Research Article https://doi.org/10.12973/ijem.7.4.683

International Journal of Educational Methodology
Volume 7, Issue 4, 683 -696.
ISSN: 2469-9632
https://www.ijem.com/

Synthetic Longitudinal Education Database: Linking National Datasets for K-16 Education and College Readiness

Jaekyung Lee*
University at Buffalo, USA

Joseph Jaeger
University at Buffalo, USA

Received: September 5, 2021 ▪ Revised: October 14, 2021 ▪ Accepted: November 3, 2021

Abstract: What are missing in the U.S. education policy of "college for all" are supporting data and indicators on K-16 education pathways, i.e, how well all students get ready and stay on track from kindergarten through college. This study creates synthetic national longitudinal education database that helps track and support students' educational pathways by combining two nationally-representative U.S. sample datasets: Early Childhood Longitudinal Study- Kindergarten (ECLS-K; Kindergarten through 8th grade) and National Education Longitudinal Study (NELS; 8th grade through age 25). The merge of these national datasets, linked together via statistical matching and imputation techniques, can help bridge the gap between elementary and secondary/postsecondary education data/research silos. Using this synthetic K-16 education longitudinal database, this study applies machine learning data analytics in search of college readiness early indicators among kindergarten students. It shows the utilities and limitations of linking preexisting national datasets to impute education pathways and assess college readiness. It discusses implications for developing more holistic and equitable educational assessment system in support of K-16 education longitudinal database.

Keywords: *College readiness, longitudinal database, machine learning, multiple imputation, synthetic data.*

To cite this article: Lee, J., & Jaeger, J. (2021). Synthetic longitudinal education database: Linking national datasets for K-16 education and college readiness. *International Journal of Educational Methodology*, *7*(4), 683-696. https://doi.org/10.12973/ijem.7.4.683

Introduction

The former United States of America (U.S.A.) President Barack Obama proposed that every American commit to attending at least one year of college in preparation for knowledge-based new economy and that the U.S. reclaim its position as the best educated nation in the world (2009, February 25). Advancing this universal college education agenda further, the current U.S. President Joe Biden has recently made a policy proposal that calls for two years of community college to be free for all Americans (2021, April 28). Further, there has been an "early college" movement that enables students to attain high school diploma and 2-year college degree simultaneously (Berger et al., 2013; Rosen et al., 2020). Although all these initiatives have potential to upgrade American public education and bridge school-to-college and school-to-work divides, what are often missing in this policy debate for college and career readiness are supporting data and indicators that inform how well all students get ready and stay on track for colleges and careers throughout their education pathways. Even if college education would become accessible for all students, some of the key questions are who are at risk of failing to meet the standards for college education and how to assess and improve their college readiness early on (ACT, 2010; Lee, 2012).

Since there remain significant achievement gaps which start very early and exacerbate through schooling, it is crucial to assess and narrow college readiness gaps during early childhood; younger students' skills, attitudes and behaviors are more malleable and early interventions have greater effects (Heckman & Lochner, 2000; Lee, 2016). There is a growing body of research on early indicator/warning systems and interventions that are designed to help students get on track for high school graduation and college readiness (Allensworth & Easton, 2007; Hauser & Koenig, 2011; Neild et al., 2007). However, the dominant model of early indicator/warning systems has focused more on middle/high school dropout prevention and thus identification of potential risk factors such as failing classes and academic disengagement. Although this approach can be useful for monitoring student behavior/performance and detecting high-risk students, it would not help identify low/moderate-risk students earlier when the achievement gaps emerge during preschool or kindergarten.

* **Corresponding author:**
Jaekyung Lee, Department of Counseling, School, and Educational Psychology. University at Buffalo, Buffalo, USA. ✉ jl224@buffalo.edu

Under the Every Student Succeeds Act (ESSA), all states and local districts are expected to redesign school accountability systems and incorporate non-academic outcome measures. However, the current school accountability system remains narrow in its scope by exclusive focus on academic indicators, ignoring other important domains of child development such as socioemotional skills, mental well-being and physical health (Lee et al., 2019; Martin et al., 2016; O'Connell et al., 2009). However, previous studies of educational assessment/accountability systems examined the achievement gap or college readiness gap problems by narrowly focusing on the standardized test scores and school grades of academic achievement rather than taking a whole-child education approach including socioemotional and physical development indicators as well as cognitive development measures (Lee, 2020; Lee & Lee, 2020).

Literature Review

Prior research and discussion on students' readiness for school/college entry or transition has remained largely separate among different levels of education: elementary school readiness (e.g., Hair et al., 2006; Lee & Burkam, 2003), middle/high school readiness (e.g., Eccles et al., 1991; MacIver & Epstein, 1991), and college readiness (e.g., Conley, 2005; Ellwood & Kane, 2000; Kirst & Venezia, 2004). The "P-16" (i.e., preschool through college) or "K-16" (i.e., Kindergarten through college) education policy movement in the U.S. over the past decade seeks to address these broken linkages among different levels of education and improve college and career readiness for all students (National Governors Association, 2007). At the same time, there have been national efforts to support the development of state and local longitudinal education databases that track students' transition over the course of P-16 or K-16 education and beyond (Data Quality Campaign [DQC], 2014).

While the collection and organization of longitudinal student data are most critical for helping meet college readiness goals, building a robust educational database system requires statewide comprehensive and systematic efforts for integrating complex dimensions of longitudinal and multilevel data. According to the DQC analysis of state education databases (DQC, 2014), there are substantial variations among states in terms of their longitudinal education data tracking capacity and many states lack some critical components to ensure effective data use.

On the other hand, there also exist national education longitudinal datasets collected by the National Center for Education Statistics (NCES), a branch of the U.S. Department of Education. The merge of existing national education datasets, linked together via statistical matching and imputation techniques, can offer a promising new way to explore students' trajectories including college pathways, which are often missing and unobservable due to data tracking inhibitors. The current state longitudinal education databases often lack generalizability and applicability beyond their own state boundaries and K-12 time frame. Data privacy and security issues also hamper public access and analysis. Synthetic national education data can help overcome those restrictions and supplement/enrich existing state longitudinal education databases. This study can help fill the gap in existing longitudinal educational databases and inform education policy for improving college readiness.

This study considers a wide range of common variables to link two separate national longitudinal datasets, that is, elementary one (grades K-8) and secondary/postsecondary one (grades 8-12 and college). Demographic and school characteristic predictors of educational attainment have been chosen based on relationships previously observed in the literature. These common variables include socio-economic status (Ladd, 2012; Polidano et al., 2013), gender (DiPrete & Buchmann, 2013; Hedges & Nowell, 1995), race (Bhopal, 2017; Henry et al., 2020), school type (Finn et al., 2002; Jack, 2014), school location and region (Owens, 2010; Sander, 2006). Prior measures of academic achievement and readiness such as the standardized test scores of reading, math and science achievement are also included as they were proven to be strong predictors of educational attainment (Lee, 2016). Further, several variables which measured aspects of socioemotional well-being (Gutman et al., 2003), locus of control (Young et al., 2011), educational aspirations (Froiland & Davison, 2016), time spent on homework (Rau & Durand, 2000), and extracurricular activities (Feldman & Matjasko, 2005) are also included in this study.

The purposes of this study are (a) to develop synthetic national longitudinal education database that helps track students' K-16 education pathways by combining existing nationally-representative longitudinal datasets; and (b) to assess kindergarten students' future college readiness holistically and equitably based on the early indicators of whole-child development (i.e., academic, socioemotional and physical development indicators). This study pioneers the development of synthetic K-16 education longitudinal database by bridging the gap between data silos (elementary education vs. secondary/postsecondary education datasets) and linking them together for college readiness early indicators assessment. Further, we will discuss policy and research implications for developing more holistic and equitable educational assessment system in support of K-16 education longitudinal database.

Methodology

Data Sources

Primary data sources are the public-use version of two preexisting NCES longitudinal education datasets (see https://nces.ed.gov/surveys/), including (1) ECLS-K (Kindergarten through 8th grade) and (2) NELS: 88 (8th grade through college and career).

ECLS-K (Early Childhood Longitudinal Study-Kindergarten): The ECLS-K database provides a nationally-representative sample of kindergartners from the fall of 1998. This study obtained data during the fall and spring of the kindergarten year, and the spring of grades 1, 3, 5 and 8. Sample sizes available for the analysis of the ECLS-K data is 21,409 students.

NELS:88 (National Education Longitudinal Study of 1988): The NELS:88 database is a nationally-representative sample of students in eighth grade in 1988. This study used data from all waves of data collection, grades 8, 10 and 12 and approximately 8 years after high school graduation (i.e., around age 25). Sample size for the analysis of the NELS:88 data is 12,144 students.

Common student demographics and school background variables are the linchpin to match ECLS-K and NELS (see Table 1); NELS:88 is chosen among secondary/postsecondary education datasets because it contains many common 8th grade survey questions with ECLS-K. Variables that appeared in both surveys were used to link the datasets.

Table 1. List of Key Variables in ECLS-K and NELS datasets

	Common Demographics and Background variables	**Early Childhood indicator variables (Kindergarten)**	**Educational attainment and work variables (Age 25)**
NELS	Gender, race/ethnicity, family SES, school type, school location, region, 8th grade reading, math and science achievement, socio-emotional well-being*, locus of control*, educational aspirations*, time spent on homework*, extracurricular activities*	Missing	On-time high school graduation, post-secondary educational attainment, and employment status
ECLS-K		Kindergarten reading and math achievement, fine and gross motor skills, print familiarity, number and shape familiarity, letter recognition, beginning and relative size, healthy weight (BMI), approaches to learning, disability status, socioemotional well-being, intrapersonal/interpersonal skills; internalizing and externalizing behaviors; being in excellent or very good health	Missing

Note. * Variables harmonized across both datasets

Missing Data Imputation for Educational Attainment

If we stack up both ECLS-K and NELS datasets organized by cases (rows) and variables (columns), we are able to find both complete and missing data patterns with N = 33,553 students (see Figure 1). We apply multiple imputation methods to impute missing data for educational attainment in ECLS-K under the assumption of Missing at Random (MAR).

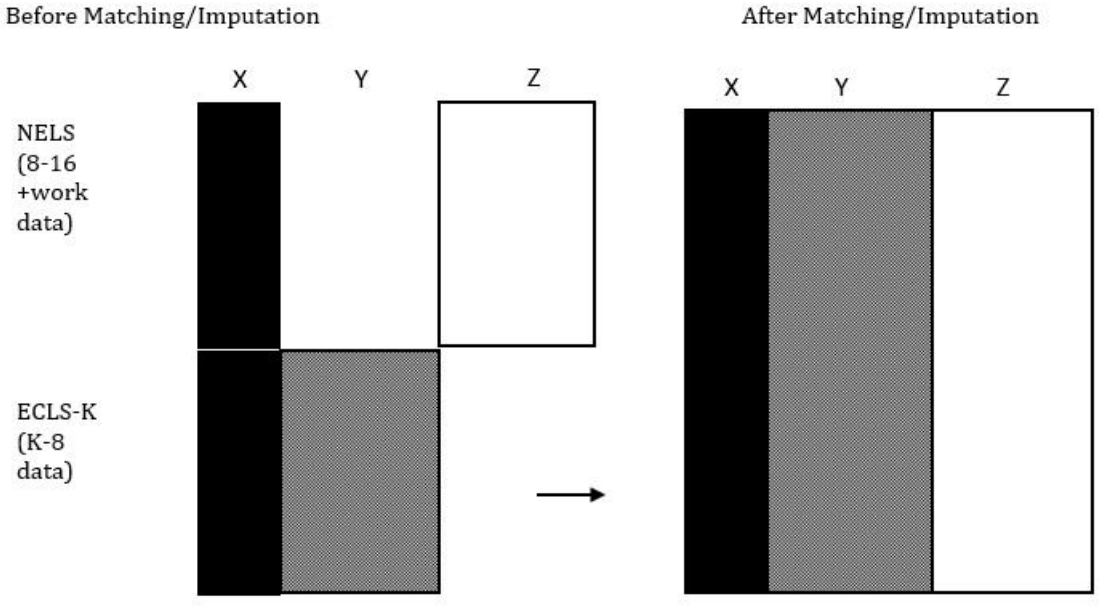

Figure 1. Schematic Illustration of Linking ECLS-K and NELS Datasets

Note.

X = demographics and background variables commonly available in ECLS-K & NELS;

Y = early childhood development indicators available only in ECLS-K;

Z = educational attainment and career variables available only in NELS;

Multiple imputation constructs the posterior predictive distribution of missing data, conditional on observed data, and then a random draw is independently made from this posterior distribution (Carpenter & Kenward, 2013; King et al., 2001; Rubin, 1987). Research demonstrated the superiority of multiple imputation methods over listwise deletion and single imputation methods in terms of reducing bias and inefficiency (Takahashi, 2017). In our case, this strategy is used to create statistically matched files between ECLS-K and NELS that share a common set of variables X. It concatenates the files of ECLS-K and NELS and then multiply imputes values for each missing Y and Z variables based on their relationship with X to reflect uncertainty in the correct value to impute (Rubin, 1987).

We harmonized the common variables X in the above two data sources and compared their marginal/joint distributions, under the assumption that they are representative samples of the same population (D'Orazio et al., 2006). Because they are different cohorts 19 years apart, we made adjustment to data including 8th grade family SES and academic achievement variables, using national trends between 1988 (NELS 8th grade year) and 2007 (ECLS-K 8th grade year) based on the Census and NAEP data.

Then, the following three methods were used for the sake of cross-examination and triangulation to impute missing data for educational attainment variable in ECLS-K: K-Nearest Neighbor (KNN) matching for imputation, Fully Conditional Specification (FCS) for multiple imputation, and Expectation-Maximization with Bootstrapping (EMB) for multiple imputation.

First, K-Nearest Neighbor (KNN) matching is a non-parametric method of imputation by which values are imputed using 'neighbors' or cases identified as similar to those with missing data. The calculated distances of each neighbor are used as weight when averaging values from neighbor cases. Thus, the more similarities that exist between a case and its neighbor for non-missing data, the more weight given to that neighbor during imputation. We used the R package VIM (Templ et al., 2016) to impute missing data for the three dependent variables in the ECLS-K dataset: on-time high school graduation, post-secondary educational attainment, and employment status. A total of thirty neighbors (k = 30) were used in this imputation process.

Second, Fully Conditional Specification (FCS) model is a parametric method for multiple imputation (van Buuren et al., 2006). In this procedure, imputed values are modeled as functions of other available information. Values are imputed many times, creating a series of datasets for analysis. Parameter estimates generated from these datasets are then pooled to create final estimates. Using SAS PROC MI, missing values were imputed fifty times (m = 50) for the dependent variables in the ECLS-K dataset: educational attainment and employment status. The same shared variables used for matching in the KNN imputation procedure were used in the multiple imputation model.

Third, Expectation-Maximization with Bootstrapping (EMB) method is a mixed approach to create multiple imputation for missing data; it combines the Expectation-Maximization (EM) algorithm with the nonparametric bootstrap (Takahashi, 2017). Bootstrap resamples are randomly drawn from the sample data with replacement, and then the EM algorithm is applied to each of these bootstrap resamples to refine parameter estimates until convergence; the Maximum Likelihood Estimates (MLE) from bootstrap resamples is asymptotically equal to a sample from the posterior distribution (Little & Rubin, 2002). Using the R-Package AMELIA II (Honaker et al., 2011), missing values were imputed fifty times (m = 50) for the missing data of educational attainment variable in ECLS-K.

Factor Analysis and Regression Analysis of Early College Readiness Indicators

Finally, the last step was to develop a model of predictors of each outcome, using data from early childhood. In order to test the efficacy of such a model, one outcome variable, educational attainment, was used as a test case. Educational attainment (0 = 'Less than high school diploma or equivalent' , 1 = 'High school diploma or GED', 2 = 'Some postsecondary education (PSE), no degree', 3 = Associates degree', 4 = 'Four year degree or higher'.) was modeled as a function of student/family demographics and a series of possible early education/development indicators selected based on prior research (Hair et al., 2006; National Research Council, 2012). These include dimensions of physical health, social and emotional development, approaches to learning, language development, and cognitive development (see Appendix). Exploratory factor analysis of the ECLS-K data with seventeen indicators measured during Kindergarten has identified three factors of early child development: academic, socioemotional and physical factors (see Table 2). A total of 5,145 cases in ECLS-K had complete data for all of the variables required for this classification analysis.

Table 2. Factor Analysis of ECLS-K Kindergarten Child Development Indicators: Rotated Factor Matrix with Three Extracted Factors and Factor Loadings

Variable Names	Academic Factor	Socioemotional Factor	Physical Factor
Reading T-Score	**.915**	.120	-.011
Math T-Score	**.894**	.142	.141
Fine Motor Skills	**.482**	.149	**.334**
Gross Motor Skills	.167	.083	**.618**
Print Familiarity	**.642**	.109	.083
Count, Number, Shape	**.634**	.117	.222
Letter Recognition	**.879**	.128	.024
Beginning Sounds	**.856**	.112	-.037
Relative Size	**.880**	.140	.138
Healthy Weight	-.021	.010	**.386**
Approaches To Learning	**.359**	**.763**	.125
Externalizing Problem Behaviors	-.038	**-.806**	-.024
Internalizing Problem Behaviors	-.116	**-.462**	-.171
Interpersonal Skills	.150	**.872**	.036
Self-Control	.093	**.909**	-.004
Child's Overall Health	-.123	-.027	**-.439**
Child without Disability	.018	.084	**.579**

Note: See Appendix for the description of variables and factors. The factor loadings with the value of 0.3 or higher are highlighted in bold. Extraction method was principal component analysis, and rotation method was varimax with Kaiser normalization.

To predict the likelihood of educational attainment, multinomial logistic regression models were trained and tested using the sample of ECLS-K cases. The data were randomly divided into a training set (80%, N = 4,117) and a test set (20%, N =1,028), including early kindergarten indicators and demographics as possible predictors of imputed educational attainment as outcome variable; categorical variables are dummy coded and continuous variables are standardized. Regularization was implemented to correct for model overfitting due to the high number of predictor variables (Zou & Hastie, 2005). In order to achieve the best solution, three models were testing using multinomial logistic regression with different regularization methods. This included ridge regression, (variables with weaker relationships are reduced toward zero), lasso regression (weaker variables are reduced to exactly zero) and elastic net regression (some variables are reduced to exactly zero, while others are reduced toward zero). We used the R package *glmnet* (Friedman et al., 2008) to train and test both unpenalized and penalized regression models. The following model illustrates a multinomial logit analysis of educational attainment:

$$Y_{mi} = \alpha\ (\text{Academic Factor})_i + \beta\ (\text{Socioemotional Factor})_i + \gamma\ (\text{Physical Factor})_i + \sum (\text{Background Variables})_i$$

where Y_{mi} is the log-odds of falling into category *m* relative to category *M* for student i; $Y_{mij} = \log(P_{mij} / P_{Mij})$ for which m = 1 for 'High school diploma or GED', 2 for 'Some postsecondary education (PSE), no degree', 3 for Associates degree', 4 for 'Four year degree or higher'. The reference group is 'Less than high school diploma or equivalent'.

For the sake of model performance evaluation and comparison, we used prediction accuracy, deviance (-2 LL), and log loss (cross entropy loss) statistics. For categorical dependent variable like educational attainment, the goal is to have a model that estimates a high probability for the target class (and a low probability for the other classes); we use cross entropy (log loss) as the cost function (Geron, 2017):

$$cross\ entropy = -\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K} y_k^{(j)} \log\left(\hat{p}_k^{(j)}\right)$$

where $y_k^{(j)}$ is equal to 1 if the target class for the jth observation is k and otherwise it is equal to 0.

For the sake of research transparency and reproducibility, we provide online resources including the documentation of data analysis steps and syntaxes (https://doi.org/10.6084/m9.figshare.16863544) along with associated data file used in this study (https://doi.org/10.6084/m9.figshare.16863532).

Findings/Results

Missing Data Imputation for Educational Attainment Outcomes

First, we matched and harmonized common variables between two datasets. We checked whether the common variables in the two data sources have the same marginal/joint distributions. We also checked the patterns of missing data in our merged dataset. Missing data analysis showed that the data missingness pattern is not missing completely at random (MCAR) but can be missing at random (MAR). Logistic regression analysis was conducted to predict missingness indicator for educational attainment variable (i.e., Missingness = 1 for ECLS-K data vs. Missingness = 0 for NELS data) based on all commonly available covariates. It showed acceptable model fit (-2LL = 4734.96, $p < .001$; classification accuracy = 94%; Nagelkerke $R^2 = .86$).

Then KNN matching, FCS and EMB multiple imputation methods were used sequentially to impute the records of missing educational attainment variable among kindergarten cohort in the ECLS-K data. For cross-validation of matching and imputation results, we used a randomly selected 10 percent of NELS data as a subset (N = 1,009) in which the original values of educational attainment were removed and thus pretended missing. In addition to matching and multiple imputation methods, random assignment method was also applied to this subset so that we were able to compare actual vs. imputed values for the same students' educational attainment variable by four different methods.

It turned out that KNN matching, FCS and EMB multiple imputation results have significantly higher accuracy of classification for educational attainment than random assignment (RA) results: FCS (57%) > KNN (55%) > EMB (47%) > RA (32%). While the overall accuracy rate was not very high, it varied substantially among the five different categories of educational attainment: bachelor's degree (86%) > less than high school (85%) > some PSE (56%) > associate's degree (16%) > high school (13%). If those lower-performing categories including high school, some PSE and associate's categories were collapsed into a single category, then the classification accuracy rate would improve substantially: FCS (73%) > KNN (69%) > EMB (65%). FCS multiple imputation performed best among the methods, with accuracy rate gain from 57% to 73%.

Further subgroup analysis of cross-validation by race/ethnicity and parental education level showed similar accuracy rates overall with some notable variations among groups depending on the target levels of educational attainment. Racial minority and disadvantaged student groups tend to have relatively more accurate matching/imputation results for lower education level vs. less accurate results for higher education level; this pattern is an artifact of racial and social inequalities of educational attainment which caused uneven frequency distributions of observed education variable among those subgroups (e.g., more sparse data of Blacks and low-income students with bachelor's degree holders).

Further, we also conducted external validation of these matching and imputation methods by using the Census-based report of educational attainment, which allowed us to compare expected vs. imputed data distributions of ECLS-K cases. ECLS-K kindergarten cohort students who were around age 5-6 in 1998 would become age 24-25 in 2017, which is the same age of students as of NELS final follow-up. According the U.S. Census Bureau, Current Population Survey, 2017 Annual Social and Economic Supplement, educational attainment of the population among 25 – 29 year-old individuals were 7.5% for less than high school, 56.8% for high school/some PSE/associate's degree, and 35.7% for bachelor's degree or higher. Multiple imputation using FCS method produced more favorable results for their educational attainment: 6.6% for less than high school, 43.6% for high school/some PSE/associate's degree, and 49.8% for bachelor's degree or higher. the above Census statistics of educational attainment. The results indicate the possibility of potential upward bias in our imputed ECLS-K data for kindergarten students' future education attainment variable.

Machine Learning Search for College Readiness Early Indicators

Using newly created synthetic national K-16 education data (with educational attainment variable imputed via FCS multiple imputation method), we tested its usability via machine learning models to predict educational attainment based on the early college readiness indicators among kindergarten students. Multinomial logistic regression models with and without regularization methods (i.e., ridge, lasso, and elastic net), were applied to both training data and test data of two different sizes (see Table 3). In terms of our model performance, the results were highly similar among the models; unpenalized regression model produced equally good model fit as penalized models with regularization. This pattern may be related to the prior use of factor analysis which reduced data dimensions and thus prevented potential multicollinearity problems. It is also likely attributable to our data non-sparseness; the sample size used is relatively much larger than the number of model parameters estimated. However, for the smaller size data, penalized regression models produced better goodness-of-fit results (i.e., the smaller values of deviance and log loss) than unpenalized regression model.

Table 3. Multinomial logit model performance statistics with training and test datasets for the prediction of future educational attainment based on ECLS-K Kindergarten child development indicators and background characteristics

	Unpenalized Model			**Ridge Model**			**Elastic Net Model**			**Lasso Model**		
Sample	**accuracy (%)**	**deviance**	**log loss**	**accuracy (%)**	**deviance**	**log loss**	**accuracy (%)**	**deviance**	**log loss**	**accuracy (%)**	**deviance**	**log loss**
Training Data												
(N = 4117)	80	3654.05	0.44	80	3874.86	0.47	80	3654.24	0.44	80	3656.04	0.44
Test Data												
(N= 1028)	79	935.28	0.45	78	1009.88	0.49	79	936.03	0.46	79	937.39	0.46
Training Data												
(N = 1179)	81	1296.03	0.55	81	1408.17	0.60	81	1347.03	0.57	82	1339.26	0.57
Test Data												
(N= 294)	77	309.66	0.53	77	299.45	0.51	77	292.21	0.50	77	291.99	0.50

Note: Model performance statistics including prediction accuracy, deviance statistic (-2 LL), and log loss (cross entropy loss) values are reported for both training and test datasets using unpenalized and penalized (Ridge, Elastic Net, Lasso) models of multinomial logistic regression.

Statistically and practically significant predictors of educational attainment include some key student demographic and geographic background variables (gender, SES, race/ethnicity, school type, region), as well as Kindergarten child development measures of academic achievement, socioemotional well-being, and physical health (see Figure 2). In terms of the strength of relationships (as measured by the size of standardized logit regression coefficients), student demographics and background variables, particularly SES, appear to be more powerful predictors. However, Kindergarten students' academic, socioemotional and physical readiness measures in combination are equally or more powerful in comparison with family SES and race/ethnicity variables. Among those early indicators of educational attainment, academic factor goes first, socioemotional factor second, and physical factor last in terms of relative effect size rank order. Figure 3 visually demonstrates the distribution of imputed future educational attainment values among kindergarten students based on the combination of their academic, socioemotional and physical readiness factors.

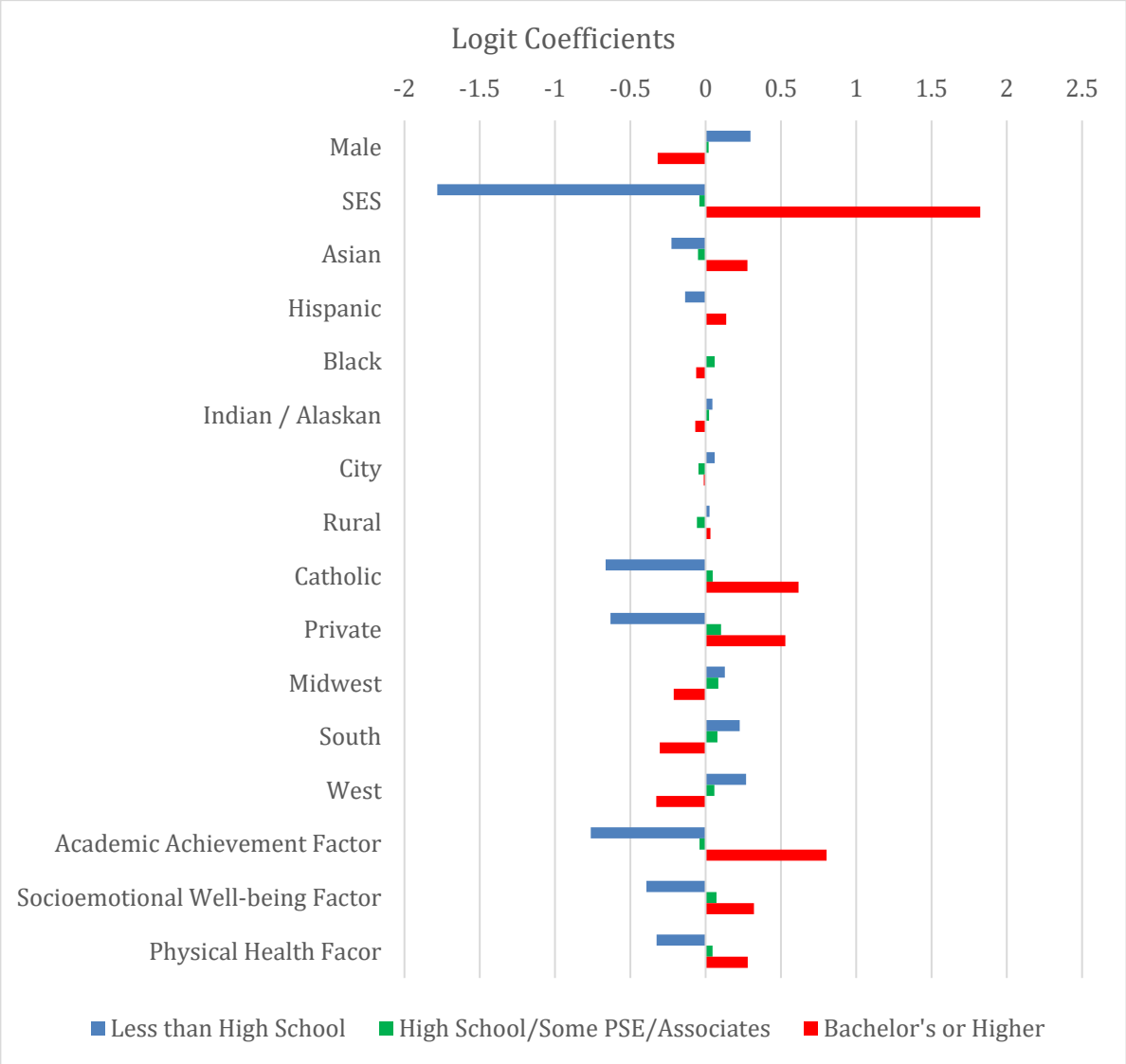

Figure 2. Multinomial logistic regression analysis results: logit coefficients for the predictors of educational attainment (three combined categories)

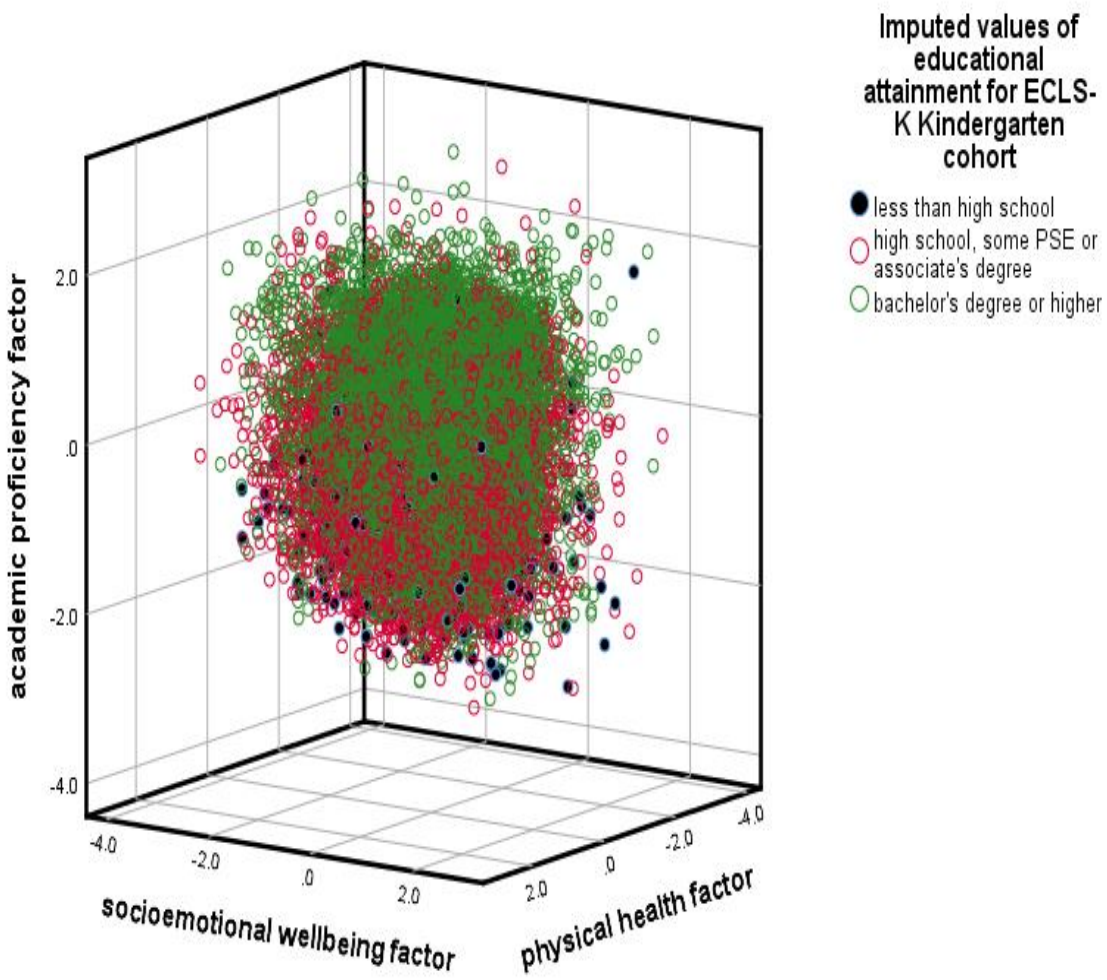

Figure 3. 3-D plot of the relationship between early childhood development indicators (academic, socioemotional and physical factors) and future educational attainment (imputed variable)

Discussion

We summarize the key findings and discuss their implications here. First, the assumption of missing at random (MAR) was tenable based on the pattern of missingness in ECLS-K and NELS datasets; all common variables for which there are differences between the two preexisting datasets were included in the missing data analysis and all biases related to them were removed. Then, the choice of imputation methods influenced the accuracy of imputation for missing educational attainment variable in ECLS-K data. In our cross-validation results, it turned out that FCS multiple imputation method performed relatively better than both EMB and KNN methods.

However, the accuracy of missing data imputation was highly uneven for different levels of educational attainment, particularly lower accuracy among the categories of high school graduates without any postsecondary education and high school graduates with 2-year college education (i.e., associate's degree). This misclassification error issue is likely attributable to the fact that these two groups are not highly distinguishable from each other by academic and other qualifications, since most 2-year colleges are open and non-selective to high school graduates. The broader classification scheme of educational attainment (e.g., 3 combined categories including less than high school, high school diploma and 2-year college degree, 4-year college degree and higher) would help reduce misclassification errors substantially. Further, this new classification scheme would be consistent with the framework of afore-mentioned "early college" education policy initiatives, that is, integrating high school and 2-year community college into a 6-year program that ensures students' attainment of high school diploma and associate's degree together.

Second, modeling imputed educational attainment variable based on the ECLS-K early childhood development indicators, via machine learning analytics, worked well with good model fit results. The choice of regularization methods (i.e., Ridge, Elastic Net, Lasso) hardly made a difference in the goodness of model fit with test data. This finding is likely attributable to the use of factor analytic method for data reduction in advance of regression analysis as well as non-sparse nature of our data with relatively larger sample size compared to the number of model parameters. Penalized regression methods (i.e., constraining a model) were expected to alleviate model overfit with training data and improve model fit with test data, thus reducing the generalization error (i.e., bias-variance tradeoff).

By monitoring and evaluating how well the model will perform on instances it has not seen before, we can build more confidence about its validity and usability. For instance, it remains to be seen how well the system of early college readiness indicators will perform with small or noisy data, particularly when the data size is relatively small (e.g., small school district or school setting where its Kindergarten cohort has sparse data for model fit).

Third, it is worth noting that Kindergarten students' academic, socioemotional and physical readiness measures in combination are equally or more powerful predictors of future (imputed) educational attainment in comparison with the effects of family SES and race/ethnicity background variables. While those demographic and socioeconomic markers are not policy-manipulatable variables in nature, using them as the predictors of educational attainment may have the risk of potential biases which may result in misclassification and discrimination such as negative stereotyping of certain disadvantaged and underrepresented groups of students (e.g., low-income, racial or ethnic minority, immigrant and refugee students); they might seem to have relatively lower initial performance and thus poorer chances of school success in spite of untapped potential for future growth.

Therefore, it is crucial for states and local school districts to develop a holistic and equitable assessment system of whole-child development, aligned with college/career readiness performance standards and benchmarks, which can help facilitate more accurate and timely diagnosis of developmental gaps/needs and personalize support/interventions for off-track students. School districts use several types of assessments to determine the progress of their students. The data generated from state accountability assessments as well as local district assessments are intended to give educators a picture of the student's achievement level that indicates if students are on track to be college and career ready by high school graduation (Anderson & Fulton, 2015; Dougherty & Mellor, 2010). However, the current gap in many school districts' data system is a systematic assessment mechanism to determine if students are meeting targets indicating college and career readiness (Jiao & Lissitz, 2016). The school districts administer several measures but educators find it difficult to determine if diverse students are on track to meet college readiness standards/benchmarks, and determine if their performance in lower grades map to the higher standards in upper grades. It is crucial to train and support educators with the use of various assessment tools and technologies.

Conclusion

In the midst of the U.S. education policy movement towards universal college education (ensuring access to at least two years of community college for all eligible students), what are often missing in this policy debate are supporting data and evidence on how well American education system gets all students ready for college and career through their K-16 education pathways. Although this policy has great potential to help significantly upgrade educational attainment and close the achievement gaps among racial and socioeconomic groups (Lee, 2016), its success depends partly on the quality of student assessment and decision support system. Thus, it is idealistic to build P-16 or K-16 education database tracking individual students' educational pathways, all the way from preschool/Kindergarten through postsecondary education and career. However, the NCES as well as state education agencies have not yet created such a seamless longitudinal education database due to financial and logistical difficulties.

In light of these concerns, our study sheds new light on the feasibility of creating synthetic K-16 education longitudinal database by linking two separate and preexisting longitudinal education datasets available from the NCES (i.e., ECLS-K and NELS). The application of novel research design and data analysis methods would help harmonize different datasets and link them together; the current data/research silos between elementary vs. secondary or postsecondary education levels remain as barriers to evidence-based policy and practice towards the goal of improving college readiness and success for all students. The K-16 education longitudinal database, if appropriately designed and used, will help inform educational policy for accountability and equity to improve all students' college and career readiness and close the achievement gaps among diverse student groups.

Recommendations

Our pilot study results provide implications and caveats about the feasibility of creating national synthetic longitudinal data on K-16 education pathways, including college and career readiness early indicators. The focus of this study was for early childhood academic, socioemotional, and physical development indicators to predict and inform later educational attainment pathways. Conversely, we anticipate that the reverse direction could work as well. Eventual college and career success measures could be used to identify early childhood precursors that may show promise for early prevention and intervention. We recommend further research and development efforts in order to link and correlate the variables between elementary and secondary/postsecondary education datasets in both directions (i.e., predictive/prospective validity on one hand and postdictive/retrospective validity on the other hand) for cross-validation. While the current policy debates on college readiness indicators and college admissions system tend to focus on the validity and utility of standardized test scores (e.g., SAT) and high school course grades (Amo & Lee, 2013; Glancy et al., 2014), this discussion needs to expand to multiple measures including non-cognitive development markers (e.g., educational and career aspirations, socioemotional skills, mental and physical health) and more diverse and long-term indicators of college and career success.

Limitations

Our study has limitations in that it relies on the validity of methodological assumptions as well as the validity of data/information that are available in preexisting NCES datasets. In this study, the relationships between ECLS-K and NELS variables are assumed to be conditionally independent (i.e., conditional on X, common demographics and background variables). The question is how to address uncertainty due to potential violation of the conditional independence assumption. In order to avoid the conditional independence assumption, the statistical matching should incorporate some auxiliary information concerning the relationship between Y and Z (D'Orazio et al., 2006). Further study needs to draw upon information from prior research that extends from kindergarten through high school or college and beyond; they might include the longitudinal studies of early intervention programs such as Perry Preschool Program (Schweinhart & Weikart, 1998) and Tennessee Project STAR (Finn et al., 2001).

The results of this study suggest that the currently available national education longitudinal datasets have some inherent limitations in terms of data missingness and incompatibility issues. Although there exist some commonly available demographic and socioeconomic background factors across the different datasets, they are only proxy (and potentially biased) indicators of students' readiness for college and career. Further, the datasets used in this study are outdated and only applicable to the U.S. student samples of specific time periods. Subsequent study needs to update the results with more recent data and address similar issues and challenges in different populations and settings.

Authorship Contribution Statement

Lee: Concept and design, data analysis / interpretation, drafting manuscript, critical revision of manuscript. Jaeger: Statistical analysis, drafting manuscript, critical revision of manuscript.

References

ACT. (2010). *Mind the gaps: How college readiness narrows achievement gaps for college success*. https://bit.ly/3mJm8QJ

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on-track and graduating in Chicago public high schools: A close look at course grades, failures, and attendance in the freshman year.* Consortium on Chicago School Research. https://eric.ed.gov/?id=ED498350

Amo, L., & Lee, J. (2013). Review of "SAT wars: The case for test-optional college admissions". *The Review of Higher Education, 36*(3), 405–406. https://doi.org/10.1353/rhe.2013.0031

Anderson, L., & Fulton, M. (2015). *Multiple measures for college readiness.* Education Commission of the States. https://www.ecs.org/clearinghouse/01/17/37/11737.pdf.

Berger, A., Turk-Bicakci, L., Garet, M., Knudson, J., & Hoshen, G. (2013). *Early college, early success: early college high school initiative impact study*. American Institutes for Research. https://eric.ed.gov/?id=ED577243

Bhopal, K. (2017). Addressing racial inequalities in higher education: equity, inclusion and social justice. *Ethnic and Racial Studies*, *40*(13), 2293–2299. https://doi.org/10.1080/01419870.2017.1344267

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Wiley. https://doi.org/10.1002/9781119942283

Conley, D. T. (2005). *College knowledge: What it really takes for students to succeed and what we can do to get them ready.* Jossey-Bass.

Data Qualtiy Campaign. (2014). *Data for action 2014: Paving the path to success.* https://bit.ly/3mHTqQd

DiPrete, T. A., & Buchmann, C. (2013). *The rise of women: the growing gender gap in education and what it means for American schools.* Russell Sage Foundation.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice.* John Wiley & Sons. https://doi.org/10.1002/0470023554

Dougherty, C., & Mellor, L. (2010). Preparing students for advanced placement: It's a P-12 issue. In P. Sadler, R. Tai, K. Klopfenstein & G. Sonnert (Eds.), *Promise and impact of the advanced placement program*. Harvard Education Press.

Eccles, J. S., Lord, S., & Midgley, C. (1991). What are we doing to early adolescents? The impact of educational contexts on early adolescents. *American Journal of Education*, *99*(4), 521-542. https://doi.org/10.1086/443996

Ellwood, D. T., & Kane, T. J. (2000). Who is getting a college education? Family background and the growing gaps in enrollment. In S. Danziger & J. Waldfogel (Eds.). *Securing the future* (pp. 283-324). Russell Sage Foundation.

Feldman, A. F., & Matjasko, J. L. (2005). The role of school-based extracurricular activities in adolescent development: A comprehensive review and future directions. *Review of Educational Research 75(*2), 159–210. https://doi.org/10.3102/00346543075002159

Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (2001). The enduring effects of small classes. *Teachers College Record, 103*(2), 145-183. https://doi.org/10.1111/0161-4681.00112

Finn, J. D., Gerber, S. B., & Wang, M. C. (2002). Course offerings, course requirements, and course taking in mathematics. *Journal of Curriculum and Supervision*, *14*(4), 336-366. https://eric.ed.gov/?id=EJ648747

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22. https://doi.org/10.18637/jss.v033.i01

Froiland, J. M., & Davison, M. L. (2016). The longitudinal influences of peers, parents, motivation, and mathematics course-taking on high school math achievement. *Learning and Individual Differences*, *50*, 252–259. https://doi.org/10.1016/j.lindif.2016.07.012

Geron, A. (2017). *Hands-on machine learning with Scikit-learn & tensor flow*. O'Relly.

Glancy, E., Fulton, M., Anderson, L., Zinth, J., Millard, M., & Delander, B. (2014). *Blueprint for college readiness*. Education Commission of the States. http://www.ecs.org/docs/BlueprintforCollegeReadiness.pdf.

Gutman, L. M., Sameroff, A. J., & Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: effects of multiple social risk factors and preschool child factors. *Developmental Psychology*, *39*(4), 777–790. https://doi.org/10.1037/0012-1649.39.4.777

Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., & Calkins, J. (2006). Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. *Early Childhood Research Quarterly*, *21(4)*, 431–454. https://doi.org/10.1016/j.ecresq.2006.09.005

Hauser, R., & Koenig, J. A. (2011). *High school dropout, graduation, and completion rates: Better data, better measures, better decisions.* National Academies Press.

Heckman, J., & Lochner, L. (2000). Rethinking education and training policy: Understanding the sources of skill formation in a modern economy. In S. Danziger & J. Waldfogel (Eds.), *Securing the future* (pp. 47-83). Russell Sage Foundation.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41–45. https://doi.org/10.1126/science.7604277

Henry, D. A., Betancur Cortés, L., & Votruba-Drzal, E. (2020). Black–white achievement gaps differ by family socioeconomic status from early childhood through early adolescence. *Journal of Educational Psychology*, *112*(8), 1471–1489. https://doi.org/10.1037/edu0000439

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. https://doi.org/10.18637/jss.v045.i07

Jack, A. A. (2014). Culture shock revisited: The social and cultural contingencies to class marginality. *Sociological Forum*, *29*(2), 453–475. https://doi.org/10.1111/socf.12092

Jiao, H., & Lissitz, R. W. (2016) (Eds.) *The next generation of testing: common core standards, smarter-balanced, PARCC, and the nationwide testing movement.* Information Age Publishing.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, *95*(1), 49-69. https://doi.org/10.1017/S0003055401000235

Kirst, M. W., & Venezia, A. (2004). (Eds.) *From high school to college: Improving opportunities for success in postsecondary education.* Jossey-Bass. https://doi.org/10.1037/e565212006-013

Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management, 31*(2), 203–227. https://doi.org/10.1002/pam.21615

Lee, J. (2012). College for all: gaps between desirable and actual P-12 math achievement trajectories for college readiness. *Educational Researcher, 41*(2), 43-55. https://doi.org/10.3102/0013189X11432746

Lee, J. (2016). *The anatomy of achievement gaps: Why and how American education is losing (but can still win) the war on underachievement.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217648.001.0001

Lee, J. (2020). What's missing from the nation's report card. *Phi Delta Kappan, 102*(4), 46-51. https://doi.org/10.1177/0031721720978067

Lee, J., Kim, N., Cobanoglu, A., & O'Connor, M. (2019). *Moving to educational accountability system 2.0: Socioemotional learning standards and protective environment for whole child development.* The Rockefeller Institute of the Government. https://eric.ed.gov/?id=ED605689

Lee, J., & Lee, M. (2020). Is 'whole child' education obsolete? Public school principals' educational goal priorities in the era of accountability. *Educational Administration Quarterly, 56*(5), 856-884. https://doi.org/10.1177/0013161X20909871

Lee, V. E., & Burkam, D. T. (2003). Dropping out of high School: The role of school organization and structure. *American Educational Research Journal, 40*(2), 353–393. https://doi.org/10.3102/00028312040002353

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119013563

MacIver, D. J., & Epstein, J. L. (1991). Responsive practices in the middle grades: Teacher teams, advisory groups, remedial instruction, and school transition programs. *American Journal of Education, 99*(4), 587-622. https://doi.org/10.1086/443999

Martin, C., Sargrad, S., & Batel, S. (2016). *Making the grade: A 50-state analysis of school accountability systems*. Center for American Progress. https://eric.ed.gov/?id=ED567858

National Governors Association. (2007). *Principles of federal preschool-college (P-16) alignment*. Stark Education Partnership.

National Research Council. (2012). *Education for life and work.* The National Academies Press.

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational Leadership, 65*(2), 28-33.

O'Connell, M. E., Boat, T., & Warner, K. E. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth, and Young Adults: Research Advances and Promising Interventions. The National Academies Press.

Owens, A. (2010). Neighborhoods and schools as competing and reinforcing contexts for educational attainment. *Sociology of Education*, *83*(4), 287–311. https://doi.org/10.1177/0038040710383519

Polidano, C., Hanel, B., & Buddelmeyer, H. (2013). Explaining the socio-economic status school completion gap. *Education Economics*, *21*(3), 230–247. https://doi.org/10.1080/09645292.2013.789482

Rau, W., & Durand, A. (2000). The academic ethic and college grades: Does hard work help students to 'make the grade'? *Sociology of Education, 73*, 19-38. https://doi.org/10.2307/2673197

Rosen, R., Byndloss, D. C., Parise, L., Alterman, E., & Dixon, M. (2020). *Bridging the school-to-work divide: Interim implementation and impact findings from New York City's P-TECH 9-14 schools*. MDRC. https://eric.ed.gov/?id=ED605308

Rubin, D. B. (1987) *Multiple imputation for nonresponse in surveys*. John Wiley & Sons Inc. https://doi.org/10.1002/9780470316696

Sander, W. (2006). Educational attainment and residential location. *Education and Urban Society, 38*(3), 307–326. https://doi.org/10.1177/0013124506286944

Schweinhart, L. J., & Weikart, D. P. (1998). High/ scope perry preschool program effects at age twenty-seven. In J. Crane (Ed), *Social programs that work* (pp. 148-162). Russell Sage Foundation.

Takahashi, M. (2017). Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: assessing the effects of between-imputation iterations. *Data Science Journal*, *16,* 1-17. http://doi.org/10.5334/dsj-2017-037

Templ, M., Alfons, A., Kowarik, A., & Prantner B. (2016). *VIM: Visualization and imputation of missing values. R package version 4.6.0*. https://CRAN.R-project.org/package=VIM.

van Buuren, S., Brand, J. P. L., Groothius-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049–1064. https://doi.org/10.1080/10629360600810434

Young, A., Johnson, G., Hawthrone, M., & Pugh, J. (2011). Cultural predictors of academic motivation and achievement: A self-deterministic approach. *College Student Journal*, *45*(1), 151–163.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

Description of ECLS-K Kindergarten Data and Variables for College Readiness Early Indicators

Academic Readiness Indicators

ECLS-K indicators of academic readiness among kindergarten students include overall measures of academic performance (Reading T-Score; Math T-Score) and sub-domain measures of knowledge and skills in reading and math (Print Familiarity; Count, Number, Shape; Letter Recognition; Beginning Sounds; Relative Size). Reading and math achievement measures were based on Item Response Theory (IRT) scale scores based on students' answers to multiple choice questions in each subject area. The reliability of scores for the reading assessments ranged from .91 to .95. The reliability of scores for the mathematics assessments ranged from .92 to .94. The first factor (Academic Readiness) has an eigenvalue of 6.07 and explains 36 percent of the combined variance.

Socioemotional Readiness Indicators

ECLS-K indicators of socioemotional readiness among kindergarten students include measures of socioemotional skills (Approaches to Learning: Interpersonal Skills; Self-Control) and problem behaviors (Externalizing Problem Behaviors; Internalizing Problem Behaviors). These measures were produced by teachers' ratings of their students in classrooms. The approach to learning scale measures behaviors that affect the ease with which children can benefit from the learning environment. The self-control scale is indicative of the child's ability to control behavior by respecting the property rights of others, controlling temper, accepting peer ideas for group activities, and responding appropriately to pressure from peers. Ranges of reliability coefficients (Cronbach's alpha) are as follows: Approach to Learning (.91), Self-control (.79-.82), and Interpersonal skills (.85-.87). The second factor (Socioemotional Readiness) has an eigenvalue of 2.48 and explains 15 percent of the combined variance.

Physical Readiness Indicators

ECLS-K indicators of physical readiness among kindergarten students include measures of motor skills (Fine Motor Skills; Gross Motor Skills), BMI in the range of not being underweight or overweight (Healthy Weight); parent's rating of child' overall health status (Child's Overall Health); parent's report of child's not having daily functioning difficulties (Child without Disability). Fine motor skills consisted of seven tasks: build a gate, draw a person, and copy five simple figures. Gross motor skills consisted of balancing, hopping, skipping and walking backward. Alpha coefficients (reliabilities) were 0.57 for fine motor skills and 0.51 for gross motor skills. The third factor (Physical Readiness) has an eigenvalue of 1.13 and explains 7 percent of the combined variance.