Secondary analysis of large-scale datasets in early childhood educational research

Introduction

Educational data has become increasingly complex, available, and extensive in response to the pressures of educational reform and the growing ease of data collection and analysis due to advancing technology. Large-scale data pertaining to U.S. student achievement has been collected on a national scale for over half a century: the National Assessment of Educational Progress (NAEP) was first administered to a nationally representative sample of 9-, 13-, and 17-year-olds in 1969, and continues to assess 4th, 8th, and 12th grade students every two years (Vinovskis, 1998). In the past 30 years however, the advent of standards-based accountability, the desire for measures of international comparison, and advances in computing that have continued to make collection, storage, and analysis of large amounts of data more attainable has led to the development of a great number of ongoing studies that produce large-scale datasets. Some of these studies, like the NAEP, the Trends in International Mathematics and Science Study (TIMSS), and the Program for International Student Assessment (PISA) repeatedly collect data on students at specific grade levels in order to assess educational trends. Other studies, like the High School Longitudinal Study (HSLS), the Early Childhood Longitudinal Studies (ECLS), and the National Education Longitudinal Study (NELS) follow a representative sample of students for several years to understand more about students’ educational pathways and experiences through time.
Educational researchers, who now possess new technologies that make large datasets more manageable, have taken full advantage of datasets such as TIMSS and NAEP and answered important pedagogical and policy questions. However, similar research efforts in early childhood education (ECE) have been less prolific even with the availability of large-scale data sets such as the Head Start and Early Childhood Longitudinal Studies (ECLS). The reasons appear to be manifold. First, some have suggested that many early childhood (EC) educators, and consequently EC educational researchers, may be uncomfortable understanding and using data from large repositories (Lake & Kelly, 2014; Lee & Ginsburg, 2009; Balfanz, 1999). Second, although researchers in developmental psychology and economics have made good use of these datasets to ask questions relating to their respective disciplines, the majority of educational researchers studying EC may be more drawn to qualitative methodologies that are more able to embrace the “messy complexity of classroom contexts” (Parks & Wager, 2021, p.76).

While it is true that research using large-scale educational datasets will never be able to approach the level of nuance and specificity that qualitative data can produce, these datasets do have the potential to convey a great deal of contextual information and yield fascinating insights into the ecosystem of ECE. In the following sections I hope to make the case for educational researchers to increase their use of secondary analysis of large-scale datasets to explore ECE. I will first discuss the affordances and constraints present in this type of research, what types of analyses these datasets make possible, and what datasets are readily available. I also provide a more in-depth look at the methodological considerations that underpin these types of studies, using the ECLS-K:2011 as an example, and discuss some ways this dataset has been used by educational researchers. Finally, I discuss the necessity of this type of research as access to public pre-kindergarten education expands nationwide.
Affordances of large-scale educational datasets

There are many reasons for educational researchers to take an interest in large-scale educational datasets. Some of these advantages are logistical, while others are methodological or theoretical. One straightforward advantage of using this type of data is the convenience and economy of performing secondary analysis. Researchers working with this type of data are able to avoid the cost and time associated with data collection and may move directly into the analytical phase of their research. Research consisting of secondary analysis of large-scale datasets usually qualifies for exemption from the stringent requirements that researchers collecting their own data must conform to in order to receive clearance from an Institutional Review Board (IRB). Additionally, while researchers collecting their own data may have to test multiple versions of their survey and assessment instruments, the instruments used to collect data in large-scale educational studies already include information on reliability and validity of the measures used.

The size of these datasets is one of their greatest advantages; researchers using large-scale educational datasets provide accurate effect size estimates, allowing researchers to perform more complex analyses and make more confident claims. The samples of these datasets are usually designed to be nationally representative, which allows researchers to assert the generalizability of their findings. Analysis of these datasets can illuminate trends and patterns that might otherwise have gone unnoticed, which can be invaluable in understanding how social, cultural, and political forces drive change in the EC educational context. In a society marked by systemic inequity, having this type of “birds-eye view” of educational systems is powerful and invaluable. Price (2019) writes of the clout that analyzing large-scale datasets lends to researchers’ claims about inequity:
“The abundance of data points used in large-scale quantitative analyses can provide an avenue for researchers to shelve their preconceived notions of how things appear to operate and instead focus squarely on the patterns in the historical, structural, institutional, and organizational data. With these aspects, the interpretation can be less ridiculed for being prone to interpretation in the eyes of the beholder and instead can be revered as providing the 20/20 lens to clearly see the patterns that undertow our social systems” (p. 204).

Mosqueda and Maldonado (2020), writing about the value of large-scale educational data analysis in improving educational outcomes for minoritized students, agree with Sobe’s message: “We believe that studies of large-scale data that are attentive to the methodological considerations of complex sampling, data clustering, and causal inference will contribute nuanced perspectives of promising policy and practice directions” (p. 38).

Through the types of sensitive analysis advocated by researchers like Price, Mosqueda and Maldonado, secondary analysis of large-scale educational datasets can give researchers the opportunity to answer unique questions with rich analyses that incorporate a wealth of contextual information. In turn, these analyses are invaluable tools in providing educational stakeholders with the evidence needed to promote more equitable policies and practices.

**Constraints of large-scale educational datasets**

Not all critical scholars agree that analysis of large-scale educational datasets is useful or beneficial. Many researchers have criticized work with large-scale data as being divorced from the daily realities of teaching and learning, arguing that too many analyses are means-focused and therefore describe a mythical “average student” who does not exist in the real world (e.g. Parks & Wager, 2021; Walker, 2016). Researchers working with large datasets may be tempted to take the data as a given and neglect to critically consider theoretical or broader contextual factors that could be playing a role in the data’s presentation. Sobe (2018), in a reflection on how comparative research is evolving in response to ever-increasing amounts of data, writes that
“there is an inherent conservatism to Big Data in that by using numbers as they are given, we are stuck with what is rather than what should be or what could be” (p. 561). Parks and Wager, in a 2021 call for more humanizing research in ECE, assert that the majority of studies based on large-scale survey and assessment data erase much of the context and complexity of the EC classroom, resulting in research that is lacking in actionable information that would lead to real improvements in ECE. Reviewing the three most frequently cited articles in the journal Early Childhood Research Quarterly (ECRQ), all based on large-scale educational datasets, they assert that “using this as our research foundation makes invisible the individual nature of the child” (p.79). Of course, Parks and Wagers’ argument hinges on a belief that rich, qualitative analysis is inherently superior to the types of analyses performed using large-scale datasets. While the latter may not be the ideal instrument to measure discursive interactions within a classroom, for example, the former is likewise an inadequate instrument for exploring larger social and political forces as they act on the educational system on the macro scale. Further, as discussed by Mosqueda and Maldonado, there are ample opportunities for nuanced contextual analysis for researchers who take full advantage of the wealth of data points present in large-scale datasets.

Although the wide-angle view given by large-scale educational datasets provides useful information that counterbalances the fine-grained analysis of small-scale, qualitative studies, there are other constraints present in this type of research. While secondary research does have the aforementioned logistical and economic benefits over a self-designed study, researchers using these datasets are bound to the information collected by the original researchers and may find a mismatch between their own research goals and preferred methodologies and those of the researchers who originally designed the studies. Questionnaires, for example, are blunt instruments, and researchers may find the reliance on pre-collected data limiting. Many
researchers who want to answer “why” rather than “what” questions may find it difficult to accomplish their research objectives with secondary analysis alone. Although the measures in these datasets include information on reliability and validity, some EC researchers argue that use of assessment data as evidence of proficiency is inappropriate when investigating young children, who have a wide range of developmental skills and levels of experience with assessments. Harrison and Wang, in their overview of quantitative analysis in the International Handbook of Early Childhood Education (2018), write that researchers planning to undertake this type of research must consider “the appropriateness of quantitative methods for collecting data with young children; the cultural appropriateness of some assessment tools; the reliability of some measures, such as self-report questionnaires, which can be affected by social desirability; measurement error; and the need for multiple sources of data to avoid bias in testing measured outcomes” (p. 299).

Finally, one unavoidable constraint of secondary analysis is the fact that the data begins to age as soon as it is released. This problem is especially relevant for ECE research, as the ECE system is undergoing a period of unprecedented rapid growth and change (McElrath, 2021). Due to the time, effort, and cost of administering these large studies and making the data available for analysis, these large-scale studies cannot respond as quickly as self-designed studies to new conditions of schooling, such as the massive changes wrought by disruptions due to the COVID-19 pandemic. Fortunately, the ECLS program will continue to administer new waves of data collection that will keep data more current, with ECLS-K:2024 currently in the recruitment phase.

Scholars choosing to perform secondary analysis of large educational datasets in their research must be proactive in examining their positionality, that of the researchers who originally
designed the studies, and maintaining awareness of the limitations of the instruments used to obtain the data. However, by maintaining an open mind, the ability to take multiple perspectives, and a strong background in theory and research methodology, EC educational researchers should be able to take advantage of the many opportunities for meaningful analysis afforded by secondary analysis of large-scale educational datasets.

**Methods of Analysis**

The following table provides an overview of some of the most popular methods of analysis used in conjunction with large-scale educational datasets. It is by no means a comprehensive list of possible methodologies, but instead provides a review of some affordances and critiques of these approaches to analysis as well as examples of studies that employed these analyses to answer ECE research questions. It is important to note that most studies will employ multiple forms of statistical analysis in order to obtain the fullest understanding of the data and most complete answers to research questions.
<table>
<thead>
<tr>
<th>Method of Analysis</th>
<th>Affordances</th>
<th>Constraints / Critiques</th>
<th>Example from ECE Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA, MANOVA</td>
<td>Comparing three or more groups Allows for analysis of relationships between categorical predictor variable(s) and continuous outcome variable(s)</td>
<td>While most of these critiques do not apply to very large sample sizes, researchers should keep in mind that violations of normality assumptions are common and can result in Type I and II errors. Given how common unequal group sizes are in education studies, researchers must be aware of the properties of their data and choose robust methods of analysis.</td>
<td>Djonko-Moore, 2022: <em>Diversity education and early childhood teachers’ motivation to remain in teaching: An exploration.</em> Teacher survey responses related to the intention to remain teaching and preservice teacher education coursework related to working with culturally and linguistically diverse students were used to compare teachers with high, medium, and low motivation to continue working in ECE. The author found that teachers with high motivation to remain teaching had taken more courses related to working with diverse students, and discussed implications for EC teacher preparation.</td>
</tr>
<tr>
<td>HLM / MLM</td>
<td>Allows for subjects to be nested within contexts (e.g. students within classrooms within schools within communities, etc.), or in longitudinal datasets to have observations nested within individuals and individuals nested within groups. This affordance is especially useful given the nature of educational systems.</td>
<td>Functionally similar results may be obtained from simpler forms of analysis, and there is not yet agreement on how many levels of analysis are necessary, or how much variation is appropriate at each level for the HLM to be useful (Niehaus, 2014).</td>
<td>Blöchliger &amp; Bauer, 2018: <em>Correlates of burnout symptoms among child care teachers. A multilevel modeling approach.</em> This Swedish study used two levels of analysis - the individual and organizational level. The researchers explored how burnout symptoms were clustered within childcare centers, and identified characteristics that were significantly related to teacher burnout at both levels. The individual level included age, control (over classroom activities), and reward (salary), while the organizational level included workload, community engagement, perceived fairness, and the childcare centers’ alignment with its stated values. The results of HLM showed that at the individual level, lower age, higher reward, and more control were all negatively related to burnout. At the organizational level, analysis showed only workload to be significantly related to teachers’ burnout symptoms.</td>
</tr>
<tr>
<td>SEM</td>
<td>Variables can act as both predictor and predicted variables (mediator variables) within a hypothesized model and models can have more than one outcome variable. SEMs can include regression relationships that include latent (unobserved characteristics/constructs) and observed variables; and can model correlated prediction/measurement errors (Bowen &amp; Guo, 2012).</td>
<td>Some critics believe that it is never appropriate to specify a causal model a priori. Researchers must have strong theoretical grounding for their hypothesized model. Without a sound rationale for why different variables are placed as they are, a directional hypothesized model may be more likely to prove itself as a statistical exercise, rather than revealing meaningful information (Freedman, 1987).</td>
<td>Loomis, Freed, &amp; Coffey, 2022: <em>Inhibitory Control, Student–Teacher Relationships, and Expulsion Risk in Preschools: An Indirect Effects Path Analysis</em> Using data from three instruments - the Child Behavior Questionnaire, the closeness subsection of the Student-Teacher Relationship Scale, and the Preschool Expulsion Risk Measure (PERM)- this study explored how Head Start students’ inhibitory control and closeness with their teachers related to their expulsion risk. Results showed direct, negative relationships between inhibitory control and expulsion risk and between student-teacher closeness and expulsion risk, and also found that student-teacher closeness acted as a moderator of the effect of inhibitory control on expulsion risk.</td>
</tr>
</tbody>
</table>
### Table 1, cont’d: Affordances, critiques, and examples of popular methods of statistical analysis for ECE research

<table>
<thead>
<tr>
<th>Method of Analysis</th>
<th>Affordances</th>
<th>Constraints / Critiques</th>
<th>Example from ECE Research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RD Regression Discontinuity</strong></td>
<td>In appropriate contexts, allows researchers to achieve a quasi-experimental research design. A cut-off value is used to divide participants into multiple groups, and researchers can observe the difference in trajectory between groups in response to a treatment/intervention. It is a useful design in education, as allocation of services is often decided by a specific cut point that works well with RD.</td>
<td>There is a danger of misspecified functional form leading to misinterpretation of results. For example, if a relationship is quadratic, but is assumed to be linear, there may not actually be a discontinuity in the relationship (Cook, 2008). Researchers have also pointed out that many RD studies have not sufficiently adjusted for pretreatment variables (Imbens &amp; Lemieux, 2008).</td>
<td>Umansky, 2016: <em>To Be or Not to Be EL: An Examination of the Impact of Classifying Students as English Learners.</em> Using longitudinal data from students in a large California school district, Umansky compared students who were close to the cut point between being classified as EL or not, a determination that is made based on their score on the California English Language Development Test. She found that EL classification had a fairly strong negative effect on students’ scores on state mathematics assessments through 7th grade, and on state English language arts assessments through 10th grade. When comparing this effect across classroom models, she found that it was concentrated in English immersion classrooms.</td>
</tr>
<tr>
<td><strong>VAM Value-added Modeling</strong></td>
<td>VAMs in education generally use student achievement data as the “value” or output, and seek to determine how different aspects of a students’ educational experience add to or detract from their performance on an achievement test. Many researchers argue that these models are fairly robust, show only small levels of bias, and are indispensable in ensuring that children are receiving the high quality education to which they are entitled (see Koedel et al., 2015, for a review).</td>
<td>VAMs in educational research have been used most famously in an attempt to determine individual teachers’ effects on student achievement as part of the No Child Left Behind teacher quality-based reforms (Koedel et al., 2015). This method of determining teacher effects, and therefore teacher quality, has been extremely controversial. There are many educational researchers who question the reliability and validity of VAMs as measures of teacher quality and caution that these analysis should not be used as the primary rationale for high-stakes employment decisions (e.g., Haertel, 2013; Newton et al., 2010; Baker et al., 2010)</td>
<td>Gottfried, Garcia, &amp; Kim, 2019: <em>Peer tutoring instructional practice and kindergartners’ achievement and socioemotional development.</em> Employing VAM with school fixed-effects, this study used ECLS-K:2011 data to examine the effect of frequency of peer tutoring on kindergarteners’ academic and socioemotional outcomes. The researchers found that higher frequency of peer tutoring positively predicted students’ assessed social skills, but was not predictive of their achievement on English language arts or mathematics assessments.</td>
</tr>
</tbody>
</table>
Missing Data

Missing data is unavoidable in large-scale studies, especially those with a longitudinal design. The longer a study lasts the more researchers must deal with participant attrition. Subjects may not fill out a questionnaire, or may neglect to answer specific questions, children may be absent on assessment day, or there may be multiple versions of a questionnaire over time resulting in some questionnaires missing certain items. Researchers must first determine if missing data is missing at random (MAR), missing completely at random (MCAR), or if it is not missing at random (NMAR). If data is found to be not missing at random (NMAR), and instead the probability of data missing from a variable is found to be related to the value of that variable, researchers must model the missing data mechanism as part of their analysis (Cheema, 2014). If data is MAR or MCAR, researchers can deal with missing data either by deletion, imputation, or through using a full information maximum likelihood procedure.

Listwise deletion involves deleting any participant who is missing data from a sample, while pairwise deletion involves analyzing all nonmissing data without deleting participants who are missing some data (Warner, 2007). Deletion methods can be problematic. Reducing sample size leads to wider confidence intervals and less powerful hypothesis tests, and if data is NMAR deletion can result in a sample no longer being generalizable to the population under study (Cheema, 2014).

Imputation, the process of assigning replacement values for missing data, allows researchers to retain a sample’s size and statistical power. There are a variety of methods of imputation researchers can choose to employ based on their specific research context. Some forms of imputation are simple, such as median imputation in which a variable’s median is substituted for the missing value, but these may yield misleading results. Multiple imputation,
which can be used when data is MAR or MCAR, “simulates the natural variation in missing data by imputing such missing data several times” based on relationships between values found in the observed data, and then averaging these estimated values into a single set of estimates that can replace the missing data (Cheema, 2014, p. 494).

A final method of dealing with missing values when data is MAR or MCAR is through using a full information maximum likelihood (FIML) estimation procedure. FIML is extremely useful as it allows researchers to use all available data from all participants. SEM programs, such as lavaan in R, can be used to apply FIML to a range of problem types.

**Selected available datasets**

The following section contains basic information on relatively recent datasets pertaining to ECE that are currently available for analysis. This is not an exhaustive list, but is rather intended to provide EC researchers with a starting point for exploring large-scale data. The following datasets were all collected from subjects in the United States, but researchers interested in international participants will find that there are a number of datasets available to them from around the world.

Many large-scale datasets, and all of those listed below, have both public-use and restricted-use files. Public-use files are freely available for researchers to download, but have had several variables recoded or suppressed to maintain participant confidentiality. Restricted use files contain all data in its original format, but require an often lengthy application process. Researchers must prove to the organizations overseeing the data that they have extensive security measures in place to prevent a breach of participant confidentiality. Some datasets, like the North
Carolina Longitudinal Data System (NCLDS), require a substantial fee for access. The datasets discussed below have their public-use files freely available for download.

*Early Childhood Longitudinal Studies (ECLS) Program*

The National Center for Education Statistics (NCES) is the United States’ primary organization responsible for collecting and analyzing educational data, both domestic and international. NCES is housed within the U.S. Department of Education and Institute for Education Sciences (IES). NCES has completed a series of studies under the Early Childhood Longitudinal Study (ECLS) program, which includes assessment, demographic, and other survey data from children, their teachers, school administrators, parents, care providers, and other contextual information. The original ECLS-K:1998 as well as the ECLS-K:2011 collected data twice a year on a nationally representative cohort of children between kindergarten and fifth grade, while the ECLS-B, initiated in 2001, deals with children from birth to kindergarten entry and collected data every other year beginning at 9 months. The ECLS-K: 1998 surveyed approximately 22,000 children, the ECLS-K:2011 surveyed approximately 20,250 children, and the ECLS-B surveyed about 10,700 children. NCES is currently recruiting participants for the upcoming ECLS-K: 2024 (NCES, 2022).

*Head Start and Early Head Start Studies*

Head Start, established in 1965, and Early Head Start, established in 1994, provide childcare to approximately one million children from birth to age five. These programs, serving families with low income, are overseen by the U.S. Department of Health and Human Services’
(DHHS) Administration for Children and Families (ACF) and have an extensive body of research spanning the past 50 years of program implementation. Two notable datasets include the Head Start Impact Study (HSIS) and the Family and Child Experiences Study (FACES). The HSIS is a longitudinal study conducted between 2002 and 2008, followed approximately 5,000 three- and four-year-old children who were randomly assigned to either a treatment group, in which they received Head Start services, or a comparison group, in which they did not receive Head Start services but may have received services from other community organizations. Data was collected in the form of classroom observations, in-person interviews with parents, teacher ratings of children, and assessment data relating to the children’s cognitive and developmental abilities, and was collected in the fall and spring until the children reached first grade, with a follow up in the spring of their third grade year. The FACES study collected descriptive data pertaining to children beginning Head Start, their families, and their classrooms and teachers. FACES is not a longitudinal study, but has collected fall and spring data from seven different cohorts ranging in size from 2,000 to 22,000 children between 1997 and 2020. Data was collected in the form of classroom observations and data, teacher and parent interviews and surveys, and child interviews, assessment scores, and other data. The program is set to continue into 2026 (DHHS, ACF, 2022).

The State Preschool Yearbook

Beginning in 2001, the State Preschool Yearbook is an annual survey study undertaken by the National Institute for Early Education Research (NIEER) that tracks access, policies, funding, and characteristics pertaining to state-funded prekindergarten programs in the United States. The data is primarily drawn from surveys of state preschool administrators from all fifty states, as
well as some territories. NIEER, housed in Rutgers University’s Graduate School of Education, oversees several research projects that may be of interest to researchers studying EC educational issues, especially those pertaining to policy.

*The National Early Intervention Longitudinal Study (NEILS)*

Sponsored by NCES, the NEILS focuses on children with diagnosed disabilities who participated in an early intervention program between 1998 and 2004. The researchers sought to better understand early intervention, the accessibility of services, and the outcomes for participating children. It is a nationally representative, longitudinal study of 3,400 young children from entry into early intervention until entry to kindergarten. Data was collected in the form of questionnaires and surveys of children’s teachers and intervention service providers, as well as parent interviews (Hebbeler et al., 2007).

*Pre-Elementary Education Longitudinal Study (PEELS)*

The PEELS is another dataset that explores the educational experiences of young children with disabilities. Also sponsored by NCES, the PEELS is a longitudinal, nationally representative study of approximately 3,000 children between the ages of 3 and 5. Data was collected between 2003 and 2009, and consists of assessment data, parent interviews, and teacher, principal, and educational administrator questionnaires.

*Child and Family Data Archive*

An excellent resource for researchers to be aware of is the Child and Family Data Archive. This archive is a collaboration between the Office of Planning, Research, and Evaluation (OPRE),
who oversees research on all ACF programs, and the Inter-University Consortium for Political and Social Research (ICPSR). ICPSR hosts an extensive data repository on a wide range of topics, and has a variety of resources for those interested in learning more about data analytical techniques. The Child and Family Data Archive, launched in 2019, allows researchers to search through a vast collection of OPRE data, reports, publications, and presentations relating to young children. Researchers can search specific datasets by name, or they can search topics or variables that suit their research interests.

**Looking more deeply into a large dataset - the ECLS-K:2011**

To better understand the planning and methodological considerations that go into creating a large-scale dataset, we will now look more closely at the ECLS-K:2011. The ECLS programs have provided some of the most widely used U.S. data in the study of ECE, and this section is intended to give the reader more insight into how a nationally representative, longitudinal study is carried out, as well as how educational researchers have made use of the data.

**Sampling**

The ECLS-K:2011 used a multistage sampling design to obtain its nationally representative sample of 20,250 children attending 1,320 schools. For the first stage of sampling, researchers began by using a list of all 3,141 U.S. counties to create a second list of primary sampling units (PSUs) to select the sample from (Mulligan, 2017). The ten PSUs with the largest numbers of 5-year-olds were included in the sample, while the second group of PSUs were sampled using a stratified sampling method based on 40 different strata determined by: “MSA [metropolitan statistical area] status, census geographic region, size class (defined using the
measure of size), per capita income, and the race/ethnicity of 5-year-old children residing in the PSU” (Mulligan, 2017, p. 3). From each of these 40 strata two PSUs were selected with probability proportional to their size (Mulligan, 2017). By clustering via PSUs, NCES was able to reduce field costs as sampled students were more likely to attend the same schools due to living in the same geographic area (NCES, n.d.). A simple random sample, in contrast, would have pulled children from a larger geographical area and yielded a sample with higher variation.

In the second stage of sampling, researchers created a list of all public and private schools that served 5-year-olds in each sampled PSU, and sampled from these lists with schools selected with probability proportional to the size of the schools’ kindergarten enrolment. In the final stage of sampling, researchers selected 23 kindergarteners from each selected school. NCES oversampled students who identified as Asian-American / Pacific Islander (API), with these children being sampled at 2.5 times the rate of other children (NCES, n.d.). This was in order to have enough students in each from the API group to have sufficient statistical power when analyzing and making predictions connected to this group (Mulligan, 2017).

In addition to oversampling, subsampling, the practice of taking a sample of a sample, was also used in collecting data for the ECLS-K:2011 in order to reduce field cost (NCES, n.d.). A 30% subsample of the base sample (the original sample of entering kindergarteners and their schools) was used for data collection for fall first grade and fall second grade, and a 50% subsample was used for students who moved schools after the base year (NCES, n.d.).

Data collection

Data for the ECLS-K:2011 was collected from the students, their families, their classroom teachers, their before/after school care providers, and their schools and school administrators. Researchers collected data through self-report questionnaires, as well as through
in-person or telephone interviews and one-on-one assessments. Data was collected every fall (September-December) and every spring (March-June) between September 2010 and June 2016. Questionnaires were distributed to parents, teachers, before and after school care providers, and to school principals. Teachers filled out questionnaires on their own characteristics, beliefs, and teaching practices, as well as a separate questionnaire for each participating student in their class. Students were assessed on mathematics, English language arts, and science, as well as on their health (height, weight, hearing) and executive function ability. The research team conducted field tests of all instruments in the fall of 2009 to ensure reliability and validity of instruments and to estimate the parameters of all items (Mulligan, 2017).

**Weighting**

Researchers use weighting, the practice of statistically compensating for situations like missing data, nonresponse, or disproportionate sampling, in order to reduce bias and make samples that better represent the target population, thereby improving accuracy of predictions. The ECLS-K:2011 researchers developed sample weights in multiple stages. In the first stage, researchers assigned weights to each PSU equal to the inverse of the PSU’s probability of being selected (Mulligan, 2017). In the second stage, researchers assigned weights to the schools within the sampled PSUs by multiplying the PSU weight by the school’s inverse probability of being selected from its PSU. The resulting weights are the schools’ base weights, which were adjusted for nonresponse (Mulligan, 2017). The third stage of weighting involved computing base weights for each student selected for the sample, a process that also used probability of selection based on total number of students versus number of students sampled. This calculation
was separate for API and non-API students, due to the previously mentioned oversampling of API students (Mulligan, 2017). These student-level weights were also adjusted for non-response.

While this process seems extensive as is, NCES is clear that more weights based on nonresponse “could be produced from every component of the study […] and for every combination of components” but that “creating all possible weights for a study with as many components as the ECLS-K:2011 has would be impractical” (Mulligan, 2017, p. 8). As a result, the NCES researchers had to decide which weights to provide based on what might be most useful for researchers, and researchers performing secondary analysis must decide which weight to use given their specific analytical needs. NCES directs researchers to consider three components of their study when determining which weight to use: the level of analysis; the round(s) of data to be analyzed; and the specific components that make up the data for the analysis (NCES, n.d., p. 8). Researchers should choose weights that are adjusted for nonresponse for the appropriate data collection rounds as well as for the greatest number of components in the analysis (NCES, n.d.). NCES cautions that researchers should use the same sample weight for all analyses within a single research study so that all estimates are being produced from the same underlying sample (NCES, n.d., p. 14).

Missing Data and Imputation

The ECLS-K:2011 contains various codes for missing data that allow researchers to know why the data is missing (see figure 1). Before analyzing the data it is important that researchers recode missing data to avoid having the missing data codes be treated as legitimate values by their statistical analysis program. How researchers recode missing data values should be based on the specifics and objectives of their studies.
In the ECLS-K:2011, there was a relatively high rate of nonresponse for several variables in the parent questionnaire, leading NCES researchers to impute the values for the ECLS:K-2011 dataset. For example, in the base year 15.3% of parents did not respond to the item asking about household income (Mulligan, 2017). For the parent questionnaire, missing data was imputed for each missing component using the hot deck method. Hot deck is a type of imputation in which missing values for a non-responding participant are replaced with observed values from another participant who is similar to the non-respondent on one or several characteristics (Andridge & Little, 2010). For each missing component similar respondents and non-respondents were grouped into “imputation cells” that were defined by a number of characteristics that were found to be the best predictors of that component (e.g. school type, census region, household type, race/ethnicity, age) (Mulligan, 2017, p. 8). Within imputation cells respondents’ values were randomly donated to serve as non-respondents’ values, with no respondent serving as a donor more than once (Mulligan, 2017).

ECLS:K-2011 researchers used a different method to impute missing values from the school administrator survey. Some school administrators did not respond to the questionnaire, while others neglected to answer the three questions required to calculate the composite score for percent of students approved for free or reduced price lunch. To impute these values NCES researchers took the data from the Common Core of Data (CCD) for most public schools, and used hot-deck imputation for public schools that did not have CCD data available (Mulligan, 2017).
<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>Not applicable, including legitimate skips</td>
</tr>
<tr>
<td>-2</td>
<td>Data suppressed (public-use data file only)</td>
</tr>
<tr>
<td>-4</td>
<td>Data suppressed due to administration error</td>
</tr>
<tr>
<td>-5</td>
<td>Item not asked in School Administrator Questionnaire form B</td>
</tr>
<tr>
<td>-7</td>
<td>Refused (a type of item nonresponse)</td>
</tr>
<tr>
<td>-8</td>
<td>Don’t know (a type of item nonresponse)</td>
</tr>
<tr>
<td>-9</td>
<td>Not ascertained (a type of item nonresponse)</td>
</tr>
<tr>
<td>(blank)</td>
<td>System missing (unit nonresponse)</td>
</tr>
</tbody>
</table>

Figure 1. Missing data codes in the ECLS-K:2011 (NCES, n.d., p. 19)

Applications of ECLS in educational research

In this section I will present some notable examples of studies in which educational researchers used ECLS data to answer interesting and important questions about teaching and learning in EC contexts, and how their findings may be evidence for needed policy changes. It is not difficult to find examples of ECLS data being used by developmental psychologists- there is a wide body of research dealing with psychological concepts like executive function (e.g. Grissmer et al., 2010; Little, 2021; Eilloughby et al., 2019), locus of control (e.g. Han, 2010; Hooper et al., 2010; Morgan et al., 2008), and motivation (e.g. Marshik et al., 2017; Froiland & Oros, 2012; Coleman & McNeese, 2009). While developmental psychology obviously plays an integral part of the ECE ecosystem, researchers coming from an education disciplinary background may tend towards research questions that have a systems-level (e.g. classroom-, school-, district- level analysis) approach to educational issues, rather than tending towards a focus on the individual traditional to psychology.
The wealth of variables that researchers performing secondary analysis of large-scale datasets like the ECLS have at their disposal allows for a virtually endless number of questions to be explored, and makes the data conducive to educational research. As mentioned previously, analyses that take into account the plethora of contextual variables present in ECLS and similar datasets can reveal how various aspects of the educational ecosystem impact students or teachers differentially based on various sociodemographic characteristics. Evidence from such equity studies are especially valuable, as these samples are designed to be nationally representative, making the findings generalizable and (hopefully) trustworthy and compelling to policymakers.

One interesting area of study is how students are perceived by their teachers, and what role bias might play in their schooling experience. For example, Umansky and Dumont (2021) used a coarsened exact matching method with data from the ECLS-K:2011 to explore whether students’ English Learner (EL) classification status in kindergarten had an effect on teachers’ perceptions of their academic ability. Through comparing students who were statistically similar on a variety of measures, including English literacy assessments, but differed in their official EL classification status, Umansky and Dumont found that teachers had lower expectations of students classified as EL unless they taught in bilingual classroom environments. They call for researchers to explore what features of the bilingual classroom context moderate the relationship between student classification and teacher perceptions, and urge educational leaders to both expand bilingual education opportunities and implement policies that ensure EL students have full access to all educational opportunities available to their monolingual peers to mitigate the negative impacts of classification. Ready and Wright (2011) explored a similar topic using a different methodology. To determine how childrens’ backgrounds relate to teacher bias the authors used hierarchical linear modeling (HLM) to compare teachers’ perceptions of students’
literacy skill levels to students’ achievement on a literacy assessment, revealing clear negative bias along expected sociodemographic lines.

Studies comparing similar students can be invaluable in understanding the validity and reliability of benchmark assessments for students with different characteristics. Robinson (2010) compared similar students using ECLS-K data to explore the effect of test translation on Spanish-speaking students’ mathematical performance. Robinson used a regression discontinuity (RD) design, comparing students just below or just above the English proficiency cutoff point to be administered the mathematics assessment in Spanish versus English. Although research on older students has shown that language of instruction outweighs home language when considering testing language, the kindergarteners performed substantially better when administered the test in Spanish. The large negative effect sizes for students taking the test in English for both spring of Kindergarten (Cohen's $d = -0.85, p = .001$) and spring of first grade (Cohen’s $d = -.122, p = .012$) is especially striking given that Kiefer et al.’s 2009 metaanalysis of test accommodations for ML showing that most accommodations had extremely small(Cohen’s $d = .05$) or nonsignificant effect sizes. Concluding that home language is more important than language of instruction when choosing a testing language that will most accurately reflect young students’ ability level, Robinson urged district leaders to determine whether the intent of assessments of young children is to predict outcomes on state tests (in English) or to determine their actual level of mathematical proficiency.

Of course, while testing and classification both have important implications, the majority of EC researchers may be more interested in exploring how specific teaching practices or teacher characteristics relate to student learning and success. Researchers have used self-report survey data from teachers to determine what types of instructional practices are most prevalent in EC
classrooms, what patterns emerge between instructional balance and short- and long-term student achievement, and how typical instructional practices vary depending on factors like classroom composition, grade level, school type, and policy context. For example, Engel et. al. (2016) wondered whether teachers’ instructional practices in kindergarten mathematics had changed between 1998 and 2011 with the advent of standards-based instruction and increased accountability measures. The researchers divided mathematics instructional content into four categories and used OLS regression to compare the ECLS-K and ECLS-K:2011 base year data. Their analysis revealed that reported instructional practices had not shifted substantially between 1998 and 2011. Despite the advent of significant policy changes pushing for increased focus on advanced, concept-focused instruction in early grades, teachers in both cohorts reported spending the majority of mathematics instructional time on “basic counting and shapes,” a practice that has been shown to be negatively associated with student achievement, and spending less time on “advanced topics,” which has been shown to be positively associated with student achievement (Engel et al., 2013, 2016). The authors suggest that preservice and inservice teacher education should focus on aiding kindergarten teachers in shifting to mathematics instructional practices that place a greater emphasis on advanced topics.

As we can see from these examples, through sensitive secondary analysis of ECLS data educational researchers have been able to make important connections between students, teachers, and contexts of schooling to promote equity-focused practices and policies. Going forward, such analyses will be crucial towards improving early learning systems as they continue to expand.

**Conclusion and Implications**
The need for EC educational researchers to increase their use of large-scale educational datasets has never been greater. The United States is experiencing an era of widespread growth in access to publicly-funded pre-kindergarten education (McElrath, 2021), at both the state and federal levels. California, New Jersey, and Virginia are among several states currently undergoing massive expansions in their early education systems (D’souza, 2021; Garver et. al, 2022; Delaney, 2021), while the Biden administration’s American Families Plan intends to allocate substantial funding to help states augment existing prekindergarten systems to accommodate all three- and four-year-olds (White House, 2021).

In light of these expansion efforts, there is a pressing need for more research performed by scholars who have a deep understanding of EC educational contexts. There is substantial evidence that attending a center-based prekindergarten has short- and long-term benefits for students (e.g. McCoy et al., 2017; Nores & Barnett, 2010; Camilli et al., 2010), especially for students from historically underserved communities (e.g. Johnson et al., 2022; Cannon et al., 2012; Vandell, 2004). This strong research base has been invaluable in winning support for initiatives that expand publicly-funded prekindergarten, but is still insufficient to inform the design and implementation of such programs. Despite the positive picture painted by the majority of longitudinal or meta-analytical studies of prekindergarten impacts, some large-scale studies of prekindergarten implementation have returned less optimistic outcomes. Some studies have concluded that the academic benefits of prekindergarten are relatively short-lived, or only benefit specific student demographics (e.g. Bartik & Hershbein, 2018; Barnett, 2022). In rare cases like Tennessee or Quebec’s public preschool expansions, evidence suggests that some programs may actually be detrimental to students (Barnett, 2021).
To inform the implementation of both current and future early education initiatives, evidence that attends to the instructional practices, environmental characteristics, and organizational features that result in effective, high-quality programs that have lasting impacts on students and their communities will be invaluable. This research must draw attention both to the impacts of systemic racism and to program features that are able to mitigate these effects. EC educational researchers are uniquely positioned to perform such research. With their deep understanding of the ECE ecosystem, EC educational researchers are able to move beyond more narrowly defined views of prekindergarten education typically seen in psychology or economics to a broader conception of teaching and learning in the EC context. Wideen, Mayer-Smith and Moon, in their seminal 1998 review of research on teacher education, called for researchers to take an ecological approach to research on learning to teach as opposed to continuing to produce more fragmented, disparate studies. Their call seems to be pertinent in this case as well: research that will truly improve outcomes for young students must show “a full appreciation of the inseparable web of relationships” (p. 170) that constitute the ECE context, including not only students but their communities, teachers, schools, and ever-evolving socio-historical and political conditions. Through more EC educational researchers taking up secondary analysis of large-scale educational datasets, prekindergarten expansion efforts will have the best chance at reaching their full potential for students, teachers, and communities.
References


