

# On the Effect of the English Educational Program in a Japanese

## Elementary School

\*Takeshi Emura and \*\*Hitomi Katsuyama

### **Abstract**

We are interested in assessing the effect of an English educational program introduced to a Japanese elementary school, School A, in Japan. We conducted English tests and surveys on students from School A, and from School B - which had not been introduced to any English education. The presence of covariates in our survey makes it difficult to interpret a pure school effect, and the imbalance of the covariates between the two groups also makes the two-sample  $t$  test biased. Under such limitation of data, counterfactual models of causality provide a sophisticated framework of studying in a pure school effect. In this paper, we analyze our dataset by using techniques available under the counterfactual framework for adjusting pretreatment covariates. Data analysis reveals the effect of an English program by effectively eliminating the possible influence of covariates. We also discuss the consequence of adjusting for posttreatment covariates and the possible effect of unmeasured covariates.

**KEYWORDS:** Causal effect, confounding factor, counterfactual model, regression analysis, school effect, two-sample  $t$  test.

## **1. Introduction**

In Japan, students start learning English for the first time in junior high schools. English has not been taught in elementary schools until recently. In 2002, the Ministry of Education, Culture, Sports, Science and Technology introduced a “Period of Integrated Study” as part of the curriculum in elementary schools. Each elementary school may include English as one of the areas in the ‘Period of Integrated Study’. However, the Japanese government has not yet decided to teach English as a mandatory subject at elementary schools, unlike Taiwan, where it has been a mandatory subject from the 5<sup>th</sup> grade since 2001 and South Korea, where it has been mandatory subject from the 3<sup>rd</sup> grade since 1997. In March 2006 the Central Council of Education in Japan suggested that English education should start from the 5<sup>th</sup> grade as a compulsory subject, which may start in 2010 at the earliest. Some problems that still remain are, such as who are going to teach, how English should be taught at elementary schools and how much time should be spent for English education.

The other reason why the Japanese government has not decided to make English as a part of the formal curriculum at the elementary school level is that there are many people including some authorities who wonder if English education at elementary schools is an effective use of valuable classroom time (Ōtsu, 2004, 2005). Our

research is motivated by the question for the effectiveness of English education in Japanese elementary schools.

The Ministry reported that because of the aims and purposes of the ‘Period of Integrated Study’, student evaluation should not be based on test scores as done with other regular classes (Ministry of Education, 2001). It also reported that the student evaluations should be based on descriptive assessments about learning conditions and progress that can be observed in the students’ degree of participation in activities and the learning processes. This can also includes what can be grasped in the students’ presentations, their enthusiasm and attitude towards learning.

Due to the nature of ‘the Period of Integrated Study’, there are few quantitative studies that objectively investigate the benefit of English education and to avoid public criticism, most elementary schools tend to prohibit quantitative evaluations on their students for the sake of research. It is rare to measure test scores both from students who have received English education at elementary school and the students who have not (Katsuyama et al., 2006). On the other hands, most Japanese educational researchers have focused on the descriptive nature of student learning on English at elementary school (Takagi, 2003).

In the study on the school effects of learning English at elementary schools it is natural to evaluate each school by comparing the results of some standardized tests.

When comparing two schools, one good measure is the difference of mean test scores given by

$$\hat{\tau}_{simple} = \bar{y}_1 - \bar{y}_2,$$

where  $\bar{y}_1$  and  $\bar{y}_2$  are the mean test scores between school 1 and school 2. If the students from each school have similar background characteristics, comparisons can be made based on the standard two-sample *t*-test. However, if students in the two schools are not homogeneous, the difference of the score may be due to extraneous variables other than the school effect. For example, high scores on the test may not reflect the educational effect produced by the school but the effect due to some external factors such as studying English outside the school or the students' scholastic years. These factors are sometimes called confounding factors. To see a pure educational effect of a particular school, we need to have more information about students' background characteristics (Raudenbush and Willms, 1995).

Counterfactual model of causality is a statistical framework of thinking about the effect of treatment between two populations. In the sociological and educational contexts, Winship and Morgan (1999) summarized the basic framework for studying the effect of treatment under a counterfactual framework. A similar review is available in Heckman et al. (1998) for econometrical studies. Counterfactual models of causality have been especially applied to evaluate the impact of introducing some

social and economic policies such as a labor training program (Dehejia and Wahba, 1999) and kindergarten retention policy (Hong and Raudenbush, 2006). In such contexts, a randomized evaluation of the policies cannot always be implemented and techniques of matching, stratification and propensity score play an important role in the fair evaluation of the policies (Rosenbaum and Rubin, 1983).

We discuss the estimated effect of introducing an English educational program in Japanese elementary schools under counterfactual models of causality. We use data from our survey research on two different elementary schools, School A and School B in Chiba prefecture, Japan. School A has introduced English education program since 2000, and School B has not yet introduced any English education (See Appendix A for more details). We identify the school effect based on the counterfactual model, and then apply estimators to control the differences in the extraneous covariates between School A and B. When we apply these estimators to our dataset we get the estimates of the school effect that are more conservative than the simple estimates  $\hat{\tau}_{simple}$ , but we still obtain a high statistical significance of effect (See Section 5.3).

This article is organized as follows. Section 2 introduces our dataset and presents the preliminary analysis. Section 3 defines the school effect under counterfactual framework. Section 4 discusses the estimation strategies for the school effect. Section 5 applies these estimators to our dataset. Section 6 considers the case when

we use covariates that are measured after school assignments and finally in Section 7, we conclude our article.

## **2. Data and Preliminary Analysis**

Our primary interest is to measure the effect of an English educational program applied to a Japanese elementary school, School A in Chiba Prefecture, Japan. We conducted both English tests and surveys to two schools: School A and School B. Our analytic sample includes 369 elementary students from School A and 146 students from School B. The students in their first scholastic year from School A did not take the test because the test date did not fit their schedule. Although the test and survey were conducted for the first scholastic year students in School B, we eliminate this part of data from our analysis.

Table 1 compares the averages of test scores after stratifications based on students' scholastic years and English learning experience at kindergarten, both of which are factors before their school assignments. To quantify the difference of the average scores between School A and B, p-values are calculated based on the standard *t*-test on each stratum and adjusted by the Holm's method for multiple comparisons. In general, the average test scores are higher in School A than those in School B. It is also evident from the table that the kindergarten experience of studying English gives rise to higher score on our tests. In addition, the students' scholastic year gives the

monotone increasing effect on the test scores. Given these results in mind, the  $t$ -test that directly compares all students in School A and B in the last row of Table 1 will be misleading, and an appropriate adjustment for these confounding factors is necessary.

The present analysis in Table 1 is useful only to examine the trend of the test scores in each stratum. Our final goal is to measure a global effect on the English program rather than individually categorized effects. Therefore, it would be preferred to have a single measure for the school effect.

### **3. School Effect in the Framework of the Counterfactual Model**

In this section, we define the school effect based on the counterfactual framework.

The counterfactual model of causality assumes that students have two hypothetical scores of the test. Let  $Y_i(0)$  denote the score for students  $i$  when he/she were educated in School B, and  $Y_i(1)$  denote the score for students  $i$  when he/she were educated in School A. Note that only one of  $(Y_i(0), Y_i(1))$  can be observed, and thus, the other is viewed as a hypothetical latent variable. Let  $T_i$  be the indicator such that  $T_i = 1$  if student  $i$  was educated in School A, and  $T_i = 0$  if student  $i$  was educated in School B. Then  $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$  is the observed test score for student  $i$ .

As mentioned earlier, external experiences of learning English or the student's scholastic years are potential confounding factors that obscure the inference of the school effect. To account for such effects, some background information was also

recorded. Let  $(X_i, Z_i)$  be a pair of pretreatment covariates for student  $i$  such that  $X_i = 1$  if student  $i$  had some experience in studying English at kindergarten,  $X_i = 0$  if student  $i$  did not have any experience of studying English at kindergarten, and the discrete variable  $Z_i \in \{2, \dots, 6\}$  represents student  $i$ 's scholastic years. Then, the school effect can be defined as

$$\tau = E(Y_i(1)) - E(Y_i(0)) \quad (1)$$

which is the main parameter of interest. Here the expectation is taken over the distribution from which  $(Y_i(0), Y_i(1), T_i, X_i, Z_i)$  are drawn. This quantity is more widely called the average treatment effect (Rosenbaum and Rubin, 1983).

Some researchers have studied the effect of school on learning through regression models. Alwin (1976) and Morgan (2001) imposed a linear regression model in which the school effect is measured by a regression coefficient. However, if the effects of covariates are allowed to vary across schools A and B, it is not entirely trivial to choose a reasonable parameter to measure the school effect (Morgan, 2001, p. 347). Our definition of the school effect in (1) based on the counterfactual model of causality is not confined to a particular regression model and hence provides a more coherent way of defining the school effect. While the regression approach would be inadequate in defining the school effect, it is still useful in estimation of the effect as we will see in the next section.

## 4. Inference Procedures

In this section we review three methods that are available in literature to estimate the school effect defined in Section 3. Although more information is available from our questionnaire, we can only use pretreatment covariates  $(X_i, Z_i)$  for adjusting selection bias (Rosenbaum, 1984b). The use of study experiences at cram schools, English conversation schools, foreign countries and experience due to school transferring is considered in Section 6.

### 4.1 Estimator by Linear Regression

If a certain regression model is imposed on the counterfactual model, the school effect is identifiable and can be expressed as a regression coefficient. This application of regression analysis is called the analysis of covariance (ANCOVA). In this section we review the analysis of covariance method mentioned in Winship and Morgan (1999) and Morgan (2001). We assume that observations  $(Y_i, T_i, X_i, Z_i)$  satisfy the regression model

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i, \quad (2)$$

where the error terms  $\varepsilon_i$ 's are independent and have mean 0. The school effect can be measured by the coefficient  $\beta_1$  and it can be estimated by the least square method. It follows that the school effect has the relation  $\tau = E(Y_1(1)) - E(Y_1(0)) = \beta_1$  under

independence between  $T_i$  and  $\varepsilon_i$ . Thus, the least square estimates, denoted as  $\hat{\tau}_{reg}$ , can consistently estimate the school effect  $\tau$ . Despite the simplicity of the analysis based on ANCOVA, the model assumption (2) may be too strong. Morgan (2001, p.346) pointed out that simple linear regression models as in (2) are not adequate to describe the theory of learning. We recommend checking the model assumption (2) before doing the ANCOVA. A method for model checking is illustrated in Section 5 with real data.

#### **4.2 Estimator by Sub-classification on $(X, Z)$ .**

As shown in Table 1, we can form  $2 \times 5 = 10$  strata based on the pretreatment variables of  $(X_i, Z_i)$ 's. Within each stratum, we take a difference in averages of the score between School A and B, then compute the sum of these ten mean differences, each weighted by the number of students in the stratum (See the textbook Rosenbaum, 2002, p.47). This estimate can be easily calculated from Table 1, and we denote this estimator by  $\hat{\tau}_{strat}$ .

#### **4.3 Matching Estimator**

Standard techniques for adjustment in observational study are matched sampling (Cochran, 1965). Since the number of students is smaller in School B, each student in School B is matched with the comparable students in School A. Nearest neighbor or exact matching methods are popular strategies for estimating the impact of social

programs (Heckman et al., 1998). Specifically, for student  $i$  in School B, we calculate the test score subtracted by the average of all comparable students in School A:

$$Y_i - \frac{\sum_l Y_l I(X_l = X_i, Z_l = Z_i)}{\sum_l I(X_l = X_i, Z_l = Z_i)}.$$

Here, the indicator variable  $I(X_l = X_i, Z_l = Z_i)$  is defined to be one when both  $X_l = X_i$  and  $Z_l = Z_i$  hold and defined to be zero otherwise. The averaged differences over all students in School B give the estimates

$$\hat{\tau}_{match} = -\frac{1}{\sum_i (1 - T_i)} \sum_i (1 - T_i) \left\{ Y_i - \frac{\sum_l Y_l I(X_l = X_i, Z_l = Z_i)}{\sum_l I(X_l = X_i, Z_l = Z_i)} \right\}.$$

#### 4.4 Strongly Ignorable Treatment Assignment

Here the essential assumption, under which  $\tau$  can be correctly identified, is the strongly ignorable assumption (Rosenbaum and Rubin, 1983):

$$T_i \perp (Y_i(0), Y_i(1)) \mid X_i, Z_i.$$

Here the notation  $\perp$  and  $\mid$  indicates that  $T_i$  and  $(Y_i(0), Y_i(1))$  are independent given the pretreatment covariates  $(X_i, Z_i)$ . Under this assumption, one can easily derive the equations

$$\begin{aligned} \tau &= E\{Y(1) - Y(0)\} \\ &= E[E\{Y \mid T = 1, X, Z\} - E\{Y \mid T = 0, X, Z\}] \end{aligned}$$

This expression implies that, by observing many values of  $Y_i$ 's in treatment and control groups at each value of  $(X, Z)$ , we can recover the information on the the

school effect  $\tau$ . Based on this equation, one can prove that the non-parametric estimators  $\hat{\tau}_{strat}$  and  $\hat{\tau}_{match}$  are unbiased for the school effect.

In general, it is not easy to check this assumption without other information besides the data. Rosenbaum (1984a) developed several methods for checking this assumption based on assumed causal models. We will use one of his ideas for the analysis of our dataset in Section 5. One desirable property of the estimators  $\hat{\tau}_{strat}$  and  $\hat{\tau}_{match}$  is that no parametric assumption is imposed on the counterfactual model beside the strongly ignorable assumption.

## 5. Data Analysis

In this section, we apply the methods in Section 4 to analyze our dataset. In our questionnaires, the experience was originally measured as “total years” (see Appendix A) but, for the present purpose, the variables were dichotomized and scored as 1 if he/she has any experience of studying English at kindergarten and 0 otherwise.

### 5.1. Model Adequacy for the ANCOVA

Before conducting the ANCOVA, we check the adequacy of the model assumption (2).

At first, we fitted three different regression models and compared  $R^2$ , the measure of goodness-of-fit (Sen & Srivastava, p. 39):

$$\text{Model I: } Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i, \quad R^2 = 0.0402$$

$$\text{Model II: } Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i, \quad R^2 = 0.2757$$

Model III (Model II plus interaction):

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 Z_i + \gamma_2 X_i T_i + \gamma_3 Z_i T_i + \varepsilon_i, \quad R^2 = 0.2769$$

Model I does not include any covariate and thus provides the least favorable value of  $R^2 = 0.0402$ . Model II fits the analysis model presented in the ANCOVA in Section 4.1. By including the covariate information, the measure of fit  $R^2$  increases to 0.2757. The regression coefficients under Model II are  $\hat{\beta}_1 = 1.7603$  (p-value  $< 10^{-6}$ ),  $\hat{\beta}_2 = 1.2476$  (p-value = 0.00104) and  $\hat{\beta}_3 = 1.6958$  (p-value  $< 10^{-6}$ ). Model III allows the regression coefficients to be different in School A and B. Under Model III, the value of  $R^2$  increases from 0.2757 to 0.2769, which seems negligible. In fact, the interaction terms under Model III are  $\hat{\gamma}_2 = -0.1656$  (p-value = 0.8413) and  $\hat{\gamma}_3 = 0.2218$  (p-value = 0.4280) respectively and are not statistically significant. Thus, Model II would be sufficient to describe the linear relation.

To check the adequacy of linearity assumption in model (2), we show the corresponding residual plot in Figure 1. This model checking procedure is described in standard text books such as Neter et al. (1999) on page 99. The residual plot displayed in Figure 1 shows no departure from the linear model (2).

## 5.2. Checking Strongly Ignorable Assumption

We adopt the idea from Section 5 in Rosenbaum (1984a) in which a diagnostic procedure for strong ignorability is constructed based on the assumed causal model.

For our dataset, we assume the causal model as follows:

$$Y(0) \leq Y(1) \quad a.s.$$

This assumption may be reasonable since the educational program in School A would not produce a negative effect on the test score for all students. Under this causal model and strongly ignorable assumption, it is easy to see that, for all  $t \geq 0$ ,

$$\begin{aligned} \Pr[Y(0) \leq t | T = 0, X, Z] &= \Pr[Y(0) \leq t | X, Z] \\ &\geq \Pr[Y(1) \leq t | X, Z] \\ &= \Pr[Y(1) \leq t | T = 1, X, Z] \end{aligned}$$

Thus, under the causal model, the strongly ignorable assumption implies the stochastic order relation:

$$\Pr[Y(0) \leq t | T = 0, X, Z] \geq \Pr[Y(1) \leq t | T = 1, X, Z] \quad (t \geq 0) \quad (3)$$

Equation (3) does not generally hold if the strongly ignorable assumption fails (Rosenbaum, 1984a). We check the equation (3) by comparing the two empirical distributions, that is,

$$\text{est.Pr}[Y(0) \leq t | T = 0, X = x, Z = z] = \frac{\sum_i I(Y_i \leq t, T_i = 0, X_i = x, Z_i = z)}{\sum_i I(T_i = 0, X_i = x, Z_i = z)},$$

$$\text{est.Pr}[Y(1) \leq t | T = 1, X = x, Z = z] = \frac{\sum_i I(Y_i \leq t, T_i = 1, X_i = x, Z_i = z)}{\sum_i I(T_i = 1, X_i = x, Z_i = z)},$$

where  $x \in \{0,1\}$  and  $z \in \{2, \dots, 6\}$ . Figure 2 represents the plots of empirical distributions for different combinations of covariates. For all the setting, the plots of  $\text{est.Pr}[Y(0) \leq u | T = 0, X, Z]$  tend to be larger than or nearly equal to those of  $\text{est.Pr}[Y(1) \leq u | T = 1, X, Z]$ .

A concern in the study is the possibility that even after adjustment for the

observed covariate  $(X_i, Z_i)$ , we still cannot compare the two schools. In the broad governmental survey research on American high schools, Coleman et al. (1982) used ten pretreatment covariates, including parents' education (see p.138 of their book) for regression adjustment. In our study, we would like to include the parents' education for adjustment since it may influence children's English study environment. However, we neglected to collect the information about the parents' educational background because it might hurt the students' feeling and could not obtain parents' understanding.

However, the following facts may support the validity of our data analysis. Firstly, we have no significant evidence for rejecting the strongly ignorable assumption based on adjustment by  $(X_i, Z_i)$  from the diagnostic plots in Figure 2. Then, strongly ignorable assumption theoretically guarantees that the  $(X_i, Z_i)$ -adjusted estimators in Section 4 provide an unbiased estimate of the school effect. Secondly, there are some limitations for parents to choose the school their children attend. There are not many private elementary schools in Japan (less than one percent). Especially, there is no any private elementary schools where School A and B are located. Basically children are expected to go to elementary schools in their school districts. The two schools are located in residential area within an hour commute from Tokyo.

### **5.3. Estimation of the School Effect**

Now we assess the school effect  $\tau$  based on the estimation procedures in Section 4. Notice that the scale of our test score ranges from 0 to 33, and so the estimates of  $\tau$  reflect the increase of the test score in this scale. The three point estimates, namely,  $\hat{\tau}_{reg}$ ,  $\hat{\tau}_{strat}$  and  $\hat{\tau}_{match}$ , and the naïve estimates  $\hat{\tau}_{simple}$  are given in Table 2. To investigate the precision of the three estimators, we performed the Bootstrap procedure. Specifically, we draw samples  $\{(Y_i^*, T_i^*, X_i^*, Z_i^*); i = 1, \dots, 515\}$  with replacement from  $\{(Y_i, T_i, X_i, Z_i); i = 1, \dots, 515\}$  for  $B=50,000$  times. For these re-sampling data, the mean and standard deviation are calculated and the percentile method is used to compute the 95 percent confidence intervals.

All three point estimates yield similar results that students in School A performed  $\hat{\tau}_{reg}=1.760$ ,  $\hat{\tau}_{strat}=1.525$  and  $\hat{\tau}_{match}=1.765$  points higher on our test than those in School B. As mentioned earlier,  $\hat{\tau}_{simple}=1.973$  estimates the incorrect school effect,  $E(Y_1(1) | T_1 = 1) - E(Y_1(0) | T_1 = 0)$ , while the other four estimates can capture the correct school effect  $\tau$  in (1). As we expected,  $\hat{\tau}_{simple}$  grossly overestimates the true school effect. Although standard deviations of three estimates were similar, the smallest was attained by the ANCOVA. 95 percent intervals of all estimates do not cover 0, and these results support the effectiveness of introducing English educational program with 5 percent statistical significance. Again, ANCOVA provides the shortest interval among those valid estimates on the school effect.

## 6. The School Effect Adjusted by Full Covariate Information

In our questionnaire, the experience of studying English outside the school is originally measured as the study experience in four categories (Appendix A). This section discusses the consequence of using these additional English experience indicators, which are measured after the school assignments.

### 6.1 A Theory Describing the Effect of Posttreatment Variables

Up to now, we have used the indicator of kindergarten experience  $X_i$  and students' scholastic year  $Z_i$  for adjusting selection bias. Let the three binary variables  $V_1, V_2$  and  $V_3$  be the experience in the following categories.  $V_1$ : Indicator for learning English at cram schools or English conversation schools.  $V_2$ : Indicator for living in a foreign country.  $V_3$ : Indicator for learning English at school. The variable  $V_{3i}$  is one for all students in School A and zero for almost students in School B. Five students in School B had  $V_{3i} = 1$  because they are transferred from the other schools where they studied English. Thus, a vector of binary covariates  $S_i = (V_{1i}, V_{2i}, V_{3i})$  is available for each student in addition to the pretreatment covariates  $(X_i, Z_i)$ . In observational studies, the three variables  $V_1, V_2$  and  $V_3$  are called posttreatment variables since their experience status may be changed after school assignments (Rosenbaum, 1984b). In this case, the direct application of regression, matching and stratification methods in Section 4 with adjustment for full variables  $(V_{1i}, V_{2i}, V_{3i}, X_i, Z_i)$  does not provide

a consistent estimator for the school effect in (1).

To explain the effect of affected covariates, we introduce counterfactual models on these covariates. Let  $S_i(0)$  and  $S_i(1)$  be a vector of the indicators of the posttreatment experience of studying English outside the school. We only observe the variable  $S_i \equiv (V_{1i}, V_{2i}, V_{3i})' = S_i(1)T_i + S_i(0)(1 - T_i)$  as in the setting for  $(Y_i(0), Y_i(1))$ . Also, let  $U_i \equiv (U_{1i}, U_{2i}, U_{3i})'$  be a vector of hypothetical pretreatment indicators for the experience of studying English before attending the schools. It is logical to think that kindergarten experience and scholastic year  $(X_i, Z_i)$  are not affected by the school assignment, and can safely be considered as pretreatment variables. Under the framework of Rosenbaum (1984b), the “net treatment effect” is defined as

$$\tilde{\Delta} = E[\Delta(S_i, X_i, Z_i)],$$

where,

$$\Delta(s, x, z) = E[Y_i(1) | S_i(1) = s, X_i = x, Z_i = z] - E[Y_i(0) | S_i(0) = s, X_i = x, Z_i = z].$$

In general,  $\tilde{\Delta} \neq \tau$ , and  $\tau$  is not directly estimable due to the unobserved confounding variable  $U_i$ . Following Rosenbaum (1984b), the sufficient conditions for  $\tilde{\Delta} = \tau$  are:

- (a) Strong ignorability:  $(Y_i(0), Y_i(1), S_i(0), S_i(1)) \perp T_i | (U_i, Z_i)$ ,
- (b) Surrogacy of  $S_i(t)$  for  $U_i$ :  $Y_i(t) \perp U_i | (S_i(t), Z_i)$  for  $t = 0, 1$ , and
- (c) Unaffected covariates:  $S_i(0) = S_i(1) = S_i$ .

If one of these conditions fails, there is no guarantee that  $\tilde{\Delta} = \tau$  holds. In this case, those methods in Section 4 using full covariates  $(V_{1i}, V_{2i}, V_{3i}, X_i, Z_i)$  no longer estimate the school effect  $\tau$ , but still consistently estimate  $\tilde{\Delta}$ , the net treatment effect.

## 6.2 Estimation of the Net School Effect

The educational effect may be attributed to many sources, a few of which are of substantial interest in evaluation of the program. Suppose that the introduction of an English educational program encourages students to study English outside the school, which in turn improves their English ability. To occupy the valuable class time in elementary schools, one may prefer to evaluate the pure educational effect of School A, which is measured by the net school effect, hereby eliminating the possible effect caused by all outside studies, including posttreatment factors. Thus, the estimates for the net school effect would still retain a reasonable interpretation for our purpose.

We applied estimators in Section 4 using full covariates  $(V_{1i}, V_{2i}, V_{3i}, X_i, Z_i)$  in place of  $(X_i, Z_i)$ . Table 3 shows the result of data analysis based on the full covariate information. The point estimates  $\hat{\tau}_{reg}=1.585$ ,  $\hat{\tau}_{strat}=1.367$  and  $\hat{\tau}_{match}=1.462$  were similar to those estimates using only  $(X_i, Z_i)$  but slightly reduced. The same Bootstrap procedure was performed for evaluating the sampling distributions for the estimators. The 95% confidence intervals for each estimator are away from 0, and

they provide the significant evidence for the positive educational effect of School A in terms of the net school effect. We are happy to see the results using the full covariates are similar to those using the pretreatment covariates since it indicates the robustness of the results due to different choice of the school effect.

## **7. Conclusion**

In this article, we have emphasized the importance of the quantitative research of the effect of English education in the elementary school in Japan. The framework of counterfactual models of causality has been shown to be a useful tool to quantify the school effect. We introduced popular methods of analysis, namely regression, stratification and matching method for estimating the school effect. Data analysis using these methods revealed a positive effect of the English educational program in School A. It is also shown that the English educational program introduced in School A brought a positive net school effect on learning English, where all possible effects caused by observed outside studies were removed.

One must notice that observed covariates in our study are only five variables, namely  $(V_{1i}, V_{2i}, V_{3i}, X_i, Z_i)$ , and the effect of unmeasured covariates is still unknown. Empirical analysis and subject matter discussion in Section 5.2 state that the pretreatment variables  $(X_i, Z_i)$  are sufficient to remove selection bias. Although there is reason to believe the adjustment is sufficient, it is not common to use only a

few covariates for adjustment in educational research. As mentioned earlier, we had difficulty collecting student backgrounds, such as parents' education, from Japanese elementary school. To collect more background information from students and their families, more large-scale research (possibly, governmental research) must be conducted. However, with the decreasing number of elementary schools without English education, the current dataset still provide useful research materials.

The results from our study, based on the rigorous quantitative evaluation through the models of causality may help appreciate the effectiveness of English education at the elementary school levels. We hope that the result of data analysis will be a useful resource for the support of English educational program as a formal subject in future.

### **Acknowledgements**

Most of this paper was completed while the first author was visiting National Chiao Tung University, Taiwan, R.O.C. The authors wish to thank Dr. Jingfang Wang and Dr. Weijing Wang for some useful comments.

### **Appendix A (Details of the Survey )**

The second author conducted all of our survey. The purpose of this survey was to determine the English ability of elementary school students of Schools A and B and to do research on the effect of English education conducted at school and after-school

institutes.

*Dates of Survey:* The survey was conducted at School A from December 15th to 17th, 2003 and at School B from February 24th and 27th, 2004 respectively, which were in the same school year of 2003. Each student answered the proficiency test which lasted 14 min. and answered questionnaire immediately after the test.

*School A:* English education has been introduced since 2000 and mainly Japanese Teachers of English (JTE) taught classes with homeroom teachers and sometimes with native English speakers called Assistant Language Teachers (ALT) once every week or every two weeks.

*School B:* English education has not been introduced yet. However, they had some events with ALT once or twice a year since 2002.

*Participants:* The number of students between the 2nd and 6th grade from School A who took our tests was 369, and that of students between 1st and 6th grade from School B was 178. The students in the first grade from School A did not take the test because the test schedule did not fit into their schedule.

*Proficiency Test:* The proficiency test consists of 33 questions which were extracted from commercially available mock tests of the Junior STEP (Jidoueiken

Challenge Book, 1994). The questions extracted from Level 1 (the highest) to 3 (the lowest) cover the various levels, which ranged from easy questions with one word to more difficult questions with dialogues or some sentences. The Junior STEP test gives a high priority to listening skill. The students would answer the multiple-choice test by listening to a tape and choosing one answer from 3 pictures (or 2 phrases or words).

### **Example of Question (Question No.16)**

The students hear “I usually walk to school.” on the tape, and choose one answer.

*Questionnaire:* The questionnaire consists of the following 4 questions.

1. Have you ever learned or are you learning English at cram schools or English conversation schools?
2. Have you ever lived in a foreign country?
3. Have you ever learned English at kindergarten?
4. How many years have you been learning English at school?

### **Appendix B (Reliability and Validity of the Test)**

We used the Cronbach’s reliability coefficient alpha for estimating reliability because the Cronbach’s alpha is commonly used in the language testing literature. Our tests consists of 33 questions ranging from easy to difficult levels. The value of the Cronbach’s alpha of our test was 0.714.

Figure 4 shows the scores gained for Questions 1 to 20 included in the proficiency test and Figure 5 shows those for Questions 21 to 33. The vertical axis shows the mean score gained by the four groups of participants (A+, A, B+ and B). The notation A and B indicates the name of schools, and + indicates the experience. For example, Group A+ represents the group of students in School A, who have experience studying English outside of school. Group A represents the group of students in School A, who do not have any experience studying English outside of school. The horizontal axis shows the question numbers of the test. As the same tendency can be seen almost all of the questions, we also can regard this test is reliable from this point.

According to the Ministry of Education, Culture, Sports, Science and Technology of Japan, ‘In elementary schools, it is essential that emphasis be placed on English terms that students encounter in their daily lives’ (2001). It also mentions that elementary school “English activities” focus on the listening and speaking of simple English terms that students know from their daily lives.

Because our test, which is mainly a listening test, set out to measure the student knowledge of vocabulary, expressions and simple conversation related to children’s life at home, at school, at their play world and society, our test has reasonable validity as an English proficiency test for elementary school students.

## References

- Akao, F. (ed.) (1994). *Jidoeiken Challenge Book*. Obunsha.
- Alwin, D. (1976). Assessing school effects: Some identities. *Sociology of Education*, 49, 294-303.
- Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. R. Statist. Soc. A* 128, 234-55
- Coleman, J. S., Hoffer, T., and Kilgore, S. (1982). *High School Achievement*. New York: Basic.
- Dehejia, R. H., and Wahba, S. (1998). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Ass.*, 94, 1053-1062.
- Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998), Characterizing selection bias using experimental data. *Ecomometrica*, 66, 1017-1098.
- Hong, G., and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Ass.*, 101, 901-910.
- Katsuyama, H., Nishigaki, C., and Wang, J. (2006). A Study on the Effect of English Teaching in Public Elementary Schools. *KATE Bulletin*, 20, 113-124.
- Ministry of Education, Culture, Sports, Science and Technology (2001). *Practical Handbook for Elementary School English Activities*. Kairyudo Publishing Co.,Ltd.,

124, 189.

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of Education*, 74, 341-374.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1999). *Applied Linear Regression Models Third Edition*. New York: McGraw-Hill.

Ōtsu, (2004). 小学校での英語は必要か Keio University Press, Inc.

Ōtsu, (2005). 小学校での英語教育は必要ない Keio University Press, Inc.

Raudenbush, S. W., and Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 307-335.

Rosenbaum, P. (1984a). From association to causation in observational studies: The role of strongly ignorable treatment assignment. *J. Amer. Statist. Ass.*, 79, 41-48.

Rosenbaum, P. (1984b). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Statist. Soc. A*. **147**, 656-666

Rosenbaum, P., and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55

Rosenbaum, P. (2002). *Observational Studies*. , New York: Springer-Verlag.

Winship, C. and Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659-706

Sen, A. and Srivastava, M. (1990). *Regression Analysis*. New York: Springer-Verlag.

Takagi, A. (2003). The Effects of Language Instruction at an Early Stage on Junior High School, High School, and the University Students' Motivation towards Learning English, *ARELE*, 14, 81-90.