Identifying Disadvantaged Schools Using a Data-Driven Approach

By STACEY H. CHEN, YU-KUAN CHEN, TIAN-MING SHEU, AND HUEY-MIN WU*

Subsidizing disadvantaged schools requires a systematic and easy-to-update method for identifying them. Existing methods impose subjective thresholds on a few geographic or population variables, and typically omit other socioeconomic factors relevant to policy-making, despite data availability. Revisions to those thresholds, rare and ad hoc, demonstrate significant policy implications on resource reallocation. We propose identifying disadvantaged schools by mapping an extensive set of covariates onto a scalar and estimating a "disadvantaged score" using a selection model and existing classifications. In an empirical application, the proposed method replaces 28% of the official list of disadvantaged campuses, shifting the disadvantaged status to schools deeper in the mountains, farther away from train stations, having higher teacher vacancy rates, and located in poorer minimally populous areas. The campus- and district-level variables used are publicly available and periodically updated in most advanced economies, and the statistical implementation is straightforward.

* Chen: Associate Professor, GRIPS, Tokyo, Japan 106-0032, (e-mail: s-chen@grips.ac.jp). Y.K. Chen: Ph.D. student, Rice University, Houston, Texas 77005, (e-mail: yc93@rice.edu). Sheu: Professor of Education, National Taiwan Normal University, Taipei, Taiwan 106, (email: behappy@ntnu.edu.tw). Wu: Associate Research Fellow, National Academy for Educational Research, New Taipei City, Taiwan 23703, (email: whm@mail.naer.edu.tw). Thanks to the county bureaus of education and the Department of Statistics at the Ministry of Education for approving the use of data for this study. Thanks also the seminar participants at the University of Tokyo, the Trans-Pacific Labor Seminar, and the Osaka Workshop on Economics of Institutions and Organizations for helpful comments. We acknowledge the funding from the JSPS Kakenhi Grant Number JP 17H02537. All remaining errors are ours.

1. Introduction

Researchers and policymakers who aim to alleviate disparities in resources among public schools must begin by measuring how disadvantaged a given school is. For developed countries like the U.S., the major challenge currently facing public schools does not necessarily lie in securing more funding, but in the need for better allocation policies (OECD 2015). Other areas that perform best in PISA (the Program for International Student Assessment) are also dangerously unequal across schools, e.g., Belgium, Germany, Poland, Shanghai, and Taiwan (OECD 2012). Although the issue of resource allocation among public schools is pressing, surprisingly few systematic methods are available for identifying disadvantaged schools. The present study fills this research gap.

Existing methods for identifying disadvantaged schools impose subjective *thresholds* on a small set of geographic or population variables while omitting socioeconomic backgrounds (e.g., percentages of low-income families). For example, the National Center for Educational Statistics (NCES) in the United States identifies rural schools based on thresholds on distances (25 miles from an urbanized area and 10 miles from an urban cluster). Admissions to support programs are then determined by locale categories and a threshold on population density (e.g., fewer than ten persons per square mile in counties for the Small, Rural School Achievement Program). These thresholds are occasionally revised to reflect progress in aging and urbanization. However, because of the discontinuity of threshold-setting, revisions do not necessarily justify a reallocation of resources.

Designating disadvantaged schools by subjective thresholds is particularly problematic when we compare disadvantage levels of school campuses across government agencies that have been using different thresholds. Before 2018, for instance, local county offices in Taiwan separately identified geographically disadvantaged schools, without a set of shared standards. Consequently, a disadvantaged school in one county might be rated otherwise according to the standards employed by another county, although the funding for disadvantaged schools all come from the central government.

In this study, our recommendation for tackling the issue of subjective/inconsistent thresholds is straightforward. If certain government agencies have rated a given school as disadvantaged, then other schools with the same need for additional support (characterized by the same observable

attributes) should receive the same rating, measured along a continuous score. After re-ranking and re-ordering according to the new rating, schools previously rated as disadvantaged do not necessarily remain on the list, while other schools previously out of the list now might be included. This simple, intuitive approach is readily applicable to other contexts, such as identifying rural/disadvantaged hospitals.

Precisely speaking, we estimate each public school's probability of being labeled as disadvantaged using a standard selection model, with the current and most stringent rating as the dependent variable. We assume that the most stringent designation of schools as being disadvantaged, which is informative but not entirely accurate, is determined by a latent index crossing an empirically determined threshold. Here the latent index captures the net social benefits of the given school gaining a disadvantage rating, while the empirically determined threshold is a linear function of an extensive set of the geographic, population, and socioeconomic covariates at the district or school level, with coefficients estimated by a standard statistical procedure, such as Probit.

It is noteworthy that the model excludes endogenous covariates (e.g., school facilities, teacher turnover, and student test scores) that can induce behavioral changes from teachers or school principals. Moral hazard problems might arise if poorer facilities, higher turnover rates, or lower test scores could help school principals receive more subsidies from the government through obtaining a disadvantage rating. For the same reason, we also recommend the exclusion of any self-reported variable from the model specification.

Essentially, our proposed approach projects the observed covariates from data onto a continuous propensity score, which we call the "disadvantage score," defined by the estimated probability of being selected into the disadvantage classification for each school campus. Policymakers can further classify the list of disadvantaged schools through setting *ex post* cutoffs according to their budget constraints along the estimated score.

The proposed method is straightforward, making it possible for frequent adjustments to take increasing urbanization and aging of the population into account. Unlike existing classification methods, which are often based on the decennial census and updated every ten years or more, our proposed approach relies on annual data on schools and districts, thereby making frequent revisions more feasible.

The Taiwanese government began adopting our recommended approach in 2017 to construct disadvantage scores for public elementary and middle schools and enacted the proposed estimation model into law in 2018. The Taiwanese case is great for illustration because the official classifications involve only two indicators, namely, the disadvantaged and the highly disadvantaged campuses. By contrast, the ratings in other countries typically contain multiple ordered categories, to which our method is still applicable, albeit less straightforward for demonstration. We reclassify all the school campuses into "disadvantaged," "highly disadvantaged," or "not disadvantaged" categories by using the estimated scores and maintaining the same proportions as in the previous classification. Our method replaces 28% and 35% of the previous official list of disadvantaged and highly disadvantaged campuses, shifting the disadvantaged status to campuses that are deeper in the mountains, having worse amenity and higher teacher vacancy rates, distant from the county office (which provides education resources sponsored by the central government) and train stations, and located in the poorest and least-populous areas.

Our strategy in constructing the disadvantage score for each school campus is related to the previous literature on "comparable worth" (Ehrenberg and Smith, 1987; Sorensen 1990; Baker and Fortin, 2004). The principle of comparable worth asserts that individuals, who perform the same jobs with identical observed characteristics, should receive equal pay. Similarly, schools of the same type, sharing identical observed geographic and demographic characteristics, should obtain the same rural rating. Our proposed method uses the most stringent indicator for being at-risk or disadvantage is closely related to the comparable worth literature, which uses continuous wage as the dependent variable. Our method is also related to recent studies that explore methods for assessing regional human welfare using remotely sensed data and extensive sets of covariates. For example, Blumenstock (2016), Jean et al. (2016), and Engstrom, Hersh, and Newhouse (2017) used satellite images to predict poverty rates across small areas by identifying a comprehensive list of spatial characteristics, such as paved roads, metal roofs, and nightlight; Mullainathan and Spiess (2017) provides an overview of this literature. Their method follows a similar framework of appraising local features using an extensive list of observables, although the role of unobserved social benefits from identifying the poverty level, which drives our main statistical argument, are not directly considered in that literature.

The remainder of this paper is organized as follows. Section 2 describes the target empirical settings for our proposed method. Section 3 lays out a data-driven approach to classifying schools and presents a model that motivates the construction of the disadvantage score. Section 4 illustrates the method by applying it to Taiwanese school data. Section 5 concludes.

2. Target Empirical Setting

A. The NCES Locale Code

The U.S. Department of Education runs the Rural Education Achievement Program (REAP) initiatives, which were designed to help rural districts that may not be able to compete effectively for Federal funds. The REAP includes both the Small, Rural School Achievement (SRSA) program and the Rural and Low-Income School (RLIS) grant program. The SRSA finances rural Local Education Agencies (LEAs) directly by funding initiatives aimed at improving student achievement, while the RLIS awards formula grants to State Education Agencies (SEAs), which then award subgrants to eligible LEAs. Eligibility to both programs is conditional on the status of schools served by the LEAs according to the NCES locale codes. The SRSA also considers population density, the number of students, and whether an LEA is determined to be "rural" by another government agency, while the RLIS takes the ratio of children ages 5 through 17 years from families that live in poverty into account.

These conditions for eligibility are largely based on location, as the NCES urban-centric locale codes eligible for the REAP indicate whether the location is determined to be an "urban cluster" or "rural territory" in the Census, and its distance from "urbanized areas" and "urban clusters." Although poverty ratios, population densities, and school size are also considered, they enter the decision process as thresholds for setting eligibility in the LEAs. The use of thresholds could potentially overlook much of the school-level variation in the available data. The method we propose incorporates a richer set of information into a simple statistical model that could aid the effective allocation of REAP funds or similar resources, which we demonstrate with an application to public schools in Taiwan.

B. Disadvantaged Taiwanese Public Schools

Before 2018, the Ministry of Education in Taiwan allocated special funds to schools that are either "disadvantaged" or "highly disadvantaged," with the status determined by county education bureaus. The thresholds for being "disadvantaged" or "highly disadvantaged" were chosen by each county individually, primarily based on campus elevation, access to transportation services, or the percentage of indigenous students (Tsai and Wang 2016). In 2015, approximately one-third of public schools were officially labeled as *disadvantaged* and 4% as *highly disadvantaged*. Some of these schools and others that are in need but have not gained a disadvantage rating struggle to fill vacancies in operations (Liu and Chiang 2013) and teaching (Fan and Chang 2015; Tsai and Wang 2016). Policies for either closing these schools or attracting teachers and principals with higher pay and faster promotion are subjects of intense policy debate (Wang and Chen 2007; Cheng, Chan, and Huang 2008).

Given the challenges these schools face, identifying schools that have the greatest need for the extra funds is a policy imperative. However, instead of an accurate measurement of how disadvantaged schools are, the previous system with locally determined school status had created incentives for setting thresholds that aid competition of funds. The fact that there are changes to the status of being disadvantaged between school years, while spatial characteristics --- which government agencies claim to have based their decisions on --- are fixed in our sample years, shows that discretion likely plays a role (Tsai and Wang, 2016). Our empirical approach, described in the next section, addresses the issue with inconsistent standards of disadvantage rating across counties by generating a *disadvantage score* for each school using a standard selection model.

3. A Data-Driven Approach

A. Generate the Disadvantage Score for Each Campus

Our empirical goal is to construct a measurement of how disadvantaged a school is, for which we employ a standard selection model. The model maps a comprehensive set of covariates from data onto a continuous measure for each campus' disadvantage level.

Although biased and problematic, the currently available ratings offer valuable information for conceptualizing the degree to which schools are disadvantaged and for reflecting the *unobservable*

characteristics that are not captured by the included explanatory variables. To minimize potential bias in the current ratings, we use the *most stringent* version of the existing classifications (or the intersection of all available ratings) as the reference point and the dependent variable of the selection model. Admission to this category often entails greater government expenditures in most settings; therefore, the allocation of this status is likely the most careful and least biased.

For the explanatory variables, we include an extensive set of covariates on school characteristics, local demographics, and socioeconomic status, in addition to standard criteria (e.g., geographical and population factors). Covariates are chosen to reflect the need for subsidies.

We consider a model where the government has classified school campus *s* as one of the *highly disadvantaged* (indicated by $D_s = 1$) because the net social benefit from campus *s* gaining such a rating exceeds 0. We measure the net social benefit by the sum of an observed component and a latent component e_s:

$$D_s = I\{X_s\delta + e_s > 0\}, \tag{1}$$

where $I\{\cdot\}$ represents the indicator function. The set of covariates X_s includes school attributes, campus location, district amenities and characteristics, and the constant term. The vector δ contains all coefficients. The key assumption of our approach is that the observable component $X_s\delta$ fully captures the school campus's need for a subsidy due to being disadvantaged geographically or socioeconomically. Although our approach requires additive separability, a latent index with additive separability between observed and unobserved variables can represent a non-additive latent index under regularity conditions (Vytlacil 2006).

Even if the government has classified school *s* as one of the highly disadvantaged ($D_s = 1$), school *s* does not necessarily need the extra financial and operational support that comes with such a rating, since unobservable factors in e_s might be *unrelated* to the need for subsidies. For example, transitory shocks in the local economy, local politics, residents' attitudes towards education, or school principals' bargaining power with the government could have played a role in the decision process. County-fixed effects or year-fixed effects can be employed to capture these unobservable or unmeasured factors. Indeed, many county-fixed effects, if included in the model, have large and significant coefficients, suggesting that existing classifications have overestimated the difficulties faced by schools in some counties and underestimated for some others. To construct

an objective disadvantage score that can be entirely justified by *observables* related to the need for subsidies, we recommend policymakers exclude any fixed effects from the model.

Assuming the latent index is a standard normal random variable, we can then use a standard probit model to estimate the propensity score, which we call the *disadvantage score,* denoted by p_s for each school *s*,

$$p_s = \Phi(X_s \delta), \tag{2}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. If the most rigorous rating and the model specification are correct, then the estimated propensity score is a valid measure for the degree of disadvantage for each school campus. The proposed method ensures those campuses with the same observed characteristics X_s, which indicate the same observed need for additional support, have the same disadvantage score. Policymakers can utilize the disadvantage scores in prioritizing targets for subsidies and designate additional support for the highly disadvantaged campuses.

It is possible that the existing classification is extremely biased, to the degree that the sign of the estimated coefficient on a certain covariate is opposite to policy intentions. For example, in our application to schools in Taiwan, school campuses in an indigenous district are known to be disadvantaged due to cultural differences and economic inequality, which fully merit supporting them with the additional resources that come with high disadvantage scores. Such normative considerations would require the coefficient on the indicator for a school being in an indigenous district to be positive. However, our estimation result shows otherwise. This counterintuitive result reflects Taiwan's unique historical context. Pre-WWII conflicts between indigenous residents and the Japanese colonial administrations had resulted in mislabeling and underinvestment in school campuses in indigenous districts (Oe and Kobayashi 1992; Takeshi et al. 1997). Post-WWII administrations followed previous policies and continued to label these districts as indigenous and mislabel their school campuses as not being disadvantaged, leading the coefficient of indigenous districts in our model to be negative conditional on other covariates. The negative coefficient would systematically lower their disadvantage ratings, contrary to what the observed socioeconomic variables would suggest. Such historical errors in the existing classifications can be addressed by either dropping the indigenous district indicator from estimation or by constraining the coefficient to be positive while estimating the probit model. As we observe a

comprehensive set of variables that could capture challenges in education that stem from local socioeconomic or demographic conditions, if the school campuses in indigenous districts share similar conditions with highly disadvantaged schools outside of the indigenous districts, leaving the indigenous indicator out would not significantly change the final allocation of resources.

B. Measuring the Degree of Misclassification

To measure the gap in classifications between the existing system and our proposed method, we calculate the degree of misclassification using information on the fraction of campuses that are mislabeled using information about Type-1 and Type-2 errors in two ways. First, we take the existing/official grouping percentages as given and compute the *mislabeling percentage*.

$$\text{MP1} = \sum_{s=1}^{n} [I_{1s} + I_{2s}]/n. \tag{3}$$

Here I_{1s} indicates Type-1 error, where a campus *s* that is determined to be disadvantaged in the selection model has been mislabeled/underrated otherwise in the original classification, while I_{2s} indicates Type-2 error for a campus *s* that is not disadvantaged in the selection model but has been mislabeled/overrated. Holding the number of disadvantaged campuses constant, the number of campuses with Type-1 errors necessarily equals the number of campuses with Type-2 errors (with differences due to rounding errors), and 0.5×MP1 indicates the fraction of campuses that is mislabeled as being disadvantaged in the original classification and replaced by other more qualified campuses for a disadvantage rating, according to the order of their disadvantage scores.

Analogs to MP1 have appeared in the previous literature, such as Park, Wang, and Wu's (2002) "targeting income gap," designed to evaluate targeting effectiveness. Like MP1, the target income gap also aggregates the percentage of entities with either Type-1or Type-2 error, using equal weights. Alternatively, the percentage of campuses with Type-1error might deserve greater weight than those with Type-2error, as Cornia and Stewart (1993) have suggested. However, how social weights vary across campuses ultimately depends on subjective assumptions about social welfare functions.

When the number of the official grouping exceed two, the new method can downgrade or upgrade campuses by more than one disadvantage levels. We further consider the magnitudes of downgrading and upgrading in another mislabeling percentage measure, MP2. Unlike MP1's

assigning a value of one to both downgrading and upgrading, *the degree of mislabeling* MP2 is designed to capture variation in the degree of mislabeling, as indicated below,

$$\text{MP2} = \sum_{s=1}^{n} [NI_{1s} + NI_{2s}]/n. \quad (4)$$

Here NI_{1s} is the number of disadvantage levels that the disadvantage score upgrades campus *s* from the existing level, while NI_{2s} denotes the number of disadvantage levels downgraded that campus *s* receives in our method. If the previous grouping method upgrades or downgrades only a few campuses by multiple levels, MP2 will almost equal MP1, as illustrated in our empirical example below.

B. Determining the Number of Categories by Optimal Clustering

When managing the allocation of public resources, policymakers often need to classify school campuses by the degree of disadvantage. However, the number of classifications is typically undetermined, and controversies arise if school campuses in different classifications are too similar in attributes to justify separate categories and unequal provisions. To address this practical issue, we recommend using the *k*-means cluster algorithm (Hastie, Tibshirani, and Friedman 2009, chapter 14) in grouping school campuses by similarities in their disadvantage levels.

When the number of clusters or classifications *k* is unknown, the *k-mean* cluster algorithm aims to partition *n* school campuses into *k* clusters or classifications, in which each campus belongs to the cluster with the nearest mean disadvantage score. The algorithm utilizes an iterative refinement in two repeated steps until convergence: (1) assign each campus to the nearest cluster (determined by the smallest distance to the cluster mean); (2) update the cluster mean in the new clusters. We adopt Makles' (2012) Stata program for implementing optimal clustering.

4. Empirical Application

We apply the proposed method to constructing the disadvantage scores for Taiwan's public elementary and middle school campuses. Compared with the rating methods in other countries with their multiple categories, the Taiwanese case is especially straightforward for illustration, as the official classification system only involves two classifications. Section 4.A describes the data,

Section 4.B explains the method for constructing the disadvantage score, and Section 4.C shows that the implied reclassification outperforms previous official ratings.

A. Data and Institutional Details

We assemble a dataset on all public elementary and middle school campuses in Taiwan for academic years of 2015 to 2017. We illustrate our main method in detail using data from 2015 because summary statistics and the resulting disadvantage scores show similar patterns across the years. To draw policy recommendations, we exclude private, experimental, and schools located on islands located off the main island (the last group belongs to the disadvantaged category with no controversy). Also, we count school branches separately from the main campuses, leading to 3,271 school campuses in our data.

Our master datafile is the universal list of public elementary and middle school campuses between 2012 and 2015, including main campuses and their branches, downloadable on the website of the Ministry of Education along with and their addresses. The administration assigns and publishes disadvantaged and most-disadvantaged ratings annually. We link these ratings to the master file by the school identifier and the branch indicator. We further include a list of pre-determined covariates to capture the variation in official ratings across campuses.

In Table 1, we summarize the characteristics of public elementary and middle school campuses by the previous rating. Officially assigned disadvantaged school campuses represent approximately 32% of the school campuses, and this ratio remains nearly constant between 2013 and 2015. Among the disadvantaged campuses, local governments further determine which campuses are highly disadvantaged according to geographical characteristics, the percentage of indigenous students, and the percentage of teacher vacancies. The highly disadvantaged campuses constantly comprise 4% of the school campuses throughout this period.

Because of Taiwan's mountainous geography and an obsolete policy of ensuring the presence of elementary schools in as many villages as possible since WWII, the government has placed many elementary schools in sparsely populated mountainous areas. For example, there are 29 elementary school campuses with elevation above 1,000 meters. Since middle schools number considerably fewer than elementary schools, middle schools are often located closer to urban clusters.

The administration requests that each campus's geographical location, indigenous student proportion, and hard-to-staff measure to be included for setting the disadvantage rating. Table 1 appears to reflect this practice. The elevation, distance to the nearest train station, and percentage of indigenous students are constantly more than double and quadruple in highly-disadvantaged campuses than in disadvantaged and non-disadvantaged campuses, respectively. While the percentage of full-time teacher vacancy among highly-disadvantaged campuses is double of the percentages in other campuses, the same measure on disadvantaged school campuses exceeds the non-disadvantaged ones by 22% to 27% (0.11/0.9-1 and 0.14/0.11-1).

Additionally, we construct three variables to capture each campus' access to public transportation facilities and other amenities. The first two are the driving distance to the nearest train station and the station's size level. Train stations are sorted into six levels in Taiwan in descending order, with Level 0 at the top indicating major railway hubs. The third variable is the district *amenity index*, which we calculate by summing the percentiles of three variables: the number of 4G transmitter stations, the number of post offices, and the number of convenience stores in the school district. By construction, the amenity index ranges from 0 to 1.

The incorporation of district-level demographic and socioeconomic variables is essential because the ultimate policy objective is to equalize educational opportunity across districts. Table 1 shows that the average percentage of students from a low-income family is highest among campuses with the highly disadvantaged rating; it is more than double of the percentage in disadvantaged campuses, and almost five times of that among non-disadvantaged schools.

We include population factors measured one year before the school year, such as the population density at district levels and the average dependency ratio of preschool children (aged 0–5) relative to the working age population (aged 15–64) among all the school villages of each campus. Also, we consider the population density of the district in 1921 when the indigenous districts were first labeled in Taiwanese history. Table 1 displays that highly disadvantaged campuses are situated in the least populous districts; the population density was 37 to 42 per square kilometer in 1921, which is only 7% to 8% of the level in the districts where the non-disadvantaged campuses are located. Strikingly, after nearly one century, the same ratio drops to as low as 1% to 2% in 2015, indicating polarized urbanization.

Because Taiwan has been experiencing urbanization along with a fertility drop over the century, both preschool children and working adults have been decreasing in number in rural areas at about

the same rate. Consequently, the dependency ratios of preschool children to the working age population across different school disadvantage levels are approximately in the same ballpark, ranging from 7% to 9%.

We intentionally exclude variables that are results of decisions by households or school principals, such as enrollment, land price, teacher vacancy ratios, and the distance to a bus station, to rule out possibilities for manipulation in constructing the disadvantage scores.

B. Marginal Effects of Covariates on the Disadvantage Score

The key idea of identifying disadvantaged school campuses is to exploit the information about the existing ratings and the covariates that can explain the most stringent rating, i.e., the official status of being a highly disadvantaged campus. As the government rates less than 5% of school campuses as highly disadvantaged and because the linear probability model performs poorly with low frequency events, we estimate probit regressions of the most-disadvantaged dummy defined by the previous rating system upon the covariates listed in Table 1. We report the estimated marginal effect of each covariate in Table 2.

Column 1 of Table 2 documents the new official model that the Taiwanese government enacted into law in May 2018 for determining the disadvantaged school rating in the next school year, 2019. The result suggests that every one percent increase in driving distance from the school campus to the county office is associated with a seven percentage points increase in the likelihood of being rated as highly disadvantaged. This coefficient is about nine to 11 folds of what the elevation and the distance to the town office can influence the new official rating probability. Also, a higher amenity index by one percentage point is associated with a six to eight percentage points decrease in the new rating probability. Each of the estimated marginal effects is statistically significant at conventional levels.

The percentage of low-income families on campuses also serves as a significant predictor for the disadvantage rating; a ten percentage increase in the fraction of students from families with low income accompanies an increase of almost one percentage point in the probability of receiving the rating of being highly disadvantaged. This effect is considerable because it accounts for about 20 percent (=0.01/0.05) of the unconditional probability of being rated as a highly-disadvantaged campus. Contrary to what one might expect, population density has almost no impact on the rating

probability when we also control for geographic factors, an index for public amenities in the district, and the percentage of students from low-income families.

In columns 2 to 4, we explore additional covariates related to public amenities and population factors to examine the sensitivity of the probit model. In Column 2, we replace the district population density in the previous school year with the one measured in 1921 when the indigenous districts were initially labeled. The idea of including this unique variable is informed by the historical context of the official designation of disadvantaged campuses, as discussed in Section 3.A. As shown in Column 2, this variable is significantly related to the dummy for the highly disadvantaged status. A ten percent increase in the population density in 1921 is associated with a four percentage point decrease in the probability for being rated as a highly disadvantaged campus, attesting the power of path dependence in policies that categorize schools. Column 2 also includes the average dependency ratio of preschool children to working population among the school villages of each campus. We use the dependency ratio to capture the rapid development of population aging in Taiwan, where birth rates have declined significantly in the past two decades. A one percent increase in this dependency ratio is associated with a 1.7 percent decrease in the rating probability, suggesting disadvantaged campuses tend to be in areas with dwindling children population.

We repeat a similar exercise with current population density, population density in 1921, and the dependency ratio in Columns 3 and 4, where we further include the driving distance to the nearest trains station, interacted with the level of the train station. The reference level of train stations is Level 0, indicating a major railway hub. The coefficient on the distance to the nearest train station is insignificant or marginally significant if the station is Level 0. In contrast, all the interaction terms are strongly significant in both Columns 3 and 4, with similar magnitudes. A one percent increase in the driving distance from acampus to its nearest train station, when compared against campuses with its closest train station being a Level 0 major station, is associated with an increase in the probability that ranges from 1.5 to 1.7 percentage points.

The coefficients on the other variables see some changes after including the distance to train stations and their interaction terms. However, both the signs on the coefficients and the patterns of statistical significance remain the same, and the confidence intervals of these variables overlap between Columns 1 and 3, and Columns 2 and 4, except for the coefficient on population density. Given that population density is not a robust predictor of being rated as highly disadvantaged after

conditioning on location and demographic variables, Column 4 is our preferred specification and the focus of our following analysis.

In Column 4, upon including the distance to the nearest train station and its interaction terms, the coefficients on all the other variables related to geography decreased, with a one percentage point increase in the distance to the county office, the town office, and the elevation of the campus now associated with 5.2, 0.5, and 0.5 percentage points increases in the rating probability, respectively. The coefficient on the amenity index also decreased in magnitude, with now a one percentage point increase being associated with a 5.7 percentage point decrease in the rating probability. The proportion of students from low-income families sees a slight increase in its coefficient, rising to an 11 percentage points increase for every one percentage increment in the proportion. The coefficients on the population density in 1921 and the dependency ratio both increase in magnitude, with a one percentage point increase now associated with decreases of 0.5 and two percentage points in the rating probability, respectively.

Next, we compare the results from applying the disadvantage score and the previous official classification. The disadvantage scores used in the next subsection are calculated with our preferred specification in Column 4 of Table 2. We have explored the results from further including a set of covariates of education backgrounds of adults and indicators for students potentially being socially disadvantaged, such as the absence of parents. However, none of these additional variables are statistically significant after conditioning on the variables used in our preferred specification, neither do they change the general patterns of signs or statistical significance in Columns 3 and 4 of Table 2. Therefore, to keep the statistical model parsimonious and its ease of implementation, we exclude these additional variables from our final analysis.

C. Changes in Classification Results

In this subsection, we contrast the campus attributes of the previous and recommended rating methods within the same disadvantage category. To compare methods, we maintain the identical portions of disadvantage labels across categories, as in the previous classification. We first derive each campus' disadvantage score using the preferred specification (Column 4 of Table 2), rank the campuses by the scores, and then assign disadvantage statuses according to the previous portions.

Table 3 reports changes in the mean characteristics of campuses by disadvantage levels between the previous classification and our recommended rating method.

We observe marked shifts in variables that capture the geographic characteristics of school campuses. School campuses that are determined to be non-disadvantaged are now located at lower elevations and closer to county offices and train stations. The category of highly disadvantaged schools, on the other hand, sees a large increase in these characteristics. Highly disadvantaged elementary school campuses are now 22 percent higher in elevation, and middle schools are 48 percent higher. Highly disadvantaged elementary school campuses are 12 percent farther away from their closest train stations, while middle school campuses are 36 percent farther. The largest drop in the amenity index is found among elementary school campuses that are ranked to be disadvantaged, with a 25 percent decrease, followed by a 15 percent decrease for highly disadvantaged middle schools.

Middle school campuses that are determined to be highly disadvantaged show the largest increase in the proportion of students coming from low income families, with an increase of 12 percent, followed by 9 percent increases for highly disadvantaged elementary school campuses and disadvantaged middle schools. However, the mean proportion of low-income families remains constant for non-disadvantaged elementary schools, while disadvantaged elementary school campuses see a drop of 6 percent, and non-disadvantaged middle schools show a slight increase of 3 percent. This pattern suggests that when one characteristic is intuitively linked to being disadvantaged, its effect on the distribution of disadvantage ratings may be non-monotonic and differ across regions once we take other factors into account.

As for the recent population density, although it is not statistically significant and was therefore excluded from our main specification in the probit model, all categories experience large changes. There are decreases at the magnitude of 69 percent and 47 percent for highly disadvantaged middle and elementary school campuses, respectively. For the dependency ratio, even though the signs on its marginal effect are negative in Table 2, suggesting a negative association with being rated as highly disadvantaged, highly disadvantaged elementary and middle school campuses, in fact, show increases in the dependency ratio, with 11 and 13 percent increase respectively. This result also has a policy implication beyond its descriptive aspect; as a higher rating often entails greater share in the extra resources committed to helping disadvantaged schools, a higher dependency ratio in

campuses determined to be highly disadvantaged means that more children will benefit from the extra resources when they reach school age.

To observe what the disadvantage score might show for schools that face challenges in recruiting and retaining teachers, we also report at the bottom of Table 3 the changes in the proportion of teaching positions that are not filled by fulltime teachers and the proportion of fulltime teachers with less than five years of experience in teaching. Disadvantaged elementary school campuses show a nine percent rise in the proportion of teaching positions without full-time teachers, and there is an increase of seven percent for middle schools. As for the proportion of teachers with less than five years of experience, there is a six percent increase across the board for every category of middle schools and a four percent decrease for disadvantaged elementary school campuses.

In summary, Table 3 shows our data-driven approach reassigns the highly disadvantaged and disadvantaged status to school campuses that are up to 48% higher in elevation, 8% to 19% closer to the county office, at least 12% closer to a train station, with a 7% to 26% lower degree of amenity, located in areas that are at least 35% less populous, and with fractions of full-time teacher vacancies that are up to 9% higer. This result indicates that the previous rating system has mislabeled considerably many school campuses.

We summarize the relabeling result in Table 4, which suggests the proposed method upgrades and downgrades 348 and 349 campuses, respectively, among the total of 3,271 campuses. We further quantify the degree of improvement, using the *mislabeling percentages* MP1 and the *degree of mislabeling* MP2 as proposed in equations (3) an (4). We find that the mislabeling percentage is 0.213 while the degree of mislabeling is 0.214. The two measures are very close since only two campuses are upgraded from being non-disadvantaged to highly disadvantaged while only one is downgraded conversely. The original official classification designated 1,060 schools to be disadvantaged or highly disadvantaged, 301 of which are replaced when we apply the disadvantage score, yielding a ratio of 28.4%. The corresponding ratio for being highly disadvantaged is 35%. The 301 replaced disadvantaged schools also make up 13.6% of the schools that were originally determined as not disadvantaged. Incorporating a more extensive set of information thus leads to substantial departures from the classification based on the limited number of factors that can be considered in a threshold-setting system.

5. Conclusion

We propose identifying disadvantaged schools by estimating a standard probit model of the intersection of the previous classifications on a comprehensive set of relevant covariates. This data-driven approach improves the existing classification methods that set fixed thresholds on a limited number of variables. We apply the proposed data-driven method to data from Taiwanese public elementary and middle schools. We reclassify approximately 28% of the school campuses, shifting disadvantaged or highly disadvantaged statuses to schools in the poorest neighborhoods, that are most distant from train stations, located in most mountainous and least populous areas, and have the highest percentage of full-time teacher vacancies.

Although this paper focuses on defining rural schools, our method can be applied to determining disadvantaged hospitals or revising area taxonomies, as researchers such as Hart, Larson, and Lishner (2005) have called for solutions to problems with the discontinuity of the old threshold-setting framework. The disadvantage score can also serve as a more fine-grained alternative to rank-based measures such as the Social Vulnerability Index used by the U.S. Centers for Disease Control and Prevention (CDC) in identifying at-risk areas (Flanagan et al., 2011). The proposed classification method is tractable and easy to update and replicate, given that most of the required data are publicly available in most countries.

One potential drawback is that researchers must rely on their expertise and judgment in selecting covariates relevant to their study. However, selecting the set of covariates is likely less controversial or subjective than imposing subjective thresholds on a few covariates.

References

Baker, Michael, and Nicole M. Fortin. 2004. “Comparable Worth in a Decentralized Labour Market: The Case of Ontario.” Canadian Journal of Economics/Revue Canadienne D`Economique 37 (4): 850–78.

Blumenstock, J. E. 2016. “Fighting Poverty with Data.” Science 353 (6301): 753–54.

Cheng, Tung-Liao, Chih-Yu Chan, and Ping-Der Huang. 2008. “Research for Reconstructing Small Schools in Rural Areas.” College of Education, National Chengchi University and Ministry of Education, Taiwan.

Cornia, Giovanni Andrea, and Frances Stewart. 1993. “Two Errors of Targeting.” Journal of

International Development 5 (5): 459–96.

Ehrenberg, Ronald G., and Robert S. Smith. 1987. “Comparable-Worth Wage Adjustments and Female Employment in the State and Local Sector.” Journal of Labor Economics 5 (1): 43–62.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. 2017. “Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being.” Policy Research Working Papers. The World Bank.

Fan, Chi-Win, and Win-Cheng Chang. 2016. “Teacher Movements in Rural Hualien County and Strategies for Improvement.” Taiwan Educational Review Monthly 4 (6): 74–77.

Flanagan, Barry E., Edward W. Gregory, Elaine J Hallisey, Janet L. Heitgerd, and Brian Lewis. 2011. “A Social Vulnerability Index for Disaster Management.” Journal of Homeland Security and Emergency Management 8 (1).

Hart, L. Gary, Eric H. Larson, and Denise M. Lishner. 2005. “Rural Definitions for Health Policy and Research.” American Journal of Public Health 95 (7): 1149–55.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY: Springer New York.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. 2016. “Combining Satellite Imagery and Machine Learning to Predict Poverty.” Science 353 (6301): 790–94.

Liu, Shih-Min, and Chung-Peng Chiang. 2013. “A Study of the Elementary School Beginning Principals’ Administrative Predicaments and Coping Strategies in Remote Areas of Kaohsiung City in Taiwan.” Journal of Education of National Changhua University of Education 24: 25–49.

Makles, Anna. 2012. “Stata Tip 110: How to Get the Optimal K-Means Cluster Solution.” The Stata Journal 12 (2): 347–51.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” Journal of Economic Perspectives 31 (2): 87–106.

Oe, Shinobu, and Hideo Kobayashi. 1992. “Colonization and Industrialization.” The Iwanami Lecture on Japan and Its Colonies.

OECD. 2014. “PISA 2012 Results in Focus: What 15-Year-Olds Know and What They Can Do with What They Know.”

———. 2016. “Country Note: Key Findings from PISA 2015 for the United States.”

Park, Albert, Sangui Wang, and Guobao Wu. 2002. “Regional Poverty Targeting in China.” Journal of Public Economics 86 (1): 123–53.

Sorensen, Elaine. 1990. “The Crowding Hypothesis and Comparable Worth.” Journal of Human Resources 25 (1): 55–89.

Takeshi, Komagome, and J. A. Mangan. 1997. “Japanese Colonial Education in Taiwan 1895-1922: Precepts and Practices of Control.” History of Education 26 (3): 307–22.

Tsai, Pei-Tsun, and Mei-Yu Wang. 2016. 105 Education Investigation 0003. Control Yuan, Taiwan.

Vytlacil, Edward. 2006. “A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results.” Oxford Bulletin of Economics and Statistics 68 (4): 515–18.

Wang, Li-Yun, and Hsiao-Lan Sharon Chen. 2007. “Models of Equal Educational Opportunity Policies for Rural Areas in Taiwan: Synthesis and Reflections.” Bulletin of Educational Resources and Research 36: 25–46.

Tables

Table 1. Sample mean and standard deviation of school campus characteristics by the previous, official disadvantage rating

	Elementary Schools						Middle Schools					
	Non-Disadvantaged		Disadvantaged		Most Disadvantaged		Non-Disadvantaged		Disadvantaged		Most Disadvantaged	
Variable	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Control variables:												
Elevation of the campus (m)	66	98	148	233	441	446	55	69	118	179	388	509
Distance to county office (km)	15	13	34	21	69	22	12	12	30	18	68	34
Distance to town office (km)	3.2	2.7	6.7	6.2	17.3	18.7	2.4	2.0	4.3	4.2	11.1	32.6
Distance to nearest train station(km)	7.2	7.3	14.6	13.5	36.9	20.5	5.4	5.3	12.4	10.9	39.7	27.6
Nearest train station's grade (0 to 5)	2.3	1.6	2.9	1.4	3.6	1.3	2.0	1.5	3.1	1.4	3.2	1.5
Amenity index	0.6	0.3	0.3	0.2	0.1	0.1	0.7	0.3	0.3	0.2	0.1	0.1
Percent low-income families	4.2	4.6	9.5	9.8	19.5	16.4	5.6	4.9	11.8	10.2	18.2	12.4
District population density	4913	6795	549	772	94	171	6780	7673	541	777	89	102
District population density, 1921	471	643	148	149	37	40	617	821	161	172	42	49
Dependency ratio by school village	0.08	0.02	0.07	0.03	0.09	0.05	0.09	0.02	0.07	0.02	0.08	0.03
Additional information:												
FT teacher vacancy (%)	0.09	0.08	0.11	0.12	0.20	0.16	0.11	0.08	0.14	0.12	0.14	0.14
FT teacher with less than 5 years of experience (%)	0.16	0.11	0.24	0.18	0.38	0.19	0.18	0.11	0.31	0.17	0.48	0.17
Number of campuses	1720		735		120		491		185		20	

Table 2. Marginal effects of covariates in the probit model

Model	1	2	3	4
Log distance to county office	0.0678***	0.0655***	0.0544***	0.0516***
	(0.00800)	(0.00747)	(0.00755)	(0.00693)
Log distance to town office	0.00608***	0.00637***	0.00450**	0.00489**
	(0.00223)	(0.00220)	(0.00208)	(0.00201)
Log elevation of the campus	0.00781***	0.00741***	0.00653***	0.00492***
	(0.00226)	(0.00211)	(0.00199)	(0.00186)
Amenity index	-0.0831***	-0.0647***	-0.0833***	-0.0572**
	(0.0224)	(0.0218)	(0.0225)	(0.0226)
Log distance to nearest train station			0.00404	0.00484*
			(0.00266)	(0.00271)
Log distance to nearest train station*level1			0.0165***	0.0162***
			(0.00169)	(0.00172)
Log distance to nearest train station*level2			0.0150***	0.0151***
			(0.00168)	(0.00161)
Log distance to nearest train station*level3			0.0159***	0.0152***
			(0.00171)	(0.00170)
Log distance to nearest train station*level4			0.0150***	0.0147***
			(0.00172)	(0.00168)
Log distance to nearest train station*level5			0.0160***	0.0157***
			(0.00161)	(0.00158)
Percent low-income families	0.0912***	0.0883***	0.1190***	0.1110***
	(0.0217)	(0.0210)	(0.0214)	(0.0207)
Log district population density	-0.000260		0.00117	
	(0.00290)		(0.00290)	
Log district population density, 1921		-0.00381**		-0.00473***
		(0.00156)		(0.00156)
Log dependency ratio by school village		-0.0170**		-0.0208***
		(0.00753)		(0.00733)
Number of campuses	3,271	3,271	3,271	3,271

Notes: Robust standard errors in parentheses; symbols for statistical significance: *** p<0.01, **p<0.05, * p<0.1.
Our preferred specification is Model 4 in this table. The Taiwanese government has adopted Model 1.

Table 3. Changes in the mean attributes of school campuses from the previous rating to the recommended rating

	Elementary Schools						Middle Schools					
	Non-Disadvantaged		Disadvantaged		Highly Disadvantaged		Non-Disadvantaged		Disadvantaged		Highly Disadvantaged	
Variable	diff	% change	diff	% change	diff	% change	diff	% change	diff	% change	diff	% change
Control variables:												
Elevation of the campus (m)	-11.74	-18%	3.09	2%	99.21	22%	-3.64	-7%	21.86	18%	185.86	48%
Distance to county office (km)	-2.81	-19%	4.09	12%	5.20	8%	-1.37	-11%	9.31	31%	13.15	19%
Distance to town office (km)	0.11	3%	-0.60	-9%	0.58	3%	0.33	14%	-0.73	-17%	3.68	33%
Distance to nearest train station (km)	-1.57	-22%	2.18	15%	4.45	12%	-0.48	-9%	3.10	25%	14.46	36%
Nearest train station's grade (0 to 5)	-0.11	-5%	0.19	7%	0.04	1%	0.09	5%	-0.06	-2%	0.28	9%
Amenity index	0.03	5%	-0.07	-25%	-0.01	-7%	0.00	0%	-0.07	-26%	-0.02	-15%
Percent low-income families	0.00	0%	-0.61	-6%	1.68	9%	0.15	3%	1.06	9%	2.21	12%
District population density	190.81	4%	-191.39	-35%	-44.12	-47%	-388.72	-6%	-217.59	-40%	-61.41	-69%
District population density, 1921	10.58	2%	-4.38	-3%	-11.98	-32%	-29.02	-5%	-14.72	-9%	-30.15	-71%
Dependency ratio by school village	0.01	13%	0.00	0%	0.01	11%	0.00	0%	0.00	0%	0.01	13%
Additional information:												
FT teacher vacancy (%)	-0.01	-11%	0.01	9%	0.00	0%	0.00	0%	0.00	0%	0.01	7%
FT teacher with less than 5 years of experience (%)	0.00	0%	-0.01	-4%	0.00	0%	0.01	6%	0.02	6%	0.03	6%
Number of campuses	1682		767		126		529		153		14	

Table 4. The joint distribution of previous and recommended ratings

Recommended rating:	Previous rating: Non-Disadvantaged	Disadvantaged	Highly Disadvantaged	Total
Non-disadvantaged	1,910	299	2	2,211
Disadvantaged	300	573	47	920
Highly Disadvantage	1	48	91	140
Total	2,211	920	140	3,271