $P(Y_{b,0} = 1 \mid X = x)$. The second dimension of an impact evaluation should show how each scenario determines the policy maker behavior with respect to the possible implementation of a public policy. In other words, the impact evaluation should provide some action lines coherent with the non-empirical hypotheses underlying the impact evaluation itself. Such a "behavior" corresponds to the concept of *inductive behavior* as introduced in statistics by Neyman (1957).

## 4.3 The concept of *inductive behavior*

In the 1950s, Fisher and Neyman engaged in an intense debate around the interpretation of significance tests, as well as other statistical procedures; for details on this debate, see Gigerenzer (2004). This debate is adequately summarized by the opposition between *inductive reasoning* and *inductive behavior*. Fisher (1955) considers the inductive reasoning as a mental process different from what it is traditionally called in formal logic *deductive reasoning*. This type of reasoning supplies no essentially new knowledge, but merely reveals or unfolds the implications of the axiomatic basis adopted. On the contrary, "it is the function of inductive reasoning to be used, in conjunction with observational data, to add new elements to our theoretical knowledge. That such a process existed, and was possible to normal minds, has been understood for centuries; it is only with the recent development of statistical science that an analytical account can now be given, about as satisfying and complete, at least, as that given traditionally of the deductive processes" (p.74).

Thus, for instance, in the case of a test of significance, the new theoretical knowledge consists not in affirming that the body of data at our disposal would have passed an acceptance test at some particular level, but at what level it would have been doubtful. This learning allows to establish a genuine measure of the confidence with which any particular opinion may be held in view of the particular data at our disposal, which in turn lead to do what for Fisher is the scientific work: to look "further data, probably of a somewhat different kind, which may confirm or elaborate the conclusions we drawn; but perhaps of the same kind, which may then be added to what we have already, to form an enlarged basis for induction" (Fisher, 1955, p.74).

The concept of inductive behavior was introduced by Neyman (1936) and further developed in Neyman (1957) in order to describe in a more accurate way a particular phase of scientific research, namely the concluding phase. This phase constitutes for Fisher the inductive reasoning. However, for Neyman, this is "a misnomer, [that contributes] to the confusion regarding the nature of scientific research, and that a better term would be something like *inductive behavior*" (Neyman, 1957, p.8). As a matter of fact, Neyman describes the examination of experimental or observational data by three mental processes: (i) scanning of memory and a review of the various sets of relevant hypotheses, (ii) deduction of consequences of these hypotheses and comparison of these consequences with empirical data, (iii) an act of will, a decision to take a particular action. The process (iii) results in an adjustment of our future actions to the results of observations and accordingly it should be labeled *inductive behavior* (Neyman, 1957, p. 14).

The content of the concept of inductive behavior is the recognition that the purpose of every piece of research is to provide grounds for the selection of one of several contemplated *courses of action*. Moreover, these courses of actions depend on specific circumstances and on the subjective preferences and beliefs of the individual concerned. Whatever the scientist's beliefs and preferences may be, the knowledge of the performance characteristics of all possible decision rules will allow him to choose the one that fits his case best (Neyman, 1957, pp.14, 18). This concept of inductive behavior seems relevant and adequate in order to evaluate the impact of an educational policy.

## 4.4 Scenario 1: Apply a policy is always better than don't apply it!

Let us begin the public policy evaluation by introducing a first non-empirical hypothesis which assumes that $Y_{b,1}$ and $Y_{b,0}$ are ordered in the sense that $Y_{b,1} \geq Y_{b,0}$. This means to believe that aligning principals' and teachers' intra-school actions with instructional and transformational leadership can never harm that a school experiences pedagogical and curricular changes. In less formal terms, ordered outcomes can be expressed saying that *apply the treatment policy is always better than don't apply it*. Note that this hypothesis is not empirically refutable due to the fundamental problem of causal inference.

### 4.4.1 Consequences of the ordered outcomes condition

The fundamental problem of causal inference stems from the non identifiability of the conditional probabilities $P(Y_{b,1} = 1 \mid Z_b = 0, X = x)$ and $P(Y_{b,0} = 1 \mid Z_b = 1, X = x)$. However, the ordered outcomes condition allows to partially identify them in the sense that these probabilities are bounded by identified conditional probabilities. More specifically,

$$\begin{aligned} &\text{(i)} \quad P(Y_{b,1} = 1 \mid Z_b = 0, X = x) \geq P(Y_{b,0} = 1 \mid Z_b = 0, X = x); \\ &\text{(ii)} \quad P(Y_{b,1} = 1 \mid Z_b = 1, X = x) \geq P(Y_{b,0} = 1 \mid Z_b = 1, X = x). \end{aligned} \tag{4.3}$$

For details, see Appendix A. For details on partial identification, see Manski (2007) and Tamer (2010).

These inequalities can be used to find out all the plausible values of the probabilities $P(Y_{b,1} = 1 \mid X)$ and $P(Y_{b,0} = 1 \mid X)$ that are supported by the empirical evidence. Thus, inequality (4.3.i) used in decomposition (4.1) implies that

$$\begin{aligned} &P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) + P(Y_{b,0} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X) \\ &\quad \leq P(Y_{b,1} = 1 \mid X) \\ &\quad \leq P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z = 1 \mid X) + P(Z_b = 0 \mid X). \end{aligned} \tag{4.4}$$

Similarly, inequality (4.3.ii) used in decomposition (4.2) implies that

$$\begin{aligned} & P(Y_{b,0} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X) \\ & \quad \leq P(Y_{b,0} = 1 \mid X) \\ & \quad \leq P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) + P(Y_{b,0} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X). \end{aligned} \tag{4.5}$$

The public policy evaluation under the ordered outcomes condition is based on the comparison of (4.4) and (4.5). Before doing the comparison, let us apply these bounds to the data under analysis.

### 4.4.2 Results and discussion

In order to apply (4.4) and (4.5) to the data under analysis, it is necessary to estimate the corresponding upper and lower bounds. These bounds are estimated through a bootstrap procedure based on $10^6$ replications; we report the .05 quantile of the bootstrapped distribution of the lower bound (denoted as $L-$) and the .05 quantile of the bootstrapped distribution of the upper bound (denoted as $U+$). We also report the mean of the bootstrapped distribution of both the lower and the upper bounds (denoted as $\overline{L}$ and $\overline{U}$, respectively); for details, see Appendix B.

Table 5, first line, shows the results for public schools, whereas Table 6, first line, shows the results for subsidized schools.

Table 5: Impact evaluation for public schools

| | $P(Y_1 = 1 \mid \text{Public})$ | | | | $P(Y_0 = 1 \mid \text{Public})$ | | | |
|---|---|---|---|---|---|---|---|---|
| **Scenario** | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ |
| Ordered outcomes | 0.46 | 0.52 | 0.84 | 0.88 | 0.09 | 0.12 | 0.52 | 0.58 |
| Treatment with the larger outcome | 0.35 | 0.40 | 0.52 | 0.58 | 0.46 | 0.52 | 0.68 | 0.74 |

Table 6: Impact evaluation for subsidized schools

| | $P(Y_1 = 1 \mid \text{Subsidized})$ | | | | $P(Y_0 = 1 \mid \text{Subsidized})$ | | | |
|---|---|---|---|---|---|---|---|---|
| **Scenario** | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ |
| Ordered outcomes | 0.46 | 0.52 | 0.80 | 0.85 | 0.10 | 0.13 | 0.52 | 0.58 |
| Treatment with the larger outcome | 0.33 | 0.38 | 0.52 | 0.58 | 0.46 | 0.52 | 0.72 | 0.77 |

These results deserve the following comments:

1. For public schools, at least 52% and at most 84% of them will experience curricular and pedagogical changes under the treatment policy. The precision of these percentages is given by $L-$ and $U+$; it is quite reasonable, although is better for the upper bound than for the lower bound. The same conclusion can be stated for subsidized schools. Consequently, the impact of the treatment policy on pedagogical and curricular changes does not depend on the type of school.

2. For public schools, at least 12% and at most 52% of them will experience curricular and pedagogical changes under the absence of the treatment policy. The precision of these percentages is given by $L-$ and $U+$; it is quite reasonable. The same conclusion can be stated for subsidized schools. Consequently, the impact of the absence of the treatment policy on pedagogical and curricular changes does not depend on the type of school.

3. For public and subsidized schools, the implementation of the treatment is preferable to its non implementation. This is due to the fact that the better probability of observing pedagogical and observation changes as measured by $Y_{b,0}$ corresponds to the worse probability of observing pedagogical and observation changes as measured by $Y_{b,1}$, namely

$$P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) + P(Y_{b,0} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X). \qquad (4.6)$$

   For both public and subsidized schools, this probability is equal to 0.52. It is important to remark that (4.6) corresponds to the proportion of schools of type $X$ actually experiencing pedagogical and curricular changes. Consequently, the implementation of the treatment policy implies that the proportion of schools that will experience curricular and pedagogical changes will improve with respect to what is observed in the sample. On the contrary, its non implementation means that such a proportion will decrease with respect to what is observed in the sample.

The inductive behavior associated to the ordered outcomes assumption is characterized by the inequalities (4.3): the belief in the ordered outcomes hypothesis leads to the policy maker to accept that the probability that a school of type $x$ would experience pedagogical and curricular changes under the treatment policy when actually it is not under the treatment policy, is greater than the probability that such a school experiences pedagogical and curricular changes when the treatment policy is absent; see relation (4.3.i). And that the probability that a school of type $x$ would experience pedagogical and curricular changes under the absence of the treatment policy when actually it is under the treatment, is less than the probability that such a school experiences pedagogical and curricular changes under the treatment policy: see relation (4.3.ii).

### 4.4.3 Recommendation of educational public policy

If the Agency of Quality wants to increase the proportion of schools experiencing pedagogical and curricular changes, it should implement the treatment policy by promoting its benefits and social welfare: in promoting the treatment policy, the Agency of Quality should ensure to schools that if their principals and teachers align their intra-school actions to integrated leadership, their possibility to experience curricular and pedagogical changes can not be harmed. It could be suggested that the behavior of the Agency of Quality in promoting those leadership models (as described in Section 2.3) seems to be based on this belief. In this case, the maximum proportion of schools experiencing curricular and pedagogical changes is $84\% \pm 4\%$ for public schools, and $80\% \pm 5\%$ for subsidized schools.

## 4.5 Scenario 2: Selecting of the treatment with larger outcome

Let us now introduce a second non-empirical hypothesis which assumes that the realized treatment reflects the selection of the larger outcome; that is, $Z_b = 1$ if $Y_{b,1} > Y_{b,0}$, and $Z_b = 0$ if $Y_{b,1} < Y_{b,0}$. This means to believe that if the actions of principals and teachers are aligned with instructional and transformational leadership, it is only due to the fact that in this way it is observed pedagogical and curricular changes; and that if those actions are not aligned with those leadership models, it is also due to the fact that in this way it is observed pedagogical and curricular changes.

This condition supposes what it is the criterion used by schools in order to realize the treatment or not *before* the policy maker decides to universally implement the treatment. Consequently, it is a matter of evaluating if the universal implementation of the treatment is more benefit for the society than the own decision of the schools of implementing the treatment, or not. This hypothesis is still no empirically testable due to the fundamental problem of causal inference.

### 4.5.1 Consequences of the condition of selecting the treatment with the larger outcome

The condition of selecting the treatment with the larger outcome allows to partially identify the conditional probabilities $P(Y_{b,1} = 1 \mid Z_b = 0, X = x)$ and $P(Y_{b,0} = 1 \mid Z_b = 1, X = x)$. More specifically,

$$\begin{aligned} &\text{(i)} \quad P(Y_{b,0} = 1 \mid Z_b = 0, X = x) \geq P(Y_{b,1} = 1 \mid Z_b = 0, X = x); \\ &\text{(ii)} \quad P(Y_{b,1} = 1 \mid Z_b = 1, X = x) \geq P(Y_{b,0} = 1 \mid Z_b = 1, X = x). \end{aligned} \tag{4.7}$$

For details, see Appendix A.

These inequalities can still be used to find all the plausible values of $P(Y_{b,1} = 1 \mid X)$ and $P(Y_{b,0} = 1 \mid X)$ that are supported by the empirical evidence. Thus, inequality (4.7.i) used in decomposition (4.1) implies that

$$\begin{aligned} & P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) \\ & \quad \leq P(Y_{b,1} = 1 \mid X) \\ & \quad \leq P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) + P(Y_0 = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X). \end{aligned} \tag{4.8}$$

Similarly, inequality (4.7.ii) used in decomposition (4.2) implies that

$$\begin{aligned} & P(Y_{b,0} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X) + P(Y_{b,1} = 1 \mid X, Z_b = 1)P(Z_b = 1 \mid X) \\ & \quad \leq P(Y_{b,0} = 1 \mid X) \\ & \quad \leq P(Y_{b,0} = 1 \mid X, Z_b = 0)P(Z_b = 0 \mid X) + P(Z_b = 1 \mid X). \end{aligned} \tag{4.9}$$

The public policy evaluation under the condition of selecting the treatment with the larger outcome is based on the comparison of (4.8) and (4.9). Before doing the comparison, we apply these bounds to the data under analysis.

### 4.5.2 Results and discussion

Using a bootstrap procedure based on $10^6$ replications (see Appendix B), we estimate the bounds (4.8) and (4.9). We report the .05 quantile of the bootstrapped distribution of the lower bound (denoted as $L-$) and the .05 quantile of the bootstrapped distribution of the upper bound (denoted as $U+$). We also report the mean of the bootstrapped distribution of both the lower and the upper bounds (denoted as $\overline{L}$ and $\overline{U}$, respectively). For the pubic schools, the results are gathered in Table 5, second line; for the subsidized schools, see Table 6, second line.

These results deserve the following comments:

1. For public schools, at least 40% and at most 52% of those schools will experience curricular and pedagogical changes under the treatment policy. The precision of these percentages is given by $L-$ and $U+$; it is reasonable. A quite similar conclusion can be stated for subsidized schools, although the lower bound of public schools is greater than the lower bound for subsidized schools.

Consequently, the impact of the treatment policy on pedagogical and curricular changes slightly depends on the type of school.

2. For public schools, at least 52% and at most 68% of those schools will experience curricular and pedagogical changes under the absence of the treatment policy. The precision is given by $L-$ and $U+$; it is reasonable. A quite similar conclusion can be stated for subsidized schools, although the maximum percentage of schools experiencing changes is equal to 72%. Consequently, the impact of the absence of the treatment policy on pedagogical and curricular changes slightly depends on the type of school.

3. For public and subsidized schools, the non implementation of the treatment is preferable to its implementation. This is due to the fact that the worse probability of observing pedagogical and curricular changes as measured by $Y_{b,0}$ corresponds to the better probability of observing pedagogical and curricular changes as measured by $Y_{b,1}$, namely (4.6). As it was remarked above, for both public ad subsidized schools, this probability is equal to 0.52; it corresponds to the proportion of schools of type $X$ (in the empirical evidence) experiencing pedagogical and curricular changes. Consequently, the implementation of the treatment policy implies that the proportion of schools that will experience curricular and pedagogical changes will decrease with respect to what is observed in the sample. On the contrary, its non implementation means that such a proportion will increase with respect to what is observed in the sample.

The inductive behavior associated to the assumption of selecting of the treatment with larger outcome is characterized by inequalities (4.7): the belief that the realized treatment reflects the selection of the larger outcome by a school leads to the policy maker to believe that the probability that a school of type $x$ would experience pedagogical and curricular changes under the treatment when actually it is not under the treatment policy is smaller than the probability that such a school experiences pedagogical and curricular changes when the treatment policy is absent; see relation (4.7.i). And that the probability that a school of type $x$ would experience pedagogical and curricular changes under the absence of the treatment when actually it is under the treatment is smaller than the probability that such a school experiences pedagogical and curricular changes under the treatment policy; see relation (4.7.ii).

### 4.5.3 Recommendation of educational public policy

If the Agency of Quality wants to increase the proportion of schools experiencing pedagogical and curricular changes, it should not implement the treatment policy. Its behavior should be characterized by believing that schools' decision to apply the treatment or not is more benefit for society that a universal implementation of the treatment. In this case, the maximum proportion of schools experiencing curricular and pedagogical changes is 68% $\pm 6\%$ for public schools, and 72% $\pm 5\%$ for subsidized schools.

## 4.6 Scenario 1 under treatment, or scenario 2 under its absence?

The previous two scenarios lead to opposite policy recommendations: under scenario 1, the recommendation is that the Agency of Quality implements the treatment; under scenario 2, the recommendation is that the Agency of Quality do not implement the treatment. This conclusion is not due to the particular data under analysis, but it is intrinsic to the methodological reasoning as can be seen in Table 7, where

$$\alpha(X) \doteq P(Y_{b,1} = 1 \mid X, Z_b = 1), \quad \beta(X) \doteq P(Y_{b,0} = 1 \mid X, Z_b = 0), \quad \omega(X) = P(Z_b = 1 \mid X)$$

Table 7: Summary of the intervals for scenarios 1 and 2

| **Scenario** | $P(Y_{b,1} = 1 \mid X)$ | $P(Y_{b,0} = 1 \mid X)$ |
|---|---|---|
| **1** | $[\,\alpha(X)\omega(X) + \beta(X)\{1-\omega(X)\}\,,\ \alpha(X)\omega(X) + \{1-\omega(X)\}\,]$ | $[\,\beta(X)\{1-\omega(X)\}\,,\ \alpha(X)\omega(X) + \beta(X)\{1-\omega(X)\}\,]$ |
| **2** | $[\,\alpha(X)\omega(X)\,,\ \alpha(X)\omega(X) + \beta(X)\{1-\omega(X)\}]$ | $[\,\alpha(X)\omega(X) + \beta(X)\{1-\omega(X)\}\,,\ \beta(X)\{1-\omega(X)\} + \omega(X)\,]$ |

It can be noticed that the lower bound of both intervals under consideration (scenario 1 under implementation of the treatment; scenario 2 under non implementation of the treatment) is the same. Moreover, this common lower bound corresponds to the proportion of schools that in the actual sample report pedagogical and curricular changes: both scenarios ensure to improve the actual ratio of school experiencing those changes. A criterion to choose one of them is, therefore, to select the scenario ensuring the larger proportion of schools that would experience pedagogical and curricular changes. More precisely, scenario 1 under the presence of the treatment ($Z_b = 1$) is preferable to scenario 2 under the absence of the treatment ($Z_b = 0$) if and only if

$$\alpha(X)\omega(X) + [1 - \omega(X)] > \beta(X)[1 - \omega(X)] + \omega(X),$$

which is equivalent to

$$\frac{P(Z_b = 0 \mid X)}{P(Z_b = 1 \mid X)} > \frac{P(Y_{b,1} = 0 \mid Z_b = 1, X)}{P(Y_0 = 0 \mid Z_b = 0, X)} = \frac{P(Y_b = 0 \mid Z_b = 1, X)}{P(Y_b = 0 \mid Z_b = 0, X)} \tag{4.10}$$

(the last equality is valid because of the definitions of both $Y_{b,0}$ and $Y_{b,1}$), which in turn is equivalen to

$$P(Y_b = 0, Z_b = 0 \mid X) > P(Y_b = 0, Z_b = 1 \mid X). \tag{4.11}$$

The decision is based on the empirical evidence because both (4.10) and (4.11) can be verified using data alone. However, the inductive behavior that the policy maker should assume is based on the scenarios.

For the public schools, we have that $P(Y_b = 0, Z_b = 0 \mid \text{Public}) = 0.32$ and $P(Y_b = 0, Z_b = 1 \mid \text{Public}) = 0.16$. Consequently, inequality (4.11) is satisfied and scenario 1 is chosen. For subsidized schools, $P(Y_b = 0, Z_b = 0 \mid \text{Subsidized}) = 0.28$ and $P(Y_b = 0, Z_b = 1 \mid \text{Subsidized}) = 0.20$, and consequently the conclusion is the same. However, it should be remarked that for public schools the decision is based on more stronger results ($P(Y_b = 0, Z_b = 0 \mid \text{Public})$ is double of $P(Y_b = 0, Z_b = 1 \mid \text{Public})$) than for the subsidized schools ($P(Y_b = 0, Z_b = 0 \mid \text{Subsidized})$ is 8 points larger than $P(Y_b = 0, Z_b = 1 \mid \text{Subsidized})$).

It should be stressed that the decision criterion based on inequality (4.11) is not definitive. As a matter of fact, such a criterion implies that the interval of all possible values of $P(Y_{b,1} = 1 \mid X)$ is larger than the interval of all possible values of $P(Y_{b,0} = 1 \mid X)$. The point here is that a so large interval means that the implementation of the treatment policy is not so promisor. It is enough to think in an interval like [0.10, 0.75]. In the case under analysis, we have that the length of the interval for $P(Y_{b,1} = 1 \mid \text{Public})$ under scenario 1 is equal to .32, whereas the length of the interval for $P(Y_{b,0} = 1 \mid \text{Public})$ under scenario 2 is equal to .16. Similarly, the length of the interval for $P(Y_{b,1} = 1 \mid \text{Subsidized})$ under scenario 1 is equal to .32, whereas the length of the interval for $P(Y_{b,0} = 1 \mid \text{Subsidized})$ under scenario 2 is equal to .20.

Summarizing, both scenarios imply that the proportion of schools that will experience pedagogical and curricular changes will improve with respect to the actual proportion. However, one scenario requires the implementation of the treatment policy, whereas the other scenario not. The interval of possible values for $P(Y_{b,0} = 1 \mid X)$ under scenario 2 is shorter than the interval for possible values for $P(Y_{b,1} = 1 \mid X)$ under scenario 1. This could be read either by saying that scenario 1 will improve scenario 2, or that in scenario 2 it is more sure to attain results than in scenario 2. This information is available to the Agency of Quality; the final decision according to its beliefs and preferences.

# 5 Concluding Remarks

There exist a large corpus of educational literature claiming that transformational and instructional leadership models have impact on intra-school changes helping to improve the quality of education. This corpus is accordingly used to justify the implementation of these leadership models in several national educational systems in order to importe the quality of education. Chile is an example: in the context of the National System of Quality Assurance of Education, schools are continuously monitored and. As a consequence, schools are classified according to their performance; information and guide is provided to schools by the Agency of Quality with a main objective, namely that decisions about intra-school changes be based on the analysis of those information. The Agency of Quality explicitly promotes that this type of analysis and decision be is performed in line with transformation and instructional leadership models. Specifically, the Agency of Quality collected information showing that it is more likely that a school experiences pedagogical and curricular changes when their principal and teachers align their practices to the transformational and instructional leadership models, than schools where such practices are absent.

However, the empirical evidence is not enough to justify the implementation of a treatment policy (in our case, transformational and instructional leadership); this is due to the fundamental problem of causal inference according to which it is impossible to observe the outcomes of a same school under the treatment and under its absence. Thus, an impact policy evaluation can be performed if supplementary non empirical hypotheses are introduced in order to handle the fundamental problem of causal inference. For the case of transformational and instructional leadership practices, two type of hypotheses were considered. The first one consists in assuming that the implementation of the treatment policy is better than its non implementation. Under this scenario, it is suggested to implement the treatment policy. This action ensures that the current percentage of schools experiencing curricular and pedagogical changes will improve. The second scenario is characterized by the fact that the realized treatment reflects the choice of the larger outcome. This means to believe that if the actions of principals and teachers are aligned with instructional and transformational leadership, it is only due to the fact that in this way it is observed pedagogical and curricular changes; and that if those actions are not aligned with those leadership models, it is also due to the fact that in this way it is observed pedagogical and curricular changes. Under this scenario, it is suggested to non implement the treatment policy. This action also ensures that the current percentage of schools experiencing curricular and pedagogical changes will improve.

Two opposite conclusions were reached for a same empirical evidence. The question is, What the utility of those evaluations is? Echa of the scenarios reflect different relationships between schools and the State. In the first scenario, the implementation of the treatment policy fully depends on the policy maker (the State). In this scenario, the policy maker is confident that the treatment policy is of benefit for the society. In the second scenario, the implementation of the treatment policy depends on the schools. In this case, the policy maker (the State) believes that schools' decision to apply the treatment or not is more benefit for society that a universal implementation of the treatment. To each of these scenarios, different political ideologies can be associated. Therefore, the utility of the impact evaluations is to show the implications of a specific political orientation. This does not mean that the scientific findings sustains a political option; it solely justifies actions under such an option. The implementation of instructional and transformational practices, or not is, consequently, entirely subjective; what it is objective and scientifically sustained are their implications.

**Acknowledgements:** This article is jointly authored; authors names are listed in alphabetical order. We would like to acknowledge the partial financial support of the FONDECYT Projet No. 1141030 from the National Corporation of Science and Technology CONICYT of the Chilean Government. The content of this article does not necessarily reflect the current opinion of the National Agency of Quality of Education of the Chilean Government.

## References

ACE. (2016a). *Informe Técnico SIMCE 2014*. Santiago de Chile: Author.

ACE. (2016b). *Sistema de Aseguramiento de la Calidad. Resultados Categoría de Desempeo 2016*. Santiago de Chile: Author.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Barber, M., & Mourshed, M. (2007). *How the world's best-performing schools systems come out on top*. McKinsey & Company.

Bellei, C., Vanni, X., Valenzuela, J., & Contreras, D. (2016). School improvement trajectories: an empirical typology. *School Effectiveness and School Improvement*, *27*(3), 275–292.

Bush, T., & Glover, D. (2003). *School Leadership: Concepts and Evidence*. National College for school leadership.

Cochran, W. G., & Chambers, S. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A*, *128*(2), 234–266.

DESUC-CEPPE. (2015). *Encuesta Nacional de Percepciones, Opiniones y Actitudes de Sostenedores, Equipos Directivos y Docentes sobre la Macha Blanca del Sistema de Aseguramiento de Calidad de la Educación: Fase Implementación Ordenación* (Tech. Rep.). Santigo de Chile: Instituto de Sociologa, Pontificia Universidad Católica de Chile & Centro de Estudios de Políticas y Prácticas en Educacin.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B*, *17*(1), 69–78.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, *33*(3), 329–352.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Horn, A., & Marfán, J. (2010). Relación entre liderazgo educativo y desempeño escolar: Revisión de la investigación en chile. *Psicoperspectivas*, *9*(2), 82–104.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119–135.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81-97.

Leithwood, K., Day, C., Sammons, P., Harris, A., & Hopkins, D. (2006). *Successful school leadership: What it is and how it influences pupil learning*. Nottingham: National College for School Leadership.

Leithwood, K., & Jantzi, D. (2006). Transformational school leadership for large-scale reform: Effects on students, teachers, and their classroom practices. *School Effectiveness and School Improvement*, *17*(2), 201–227.

Maddala, G. (1983). *Qualitative and limited dependent variable models in econometrics.* Cambridge: Cambridge University Press.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, C. F. (1996). Learning about treatment effects from experiments with random assignment of

treatments. *Journal of Human Resources*, *31*(4), 709–733.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

Manski, C. F., Sandefur, G., McLanahan, S., & Powers, D. (1992). Alternative estimates of the effect of family structure durig adolescence on high school graduation. *Journal of the American Statistical Association*, *87*(417), 25-37.

Manzi, J., & Preiss, D. (2013). Educational assessment and educational achievement in South America. *International guide to student achievement*, 472–474.

Manzo, G. (2010). Analytical sociology and its critics. *European Journal of Sociology*, *51*(1), 129–170.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly*, *39*(3), 370–397.

Meckes, L., & Carrasco, R. (2010). Two decades of SIMCE: an overview of the National Assessment System in Chile. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 233–248.

MINEDUC. (2005). *Marco para la Buena Dirección. Criteris para el Desarrollo Profesional y Evaluación de Desempeo*. Santiago de Chile: Unidad de Gestión y Mejoramiento Educativo. División de Educación General. Ministerio de Educaci´n. República de Chile.

MINEDUC. (2014). *Aprueba la Metodología de Ordenación de todos los Establecimientos Educacionales reconocidos por el Estado, conforme a lo dispuesto en el inciso cuarto del artículo 17 de la Ley 20.529*. Santiago de Chile: Author.

MINEDUC. (2015). *Marco para la Buena Dirección y el Liderazgo Escolar*. Santiago de Chile: Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas, CPEIP. Ministerio de Educaci´n. República de Chile.

Mouchart, M., Russo, F., & Wunsch, G. (2010). Inferring causal relations by modelling structures. *Statistica*, *70*(4), 411-432.

Neyman, J. (1936). L'estimation statistique, traitée comme un probléme classique de probabilité. *Actualités Scientifiques et Industrièlles*, *739*, 25–57.

Neyman, J. (1957). "Inductive Behavior" as a Basic Concept of Philosophy of Science. *International Statistical Review*, *25*(1-3), 7–22.

Pont, B., Nusche, D., & Moorman, H. (2008). *Improving School Leadership, Volume 1 Policy and Practice: Policy and Practice* (Vol. 1). OECD Publishing.

Robinson, V., Lloyd, C., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly*,

*44*(5), 635-674.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Seashore, K., Leithwood, K., Wahlstrom, K., & Anderson, S. (2010). *Investigating the links to improved student learning: Final report of research findings*. The Wallace Foundation.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Boca Raton: Chapmand and Hall/CRC.

Tamer, E. (2010). Partial identification in econometrics. *The Annual Review of Economics*, *2*(1), 167–195.

Waters, T., Marzano, R., & McNulty, B. (2003). *Balanced Leadership: What 30 Years of Research Tells Us about the Effect of Leadership on Student Achievement. A Working Paper*. ERIC.

Weinstein, J., & Hernández, M. (2014). Políticas hacia el liderazgo directivo escolar en chile: Una mirada comparada con otros sistemas escolares de américa latina. *Psicoperspectivas*, *13*(3), 52–68.

Wunsch, G., Mouchart, M., & Russo, F. (2014). Functions and mechanisms in structural-modelling explanations. *Journal for General Philosophy of Science*, *45*(1), 187–208.

# A Technical Appendix

Let $Y_1$ be the outcome under the treatment and $Y_0$ the outcome under its absence (or under an alternative outcome). Using the Law of Total Probability, the following decompositions are obtained:

$$P(Y_1 = 1 \mid X) = P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) + P(Y_1 = 1 \mid X, Z = 0)P(Z = 0 \mid X); \tag{A.1}$$

and

$$P(Y_0 = 1 \mid X) = P(Y_0 = 1 \mid X, Z = 1)P(Z = 1 \mid X) + P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X). \tag{A.2}$$

As it was discussed in the main text, the sampling process does not provide information on both $P(Y_1 = 1 \mid X, Z = 0)$ and $P(Y_0 = 1 \mid X, Z = 1)$. These quantities, however, lie between zero and one. This yields the following bounds on both $P(Y_1 = 1 \mid X)$ and $P(Y_0 = 1 \mid X)$, respectively:

$$\begin{aligned} &P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) \\ &\quad \leq P(Y_1 = 1 \mid X) \\ &\quad \leq P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) + P(Z = 0 \mid X), \end{aligned} \tag{A.3}$$

and

$$\begin{aligned} & P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X) \\ & \quad \leq P(Y_0 = 1 \mid X) \\ & \quad \leq P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X) + P(Z = 1 \mid X). \end{aligned} \tag{A.4}$$

Bounds (A.3) and (A.4) correspond to the worst-cases. However, it is informative in the sense that all estimation of $P(Y_1 = 1 \mid X)$ and $P(Y_0 = 1 \mid X)$ should lie in them.

**Scenario 1:** The first scenario that was discussed is defined by the non-empirical hypotheses of ordered outcomes, namely $Y_1 \geq Y_0$. It follows that $\{Y_1 = 0\} \subset \{Y_0 = 0\}$ and consequently $P(Y_1 = 1 \mid Z = 0) \geq P(Y_0 = 1 \mid Z = 0)$. Using this inequality in (A.1), we obtain a new bound for $P(Y_1 = 1 \mid X)$:

$$\begin{aligned} & P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) + P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X) \\ & \quad \leq P(Y_1 = 1 \mid X) \\ & \quad \leq P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) + P(Z = 0 \mid X). \end{aligned} \tag{A.5}$$

Note that this interval is shorter than the interval (A.3).

Similarly, $Y_1 \geq Y_0$ implies that $P(Y_1 = 1 \mid X, Z = 1) \geq P(Y_0 = 1 \mid X, Z = 1)$. Using this inequality in (A.2), we obtain a new bound for $P(Y_0 = 1 \mid X)$:

$$\begin{aligned} & P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X) \\ & \quad \leq P(Y_0 = 1 \mid X) \\ & \quad \leq P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) + P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X). \end{aligned} \tag{A.6}$$

Note that this interval is shorter than the interval (A.4).

**Scenario 2:** The second scenario that was discussed is defined by the non-empirical hypotheses of selection of the treatment with the larger outcome, namely $Z = 1$ if $Y_1 > Y_0$, and $Z = 0$ if $Y_1 < Y_0$. Now, if $Y_1 < Y_0$ (so, $Z = 0$), then $P(Y_0 = 1 \mid X, Z = 0) \geq P(Y_1 = 1 \mid X, Z = 0)$. Using this inequality in (A.1), we obtain other bound for $P(Y_1 = 1 \mid X)$:

$$\begin{aligned} & P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) \\ & \quad \leq P(Y_1 = 1 \mid X) \\ & \quad \leq P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) + P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X). \end{aligned} \tag{A.7}$$

Similarly, if $Y_1 > Y_0$ (so, $Z = 1$), then $P(Y_1 = 1 \mid X, Z = 1) \geq P(Y_0 = 1 \mid X, Z = 1)$. Using this inequality in (A.2), we obtain other bound for $P(Y_0 = 1 \mid X)$:

$$\begin{aligned} & P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X) + P(Y_1 = 1 \mid X, Z = 1)P(Z = 1 \mid X) \\ & \quad \leq P(Y_0 = 1 \mid X) \\ & \quad \leq P(Y_0 = 1 \mid X, Z = 0)P(Z = 0 \mid X) + P(Z = 1 \mid X). \end{aligned} \tag{A.8}$$

# B Statistical procedure to estimate the bounds

The bounds of the intervals (A.5), (A.6), (A.7) and (A.8) are estimated by a bootstrap procedure as suggested by Manski, Sandefur, McLanahan, and Powers (1992). The procedure is the following:

**a.** Estimate the conditional probabilities $P(Y = i, Z = j \mid X = x)$, for $i = 0, 1$, $j = 0, 1$ and $x \in \{\text{Public, Subsidized}\}$ nonparametrically. We use the naive estimator corresponding to the relative frequency; for details, see Silverman (1986, Chapter 1) or Manski (1995, Chapter 1).

**b.** Apply the estimate $P[(Y, Z) \mid X = x]$ to draw a simulated realization of $(Y, Z)$ for each member of the sample, hence generating a pseudo-sample.

**c.** Estimate the bounds corresponding to the intervals (A.5), (A.6), (A.7) and (A.8) on the pseudosample data.

**d.** Repeat steps **b** and **c** $10^6$ times, thereby yielding a bootstrapped sampling distribution for the bounds.

**e.** Report the .05 quantile of the bootstrapped distribution of the lower bound and the .95 quantile of the bootstrapped distribution of the upper bound. We also report the mean of the boostrapped distribution of both the lower and the upper bounds.

# C Complementary results

In Section 3.4, the empirical evidence was summarized considering it was generated by $(Y_c, Z_c, X)$ and $(Y_b, Z_c, X)$. Tables 8 and 9 summarize the impact evaluation for the first case, whereas Tables 10 and 11

summarize the impact evaluation for the second case. For each case, the results for public and subsidized schools are practically the same. In both cases, it seems that the better decision is to implement the treatment policy under the inductive behavior corresponding to ordered outcomes. However, for the cases characterized by $(Y_c, Z_c, X)$, the corresponding intervals are quite large (.76) and therefore the better decision is to opt by the non implementation of the treatment policy.

Table 8: Impact evaluation under three scenarios for public schools, case $(Y_c, Z_c, X)$

| | $P(Y_1 = 1 \mid \text{Public})$ | | | | $P(Y_0 = 1 \mid \text{Public})$ | | | |
|---|---|---|---|---|---|---|---|---|
| **Scenario** | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ |
| Ordered outcomes | 0.09 | 0.13 | 0.89 | 0.93 | 0.03 | 0.05 | 0.13 | 0.17 |
| Treatment with the larger outcome | 0.05 | 0.08 | 0.13 | 0.17 | 0.09 | 0.13 | 0.24 | 0.29 |

Table 9: Impact evaluation under three scenarios for subsidized schools, case $(Y_c, Z_c, X)$

| | $P(Y_1 = 1 \mid \text{Subsidized})$ | | | | $P(Y_0 = 1 \mid \text{Subsidized})$ | | | |
|---|---|---|---|---|---|---|---|---|
| **Scenario** | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ |
| Ordered outcomes | 0.09 | 0.13 | 0.89 | 0.93 | 0.02 | 0.05 | 0.13 | 0.17 |
| Treatment with the larger outcome | 0.06 | 0.09 | 0.13 | 0.17 | 0.09 | 0.13 | 0.24 | 0.29 |

Table 10: Impact evaluation under three scenarios for public schools, case $(Y_b, Z_c, X)$

| | $P(Y_1 = 1 \mid \text{Public})$ | | | | $P(Y_0 = 1 \mid \text{Public})$ | | | |
|---|---|---|---|---|---|---|---|---|
| **Scenario** | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ |
| Ordered outcomes | 0.46 | 0.52 | 0.96 | 0.98 | 0.09 | 0.12 | 0.52 | 0.58 |
| Treatment with the larger outcome | 0.35 | 0.40 | 0.52 | 0.58 | 0.46 | 0.52 | 0.56 | 0.62 |

Table 11: Impact evaluation under three scenarios for subsidized schools, case $(Y_b, Z_c, X)$

| | $P(Y_1 = 1 \mid \text{Subsidized})$ | | | | $P(Y_0 = 1 \mid \text{Subsidized})$ | | | |
|---|---|---|---|---|---|---|---|---|
| **Scenario** | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ | $L-$ | $\overline{L}$ | $\overline{U}$ | $U+$ |
| Ordered outcomes | 0.45 | 0.51 | 0.98 | 0.99 | 0.09 | 0.13 | 0.51 | 0.57 |
| Treatment with the larger outcome | 0.33 | 0.38 | 0.51 | 0.57 | 0.45 | 0.51 | 0.53 | 0.59 |