Creating an Integrated Longitudinal Person and Household Dataset by Linking State-based Administrative Data

Deven Carlson
Robert Haveman
Barbara Wolfe
Kyong Wan Kang
Hannah Miller

Institute for Research on Poverty
University of Wisconsin-Madison

July 2015

This report draws on the work by staff and affiliates of the Institute for Research on Poverty (IRP) in the creation of the Multi-Sample Person File (MSPF) data system. Key funding for the development of the MSPF was provided by the U.S. Department of Health and Human Services (DHSS) in collaboration with the Wisconsin Department of Children and Families. Patricia Brown, Maria Cancian, and Jennifer Noyes were Principle Investigators on the DHHS contract, and the primary contributors to MSPF. A technical report on the MSPF's creation [Brown et al (2014)] is available upon request from IRP. Initial funding for accessing and matching the educational data used in our data set was provided by the U.S. Department of Health and Human Services in collaboration with the Wisconsin Departments of Children of Families and Public Instruction; Lawrence Berger, Maria Cancian and Jennifer Noyes were Principal Investigators on this contract. We also acknowledge the contributions of Dan Ross of IRP for his contribution to the cleaning of the data and the geocoding of the observations. Financial support for this project was provided by the Smith Richardson Foundation, with additional financial assistance from the Annie E. Casey Foundation. We acknowledge the help and support provided by the Mark Steinmeyer, Senior Program Officer at the Smith Richardson Foundation and Cindy Guy, Program Officer at the Annie E. Casey Foundation. Any opinions are those of the authors alone.

Table of Contents

I. Introduction

Increasingly, social scientists are making use of administrative data to address important research questions. Typically, these administrative data have been collected by individual state or federal agencies to record and document their interactions with clients (e.g., payments to or collections from individuals and households). In education, administrative records for K-12 school systems record a wide range of student and school characteristics, enabling program administrators to monitor the effectiveness of programs for which they are responsible. Although use of administrative data has some disadvantages,[1] it also has several advantages: large sample sizes; inherent longitudinal structure; and fewer problems with attrition, non-response, measurement error, and under-reporting.

Because individual administrative databases tend to be specific to particular agencies or programs, taken separately many of them are of modest use to researchers. For social scientists, the value of these databases increases dramatically when they are linked to each other. When databases are linked, researchers can associate individual or household characteristics with information regarding individual and household payments, collections, location, income sources, employment, education, and other indicators of performance. Databases formed by linking information from several administrative databases typically enable the relevant linked variables to be observed over time.[2]

In our research, we rely on records contained in databases maintained by State of Wisconsin agencies to analyze the correlates of family, school, and neighborhood characteristics with low-income youth educational achievement. We also address the question of the impact of family receipt of public means tested benefits on the educational performance of children living in these families. Our goal is to develop a more complete understanding of the determinants of success among students where success is broadly construed to include academic achievement in traditional subjects, regular attendance of school, good behavior (avoiding suspensions or expulsions) and regular promotions. In the future we hope to also focus on youth choices and success regarding post-secondary schooling.

The purpose of this paper is to document the long and difficult process that we have worked through in our formation of a reliable linked administrative data set, created from numerous databases available from several public agencies. Our hope is that our experience in confronting and solving the numerous issues in constructing a reliable single dataset from numerous and unrelated administrative databases will be of help to others considering this research strategy.

[1] For instance, complex structure, duplicate cases, limited information, etc. See Card, Chetty, Feldstein, and Saez.(2010) and Hotz, Goerge, Balzekas, and Margolin (1998) for further details.

[2] Hotz et al. (1998) is an early report on the potentials and problems of using linked administrative databases; this report describes efforts in several states to create linked administrative databases. A later but related report is Brady, Grand, Powell, and Schink (2001). Card et al. (2010) make the case that the US is losing ground in this area and provide recommendations to reverse this trend. Also, a recent paper by Einav and Levin (2014) discusses the current status of the use of linked administrative data in the social sciences. As these reports indicate, studies making use of such data range across social sciences, education, and medical sciences.

In the following two sections, we introduce the several databases that we use in our study and then present the details of the process by which the integrated data set was created. In the fourth section, we describe the challenges that we faced in the process of data construction, our decisions regarding them, and the considerations that guided these decisions. In the final section, we summarize the contents described in the previous sections and provide suggestions for researchers who are also planning on constructing a unified data set based on multiple linked administrative data sources.

II. Administrative Databases Employed

1. Multi-Sample Person File (MSPF)[3] Data System

The Multi-Sample Person File (MSPF) is a longitudinal administrative database created and maintained by the Institute for Research on Poverty (IRP) programming and research staffs at the University of Wisconsin-Madison. In 2008, the IRP programming and research staffs created the first version of the MSPF data system by "drawing on information extracts from the full universe of clients or participants in the State of Wisconsin's electronically available administrative data on public assistance, child support, child welfare, unemployment benefits, and incarceration, and merging them to create a single file of unique individuals" (Brown, 2014, p. 3). The MSPF data system refers to the "master MSPF one-record-per-individual file, along with linkable aggregation files (parent/child, and case level data) and participation files (monthly benefits, eligibility, payments/receipts, or spells)" (Brown, 2014, p. 4).

Before the creation of the MSPF data system, studies had been conducted within IRP using the State of Wisconsin administrative data in which researchers extracted a sample cases from one administrative data source and linked them to other administrative data source for that sample (Brown, 2014). However, since the creation of the MSPF, researchers are able to use the full universe of cases or individuals from one source of administrative data and add other administrative data source for their analysis (Brown, 2014). This development has broadened the scope of research questions that can be addressed by the State administrative data (Brown, 2014). The existence of the MSPF aided our study by providing a list of individuals/households in any of several databases and indicators of the databases in which that person/household is listed.

After the release of initial version of MSPF in 2008, subsequent versions were released in 2010, 2011, and 2012; our study uses MSPF 2011 (for education data) and MSPF 2012 (for housing data). MSPF lists all individuals who have been identified in seven Wisconsin administrative data sets from 1988-2012. That is, the MSPF file includes indicators for whether each individual was observed in each of the following administrative data sources, along with demographic, family composition, and location information on each individual:

- Client Assistance for Re-employment and Economic Support (CARES),
- Kids Information Data System (KIDS),

[3] This section draws heavily from Brown (2014).

- Statewide Automated Child Welfare Information System (SACWIS),
- Department of Corrections (DOC),
- Milwaukee Jail (MJ),
- Unemployment Insurance (UI), and
- Court Record Data (CRD).

Also, the MSPF provides an individual identifier—the IRPID—that can be used to link records across the databases stated above. The process of constructing the MSPF by IRP programming and research staffs is described in Section III.

2. Client Assistance for Re-employment and Economic Support (CARES)

The CARES database has been maintained by the State of Wisconsin since 1994 and contains a wide variety of detailed information—including household composition, demographics, address history, and public program participation—on all cases that apply for or receive any form of public assistance from the state. CARES data contain over 500,000 unique records annually and include participation information for the following programs:

- Wisconsin Works (W2), the Wisconsin version of TANF
- Food Stamps/Supplemental Nutrition Assistance Program (SNAP)
- Medical Assistance (Medicaid, BadgerCare, BadgerCare Plus)
- Child Care Subsidies (WI Shares)
- Supplemental Security Income (SSI), Social Security (SS), Social Security Disability Insurance (SSDI).

As we describe in greater detail below, from CARES we extract information for our dataset on individual demographic characteristics (e.g., race, marital status, and education level), household composition, benefit receipt status, and address history to create a household-based dataset. We also identify whether or not households received a Section 8 housing subsidy using information gathered when households applied for W2 or SNAP benefits.[4]

3. Unemployment Insurance (UI) Database

The UI database is maintained by the State of Wisconsin and consists of wage records, unemployment insurance benefit amounts, and benefit time periods (spells). Wage records contain quarterly total wages in dollars; benefit amounts report monthly cash benefits paid to unemployed workers who continue to search for employment; and benefit time periods (spells) report the covered time period of unemployment by indicating whether a person was unemployed in each week (Brown, 2014). The database contains over 2.5 million records annually for people who work in Wisconsin. From 2000-2012 wage records and 2007-12 plus the last quarter of 2006 benefit amounts and spells records, we constructed information on individual- and household-level quarterly earnings to be included in our integrated data set. We

[4] Subsidized housing data is gathered for applicants to W2 and Food Stamps in the CARES system; it is only gathered at time of application and perhaps at review times.

could get benefit amounts and spells records only from the fourth quarter of 2006 because IRP did not request those data prior to that time.

4. American Community Survey (ACS)

The ACS is an ongoing nationwide survey that samples a small percentage of the population every year to give communities the information they need to plan investments and services. The survey is conducted, and the resulting data are maintained, by the United States Bureau of the Census. The ACS asks about age, sex, race, family and relationships, income and benefits, health insurance, education, veteran status, disabilities, where people work and how they get there, where people live, and how much people pay for some essentials.

Specifically, we will draw upon ACS's block group characteristics, which are based on surveys conducted over a five year period and provide variety of contextual measures for these small geographic areas – block groups typically contain only 1,000 to 2,000 inhabitants. The sample for Wisconsin is approximately 4,500 unique interviews per year for residents in all of the State's Census block groups. From these data, we extracted information on Census block group characteristics of households such as the percentage of persons in poverty, the unemployment rate, median family income, the percentage of young adults with less than a high school education,, and median house value for years 2005-2012. These measures provide us with important contextual information about the neighborhoods in which students reside.

5. Wisconsin Information System for Education (WISE)

The WISE data system is maintained by the Wisconsin Department of Public Instruction (DPI) and contains information about students in all public school districts and schools in Wisconsin. WISE contains information on student-level test scores (Wisconsin Knowledge and Concepts Examination [WKCE][5] test scores) and student-level attendance rates. Also, WISEdash (WISE Data Dashboard) provides files where student-level information has been aggregated to the school level (e.g., racial composition, percent of students with disabilities, and average attendance rate). The data are available to us from the 2005-06 school year through 2011-12; we use data from both the WISE and WISEdash in constructing our integrated data set.

6. Wisconsin Information Network for Successful Schools (WINSS)

WINSS is the older existing DPI data portal and it is the authoritative source for topics not yet transitioned to WISEdash (WISEdash, n.d.).[6] It provides a wide variety of school-specific

[5] The WKCE is a statewide standardized exam given each school year to students enrolled in Wisconsin public schools to measure student achievement in core academic areas. Through 2013-14, students in grades 3, 5, 6, and 7 took tests in reading and mathematics and students in grades 4, 8, and 10 took tests in reading, mathematics, science, language arts, writing, and social studies. Beginning with the 2014-15 school year, reading, mathematics and language arts are no longer tested using the WKCE. The WKCE is only being administered in grades 4, 8, and 10 and only for science and social studies (The Wisconsin Department of Public Instruction, n.d.).

[6] Since 2000, WINSS has been an important data resource used by education stakeholders, but in fall 2013, the DPI implemented a new and updated public data analysis portal called WISEdash. So, DPI has

data about academic performance, attendance and behavior, staff and other school resources, and student demographics for years 1993-94 – 2011-12. Because WINSS provides information on high school graduation rates, we extracted this information for school years 2005-06 – 2011-12.

7. DPI Assessment Data

Because neither WISEdash nor WINSS contains an ideal school-level test score measure, we obtained scale score summary files from the DPI website to construct such a measure.[7] These files report the distribution of scale scores earned by students taking the WKCE subject area tests and contain mean scale scores, their standard deviations, and local percentiles.[8]

8. Common Core of Data (CCD)

CCD annually collects and reports descriptive information about students and staff, as well as fiscal data (including revenues and current expenditures), from all public school districts and state education agencies in the United State. This database is managed by the National Center for Education Statistics (NCES). Data are available for the 1986-87 school year and all subsequent years; we use data on school-level percent of students eligible for free/reduced price meal, county code, and county name for the 2005-06 – 2011-12 school years.

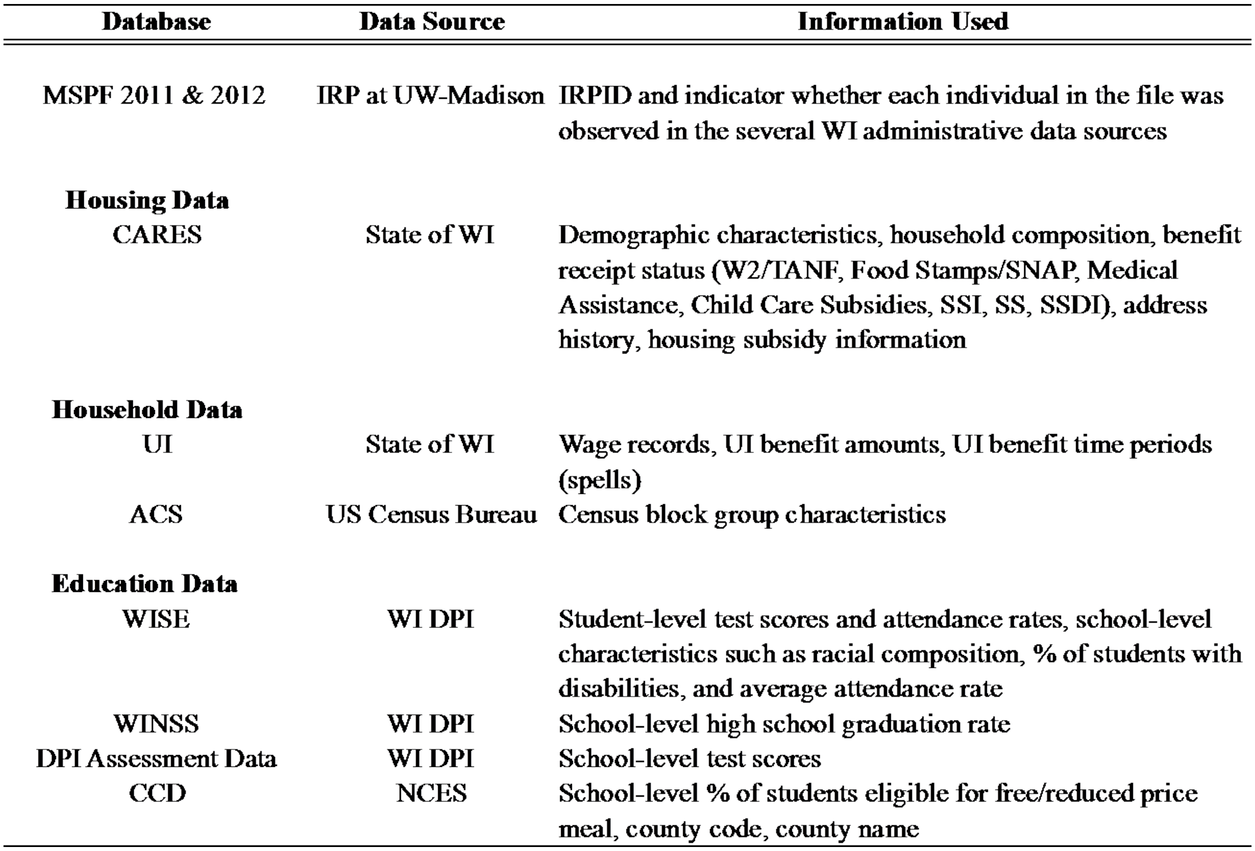
Table 1 summarizes the information that is contained in our core integrated data set. It indicates the database, data source, and the information that we extracted from each dataset and included in our core data set.

begun the process of migrating content from WINSS to the WISEdash and to the School District Performance Report. This migration will be completed in 2016 (Welcome to WINSS! n.d.).

[7] http://oea.dpi.wi.gov/assessment/data/WKCE/summaries. A scale score is a score on a numeric scale with intervals of equal size. The scale is applied to all students taking the WKCE in a particular subject at a particular grade level, making it possible to compare scores from different groups of students or individuals from year to year (The Wisconsin Department of Public Instruction [WI DPI], 2013). The scale was developed based on item response theory (IRT), which simultaneously considers test item characteristics (e.g., item difficulty) and students' performance on the items (The Wisconsin Department of Public Instruction, 2006).

[8] Local percentiles describe the location of scale scores of lower, middle, and higher performing students in each student group. For each group, 10%, 25%, 50%, 75%, and 90% of students tested scored at or below the scores reported for that group at the 10th, 25th, 50th, 75th, and 90th percentiles, respectively. For example, the 90th local percentile divides the highest 10% of the scores of students in a group from the lowest 90% of students in that group (WI DPI, 2013, p. 22).

Table 1. Description of the Dataset Used in This Study



Database	Data Source	Information Used
MSPF 2011 & 2012	IRP at UW-Madison	IRPID and indicator whether each individual in the file was observed in the several WI administrative data sources
Housing Data		
CARES	State of WI	Demographic characteristics, household composition, benefit receipt status (W2/TANF, Food Stamps/SNAP, Medical Assistance, Child Care Subsidies, SSI, SS, SSDI), address history, housing subsidy information
Household Data		
UI	State of WI	Wage records, UI benefit amounts, UI benefit time periods (spells)
ACS	US Census Bureau	Census block group characteristics
Education Data		
WISE	WI DPI	Student-level test scores and attendance rates, school-level characteristics such as racial composition, % of students with disabilities, and average attendance rate
WINSS	WI DPI	School-level high school graduation rate
DPI Assessment Data	WI DPI	School-level test scores
CCD	NCES	School-level % of students eligible for free/reduced price meal, county code, county name

III. Data Construction Process

The process that we followed in constructing our core integrated data set is complex. Some parts of the data construction work were done by IRP programmers and some parts were done by our research team members, which we reference in the text. Here, we move seriatim through the primary steps in this process.

Step 1. Development and Provision of MSPF by IRP Programmers[9]

As mentioned in the previous section, the first version of MSPF was created in 2008 by the programming and research staffs of IRP. The primary task of creating an MSPF file was to un-duplicate individuals within each of the administrative data sources and to match-merge individuals between data systems using individual characteristics, demographics, and the identity of the individual's parents or children, so that a final data file should contain only one

[9] This section draws heavily from Brown (2014).

observation per individual (Brown, 2014). This MSPF file provides an IRPID that can be used to link records across different administrative data sources contained within MSPF. The MSPF data system is accessible to researchers via secure servers running Linux and Windows, which are managed by the Social Science Computing Cooperative (SSCC) at the UW-Madison (Brown, 2014).

Step 2. Linking MSPF 2011 with DPI WISE Data by IRP Programmers

In October 2012, IRP programmers began working with DPI to merge DPI WISE student attendance and test score (WKCE) records with records from WiSACWIS[10] in MSPF 2011. This linking effort was initiated for the project of Berger, Cancian, Han, Noyes, and Rios-Salas (2014) and our research team was able to make use of the results of this data linking effort.

This data linkage was processed using four iterative steps. First, DPI provided IRP a file containing demographic information for every child in the DPI WISE database, along with their student ID (Longitudinal Data System [LDS] student key[11]). Second, IRP programmers matched those students in the DPI WISE data system to all children in the MSPF 2011, on the basis of their demographics (e.g., name, sex, date of birth, county of birth, and county of residence). Third, IRP returned the LDS student key to DPI, along with an MSPF identifier (= IRPID). This IRPID is pivotal in that it can be used to link a given student across all the relevant data sources. Lastly, DPI linked student test scores and attendance data to the IRPID and sent that data back to IRP without the LDS student key.

Step 3. Constructing Education Data and Codebook

a. Student-Level Data

In January 2014, we gained access to WISE attendance[12] and WKCE test score data for school years 2005-06 through 2011-12. Approximately one percent of cases had duplicate IDs so we cleaned the data files in order to produce the final data files with only one IRPID per individual student. The following describes our process of addressing duplicate cases in student-level attendance and test score data files:

1) Attendance Data

[10] Wisconsin Statewide Automated Child Welfare Information System. It is “a large-scale transaction based automated data system designed to permit simultaneous data entry from multiple sites and multiple workers (Berger et al., 2014, p. 40).” It includes modules for child abuse/neglect and foster care, and provides adoption analysis and reporting (Berger et al., 2014, p. 39).

[11] “An education longitudinal data system is a data system that collects and maintains detailed, high quality, student- and staff-level data that are linked across entities and over time, providing a complete academic and performance history for each student; and makes these data accessible through reporting and analysis tools (National Forum on Education Statistics [NFES], 2010, p. 7).” The assignment of a student key (= student identifier) is “a way to follow students as they move from grade to grade, and across campuses and/or districts within the state (NFES, 2010, p. 15).” The student key does not “permit a student to be individually identified by users of the system (NFES, 2010, p. 49).”

[12] Attendance data files also contained variables related to student discipline, students with disabilities, and English as a Second Language.

(Original N = 4,014,364, updated N = 3,988,488, about 0.6% of the original cases were dropped)

a) If there were duplicate cases where one record was missing and one was non-missing, we kept the non-missing one. (dropped n = 2,434)
b) If the merged the test score data indicated a duplicate case in which the test scores matched one record in the attendance data but not the other, we retained the record that was matched. (dropped n = 2,837)
c) We dropped all duplicate cases where the race differed from record to record (likely matching error). (dropped n = 5,082)
d) We dropped all duplicate cases where the sex differed from record to record (likely matching error). (dropped n = 766)
e) We kept observations where the attendance rate was not missing. (dropped n = 188)
f) Since there is no basis for dealing with the remaining the duplicates we decided to drop them all. (dropped n = 17,248)

2) Test Score Data
(Original N = 1,926,702, updated N = 1,912,941, about 0.7% of the original cases were dropped)

a) The DPI checks whether the records are from the same school and the same district. If a duplicated ID is provided by different schools or districts, it is likely that the same ID number was assigned to multiple students by accident and the DPI considers the data from these cases unusable. Therefore, if duplicate IDs were from different schools or districts, we deleted all those cases. (dropped n = 12,265)
b) If cases with duplicate IDs had a blank row of math and reading scores (likely a data entry error), we deleted the cases with that blank row. (dropped n = 330)
c) The DPI checks whether some of these cases might be duplicated because a student took both the WKCE and the WAA (the WAA-SwD[13] and the WAA-ELL[14]). In those cases, the decision rule at the DPI is to keep the WKCE results. Therefore, if duplicate IDs were generated from a student who took both the WKCE and the WAA, we deleted rows with the WAA and kept rows with the WKCE. (dropped n = 32)
d) The rest of the duplicate IDs have different WKCE or WAA scores in math or reading (or both) reported by the exact same school. Since neither the DPI nor we are able to explain why these duplicates exist, we deleted all those cases. (dropped n = 1,299)

Second, using the cleaned WKCE test score file, we created standardized reading and math score variables to enable the comparison across different grades and years.[15] To calculate

[13] Wisconsin Alternate Assessment for Students with Disabilities

[14] Wisconsin Alternate Assessment for English Language Learner

[15] Test score data files contained scores for reading, math, sciences, language arts, and social sciences, but we only use reading and math scores for our study.

standardized test scores, we obtained the statewide average test scores and statewide standard deviations of test scores by each tested grade from the DPI website.[16] For standardization, we subtracted the WI mean test score from the student's scale score and divided it by the WI standard deviation.[17]

Next, we created an education data codebook consisting of a variable list, variable definitions, variable explanations, frequency/summary tables, related notes, and glossary. This codebook was created based on 48 variables including the cleaned attendance and test score variables that we managed above. Other variables include economic status of students, whether students were retained, the number of days out-of-school due to disciplinary problems, whether students completed high school, grade level, demographics, etc. This codebook was shared with our group members as well as the separate research team of Berger, Cancian, Han, Noyes, and Rios-Salas (2014), to understand the education data to which we gained access.

b. School-Level Data

School-level test score data were obtained through scale score summary files from the DPI website (see footnote 16). The files provided the test scores by seven tested grades (e.g., grades 3, 4, 5, 6, 7, 8, and 10). Each file was a separate grade and within each grade file, there was a school-level measure of student achievement. We standardized the test scores[18] and combined those grade-level test scores into a single school score for each school year. Then, we calculated weighted reading and math scores[19] to consider the different sample sizes of each tested grade.

Next, we obtained files containing school-level attendance rate, racial composition, and percent of students with disabilities data from the WISE website.[20] We arranged these data files to have one observation for each of these variables per school per year. We then obtained a file containing a high school graduation rate variable from the WINSS data.[21] We also arranged this file so that there is one indicator per school per year. Finally, we obtained a school-level percent of students eligible for free/reduced price meals data from the CCD[22] and also organized this file so that there is one observation per school per year.

Step 4. Getting CARES and Employment Data via MSPF 2012

[16] http://oea.dpi.wi.gov/assessment/data/WKCE/summaries

[17] Formula for Standardization of Student-Level Test Scores = (Student's Scale Score – WI Mean Score) / WI Standard Deviation

[18] Formula for Standardization of School-Level Test Scores = (School's Mean Score – Mean Score of All Schools Contained in the Data) / Standard Deviation of All Schools Contained in the Data

[19] Formula for Weighting Test Scores = (N of Tested Students in a Given Grade / Total N of Tested Students in a School) × Standardized School-Level Test Score

[20] http://wise.dpi.wi.gov/wisedash_downloadfiles

[21] We could not use WISE data because high school graduation record at WISE was available only from 2009-10. And WINSS data was obtained from here: http://wise.dpi.wi.gov/wisedash_downloadfiles

[22] We decided to use CCD data because neither WISE nor WINSS had useable school level free/reduced price meal record.

While the DPI data extends to spring 2012 on a school year basis, the data in the MSPF 2011 exists only until the end of calendar year 2011. So, obtaining the data for spring 2012 from MSPF 2012 was required. However, some IRPIDs differ in the MSPF 2011 and MSPF 2012; because each new edition of the MSPF attempts to fix errors from the prior edition (e.g., one observation from the MSPF 2011 may be split into two observations in the MSPF 2012 if records were incorrectly combined in MSPF 2011), some individuals' IRPIDs are not be the same across editions.

Therefore, we used a crosswalk file created by IRP programmers to match individuals across MSPF editions. If individuals were missing in MSPF 2012, they will not appear in the MSPF data we are using and if people were missing in MSPF 2011, they will not appear in the DPI data. Hence, we dropped cases that were not in both MSPF 2011 and MSPF 2012; this eliminated 18.28% of the observations (n = 1,246,188). Then, we dropped cases if there were multiple observations of an IRPID in either MSPF 2012 or MSPF 2011; this eliminated 4.1% of the observations (n = 229,363).

Step 5. Linking Individuals with Households

We are interested in estimating the effects of housing subsidy receipt on students' educational outcomes. However, students do not directly receive a housing subsidy; rather, an adult or a household as a whole receives a subsidy. In addition, other means-tested benefits are distributed to households, rather than individuals. Therefore, we needed to determine with which household(s) each student was associated in each year. This was complicated by the fact that (1) household membership is fluid, and (2) different means-tested benefit programs group people differently based on the needs of that particular program. For example, for child support payments, nonresident parents may be associated with a given household. In contrast, for Food Stamps benefits, nonresident parents may not be associated with the same household. Also, an individual can be associated with more than one household for different means-tested programs and at different time points. Therefore, we needed to establish decision rules to link individuals with only one household at each point in time.

For each of these records, we generated year and month indicators, requiring the records to be organized into an annual format. (The record for an individual "begins" whenever information on that individual is added to the dataset; thus, the records begin on different dates and are for different lengths of time. That is, the records are not all one year or one month or two years long. The length of records varies and might be eight months, three years, or five years, etc.)

When an IRPID was associated with more than one IRPCASEID (which we think of as the "household") in the same year, we proceeded as follows:

a) We first prioritized IRPCASEIDs from the Food Stamps files if possible since most of the individuals in the housing data are identified when they sign up for Food Stamps.

b) If an IRPCASEID was not found in the Food Stamps data, we used IRPCASEIDs from CARES more broadly (see below for an explanation of the decision rules used with the CARES file).

c) In cases where the IRPCASEID was found in the Food Stamps file and an IRPID was associated with more than one IRPCASEID in the same year, we looked for the maximum value in order to determine if the value for SNAP was positive in each year. That is, we looked for the maximum Food Stamps value[23] for each IRPID-year combination in August of that year.

d) If the maximum was greater than zero, we decided to use IRPID-year-IRPCASEID links that received a non-zero Food Stamps amount. That is, we dropped any observations within the IRPID-year combination that had '0' in August of that year. We did this based on the belief that if an individual received Food Stamps through one household but not another, it is likely he/she is more strongly or immediately associated with the household in which he or she received Food Stamps. Here, we dropped 0.04% of the remaining observations (n = 3,485).

e) We dropped any remaining duplicates within IRPID-year combinations and used their IRPCASEID from the CARES data instead. We did this because the Food Stamps data files provide no additional information to choose among the remaining IRPID-year combinations. In contrast, using the CARES data, we are able to follow the decision rules below in order to hopefully make a more informed decision about which IRPID-year combination to retain. Following this step means using the CARES data rather than the Food Stamps data for n = 397,930, or 4.7% of the sample at this point.

If an IRPID was associated with more than one IRPCASEID in a given year in the CARES file, we could draw on the following information in deciding to which household the individual is to be assigned: start date, end date, and role. Therefore, we matched IRPCASEIDs with IRPIDs in August of each year, using the following decision rules if an IRPID was associated with more than one IRPCASEID in a given year:

a) We retained the observation with the later start date. The start date is the date the individual began a particular role in relation to the case/household. Therefore, a later start date would seem to indicate a more recent record;

b) We retained the observation with the lower numbered role (where "roles" indicate a person's relationship to the case – for example, mother, father, child, husband, and lower roles indicate a more direct relationship to the nuclear family);

c) We retained the observation with the later end date. The rationale for retaining the observation with the later end date is the same as for the start date (i.e., a seemingly more recent record). Note that relatively few records have end dates, which is why we do not prioritize this decision rule above the role rule – it does not help eliminate many observations; and

[23] We found the maximum simply to see if any IRPID-year had food stamps value greater than zero (i.e., if the maximum was greater than zero). We could have used many different strategies to determine whether an IRPID-year had a value that was not zero. In this sense, using the maximum to obtain the desired information was arbitrary (many other functions would have worked).

d) If IRPIDs have identical roles, start dates, and end dates, we retained the observation associated with the smaller IRPCASEID. This decision rule is arbitrary – we have no other way of deciding to which household an individual should be assigned.

If an IRPCASEID from the housing dataset could not be matched with any IRPIDs in August of that year, we repeated the procedure above using IRPCASEID-IRPID links from subsequent months in following order: September, October, November, and December.

Step 6. Indentifying Primary Person of Household

Next, we sought to identify each household's "Primary Person (PP)" in order to include information about the racial/ethnic background, educational attainment, and marital status of an adult in the house. Since there were often multiple adults associated with the household, identifying the characteristics of the household's "head" or "primary person" enabled us to provide consistent background information about the household in which the student was living. The issue here was that different data sources defined the PP differently. To address this issue, we followed these decision rules:

a) If possible, we use PP information from the Food Stamps data. When using the Food Stamps data, we identify the PP in August of a given year in order to match up with the start of the school year; if no PP is identified in August, we use earlier (then later) months in order (e.g., July, June, May, April, March, February, January, September, October, November, and December). We did it this way because it seemed to make sense to use the PP prior to the start of the school year but as close to August as possible.
b) If more than one PP is identified in Food Stamps data in a given year, we use the PP from the Food Stamps data that matches the PP from the CARES data. In three cases, more than one PP was identified in the Food Stamps data but there was no PP identified in the CARES data; in these cases, we coded the lowest IRPID as the PP.
c) If the IRPCASEID was not found in the Food Stamps data in a given year, we used PP information from the CARES data.
d) With these procedures, we identified the PP in August; if August was missing, we used the PP from September, October, November, or December in that order. Ninety six percent of the CARES observations do not have end dates; thus, when we tried to use earlier months to identify or "find" PPs that were missing in August, we did not find any. This is likely because, since most records did not have any dates, any PPs in earlier months were still coded as PPs in later months (i.e., it is not likely that someone started as a PP on a case prior to August in a given year but had a valid end date on that case by August of that year). Thus, we look for PPs going forward from August because we could not find additional PPs that were missing in August but available prior to then.

Step 7. Creating Housing Data

a. Identifying Time Period of Receiving Housing Subsidy

We identified each time period that each IRPCASEID received a housing subsidy through the following steps:

a) For each case, we generated the spell over which the case received a subsidy (spell = continuously receiving or not receiving a housing subsidy);
b) We created indicators for the first, second, third, etc. time each IRPCASEID received a housing subsidy; and
c) We generated an indicator to mark the month of first subsidy receipt.

If an IRPCASEID received both public housing and a rental subsidy, we classified the case by the type of housing subsidy receipt that occurred first (otherwise, the case would be “double-counted” as both a public housing and rental subsidy recipient in our analyses). In the treatment group, only 0.5% of the sample (n = 378) received both types of housing subsidies. In the control group, only 0.7% of the sample (n = 1,917) received both types of housing subsidies.

b. Constructing The Treatment Group

We focused on the 2006-08 school years so that it would be possible for students to have at least one year of pretreatment education data. We aggregated across cohorts to increase the sample size (while maintaining a cohort indicator). We focused on summer recipients so that we could have the cleanest identification of pre-treatment vs. post-treatment school years. For example, if a student received a housing voucher in June 2006, it is more apparent that the final pre-treatment year is the 2005-06 school year and the first post-treatment school year is the 2006-07 school year. For a student whose family received a housing voucher halfway through the school year, it is less clear how to code academic years in relation to treatment receipt. We focused on families who were newly in public housing or who received a new housing voucher so that we would have both pre- and post-treatment measures. Given that our housing data only extend back to 2000, we cannot say that these families were *never* previously in public housing or receiving a housing voucher but we can say that they were not receiving a housing subsidy immediately prior to treatment receipt.

Specifically, for the treatment group, we then identified households that received (1) a *new* housing voucher or (2) were *newly* in public housing in late spring to early fall of 2006, 2007, or 2008. We define “late spring to early fall” as April, May, June, July, August, or September. We define “new” as having a discontinuity between a prior receipt (if any) and summer 2006. This means that, while families could have received a housing voucher or been in public housing prior to April 2006, they were not receiving this voucher or living in public housing immediately prior to April 2006. We have 8,423 students in our treatment group.

c. Constructing Control Groups

For the first or our primary control group, we identify students living in families who received a housing voucher or were in public housing from 2009 to 2011; who had not previously received a housing voucher/been in public housing between 2000 and 2009; and who could be matched with the DPI attendance data at any point between the 2005-06 and 2011-12 school years. We do not restrict this group to families who received a voucher or were in public housing in "late spring to early fall" but we do include an indicator for those who received the new voucher or new public housing in April, May, June, July, August, or September of 2009, 2010, or 2011. Because this control group ultimately receives housing vouchers or public housing, it is intended to be the closest comparison to the treatment group. These are students from families who presumably are somewhat similar economically to the treatment group but who received housing subsidies at a later point in time. In fact, many of these families may have been on housing waiting lists at the time the treatment group received their subsidies. Using this approach we have 8,838 in this control group.

For the second control group, we identify students in families who are in the MSPF at any point between 2000 and 2011; who could be matched with the DPI attendance data at any point between 2005 and 2011; whose families did not receive a housing voucher or public housing in 2006, 2007, or 2008; and who had a PP associated with the case in 2006, 2007, or 2008 (i.e., who were *not missing* a PP in the year of interest). From the group of students who met all of these criteria, we then randomly selected twice as many as the number of students in the treatment group in each year. This control group is intended to provide a broader comparison for the treatment group. Through this control group, we compare outcomes for students in the treatment group to those for a broader swath of the school-aged children in Wisconsin. Using this sampling process we have 10,124 students in control group 2. A student can be in both control groups.

Step 8. Geocoding by IRP Programmers[24]

Geocoding of address history will be done by IRP programmers while the linking of geocoding results with ACS data will be done by our research team members through the following steps:

a. Sample Determination

IRP programmers obtained all IRPIDs from the DPI WISE-MSPF 2011 match. Then, they limited the sample to those with a match to at least one CARES case in MSPF 2011 because only CARES has address histories. From CARES, programmers obtained all addresses in effect on September 1st of years 2005-2013. Then, they limited the addresses to those for cases in the sample.

b. Cleaning

Addresses in administrative systems are often prone to error, whether from data entry errors or mistakes in understanding the variety of types of possible address parts and formats. Residential (as opposed to postal) addresses have fewer updates and corrections, because they are not used for sending materials by U.S. Mail and therefore do not result in returned mail or address corrections by the government agency managing the system. Rural addresses have more varied formats, so, even though standard formats exist, those entering data may not be aware of them.

[24] We thank IRP programmer Dan Ross for his contribution to this geocoding section. He has performed this task for another project using MSPF and hence has comprehensive knowledge of the required steps.

The cleaning process has several steps:

a) Limit addresses to Wisconsin addresses;
b) Inspect data and write code to correct misspelled Wisconsin cities;
c) Inspect zip codes (and corrected it if it is clearly wrong);
d) Standardize addresses, and correct them where possible:
    - Edit house number to remove apartment numbers, letters, and half (1/2);
    - Edit grid-based house numbers (e.g., N123W12345) to be compatible with geocoder;
    - Correct common street name misspellings and common yet nonstandard abbreviations;
    - Edit street names that are numeric for correct ordinal suffix (e.g., 10TH);
    - Correct street name if the street is known to have been renamed.

c. Geocoding

The geocoding process will use SAS's PROC GEOCODE software. To allow conversion from street address to geographic position (latitude and longitude), address locator data will be downloaded from the U.S. Census Bureau, specifically the 2013 TIGER/Line Shapefiles[25] for all 72 Wisconsin counties. IRP programmers will use three types of TIGER/Line file: edges (all lines), faces (topological faces), and feature name files.

Each address that is processed will result in a certain level of geographic coding (street, zip, or city), depending on how well the address conformed to known addresses in the address locator data. Only addresses that can be matched to street level will result in detailed Census geography such as Census county, tract, and block group, and thereby obtaining the Census county, tract, and block group identifiers.

Each address will also receive a numeric score and a set of indicator flags about what did or did not match correctly about it. Certain types of address (e.g., PO Box, rural routes) lack a geographic specificity and others (e.g., highways, especially county road/highway/trunk) lack a standard format to be matchable at the street level. Based on the results of geocoding and trends in what was not matched, further corrections and standardizations will be attempted of the input data, and then geocoded again, for a refined match.

After finishing further corrections and standardizations, our research team members will extract the Census block group characteristics for years 2005-2012 in which we are interested (e.g., the percentage of persons in poverty, the percentage of households receiving public assistance income, the unemployment rate, median family income, racial composition, median house value, etc.) from the ACS and merge them into our dataset using Census block group identifiers.

[25] The TIGER/Line Shapefiles are "extracts of selected geographic and cartographic information from the U.S. Census Bureau's Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) database" (U.S. Census Bureau, 2013, p. 1-5). The shapefiles include "polygon boundaries of geographic areas and features, linear features including roads and hydrography, and point features" (U.S. Census Bureau, 2013, p. 1-5).

Step 9. Merging Data Files

Finally, we will merge education data, housing data, and geocoding data into a single data file using IRPID and use this file for our analysis.

IV. Challenges, Considerations, and Decisions

In this section, we will identify a number of challenges we confronted throughout this process and describe the steps that we have taken to address these challenges.

1. Unique Identifier with Different Names/Formats

To link multiple administrative data sources, each individual or case requires "a unique (to individual/case) but common (across datasets) identifier in each dataset" (Folsom, Osborne-Lampkin, & Herrington, 2014, p. 3). Although we had an IRPID and IRPCASEID that allowed us to link records across different administrative data sources, these identifiers were only available for administrative sources that were directly linkable with MSPF (e.g., data files from CARES, UI, and WISE). So, for education data files from WINSS, DPI assessment data, and CCD that are not directly linkable with MSPF, we needed to have other unique identifiers to link them: district code and school code.

However, when linking these multiple education data files, some files had identifiers with a different name or variable type (e.g., numeric vs. string). For instance, although all WISE, WINSS, DPI assessment data, and CCD had the same district code and school code information, variable names and types differed among them. Therefore, we renamed these variables and changed their variable types to insure that those are identical when referring to the same information. This change made the linking process smooth.

2. Discrepancy in Unique Identifier

When linking multiple data files from CARES, we needed common identifiers in order to link them. However, some files only had IRPIDs (e.g., wages, unemployment benefits, educational attainment, race/ethnicity, marital status), one file only had IRPCASEIDs (e.g., the housing subsidies file), and some files had both (e.g., Childcare Subsidies, Food Stamps, Medical Assistance, W2). To address this problem, we first created a universal IRPID-IRPCASEID link for each treatment or control group using the process described in Section III-Step 5, and then we applied it across all the data files. Then, we matched each data file with the universal link file.

3. Different Time Frames Covered by Datasets

In the case of data on secondary education, we were unable to use information from the WISE database in order to obtain the school-specific high school graduation rate variable because WISE data for this variable were available only from school year 2009-10 onwards. Therefore, we decided to use WINSS data that allowed us to get a school-specific high school graduation rate over the entire period. Also, as already mentioned above, the MSPF 2011 data stops at 2011 (calendar-year basis) but the DPI WISE data includes spring 2012 (school-year timeline). Therefore, to add in MSPF data covering spring 2012, we had to use the MSPF 2012.

4. Different Time Periods of Data Collection

Some data were collected annually (e.g., DPI education data), while other data were collected quarterly (e.g., UI wage records), monthly (e.g., W2, Food Stamps, Child Care Subsidies, Medical Assistance, UI benefit amount), or weekly (e.g., UI benefit spells). Moreover, the dates that the data were collected also varied. So, we chose to transform all of these data to an annual basis.

5. Different Definitions of Case Units

Data files such as W2, Child Care Subsidies, Medical Assistance, Food Stamps defined “case” or “household” differently. For instance, the Child Care file defined cases by the child covered and the PP, while the Food Stamps file defined cases as people who are eligible members of a Food Stamps case plus people who are the PP of a case even if they are not eligible for Food Stamps. The Medical Assistance file defined cases by the family members covered under a particular PP. For this reason, an individual might be associated with one household or case in terms of Child Care Subsidies and another in terms of Medical Assistance. Therefore, when we created IRPID-IRPCASEID links, their IRPID-IRPCASEID links were sometimes different across files. Also, we believe this might happened because the cases did not have well-defined end dates as well as because individuals could be associated with multiple cases at the same time.

6. No Reference Point when Coding Data

When coding data that was not collected on an annual basis, we needed to decide when to capture the status and demographics because those values may change over time. For instance, for the housing data, we used anyone who received a housing subsidy in April, May, June, July, August, or September of a given year. For the primary parent’s education, race/ethnicity, and marital status, as well as for the number of adults and children associated with the household, we use the value from August of a given year when possible. If data are not available for August, we use data from subsequent months in order (e.g., September, October, November, etc.).

Similarly, we had to choose which household a student/individual would be assigned to if the observation was associated with more than one household at the same time (and our rules for assignment are already explained in Section III-Step 5). The difficulty of making this assignment is attributable to three primary issues:

a) Individuals change households over time;
b) The linkage of individuals to households occasionally differs across data sources;

Also, we had to decide when during the year to assign a student/individual to a household; because students may not be in the same household in August and September, we had to choose which month to use as the basis for the “annual” measure. We used data from August, when possible, in order to make it correspond to the start of the school year. We also had to decide when to identify the PP. So, for Food Stamps data, we decided to identify PP in August of a given year; if August is missing, we used earlier (then later) months in order. For the CARES data, we identified PP in August; if August is missing, we used PP from September, October, November, or December in order.

7. Duplicate Cases

Student-level attendance and test score variables had duplicate IRPIDs because the records for some students were reported by more than one school. DPI staff judged that these duplicate cases occurred because of data entry errors or matching errors. Although less than one percent of the cases had duplicate IDs, we developed a duplicate cases managing process to systematically deal with this issue; this procedure was described in the Section III-Step 3.

In the IRP data, the central reason for the duplication problem is that the MSPF collects information from multiple data sources. In addition, while some of these sources include information on when the record began or was updated, few observations contain information on when the record ended (or was no longer valid). To address duplicates, we followed the rules in Section III-Step 5.

8. Missing Values

In the secondary education data, there were some students with missing data in all school years. IRP staff presumed that they might have been students in the prior school year that did not graduate and remained on the school rosters at the beginning of the next school year, but never actually attended school and never took any scheduled exams. IRP staff also suspected that missing values might be from the students who briefly stayed in a particular school system and transferred out after a short period of time. In this case, it is possible that they have no attendance or test score records. However, because we did not know exact reasons for missing values, we retained them and labeled them as missing.

In the MSPF data, information may be missing because individuals/households did not receive a particular benefit in that period, because the data were not recorded (e.g., individual’s education, race, or marital status, and all of the means-tested benefits), or because the MSPF does not contain the person’s social security number (without the SSN, wages and unemployment benefits cannot be matched with an individual).

9. Changes in the Variable Values

The race variables are not comparable between school years due to the change in variable values. For instance, beginning with 2010-11, the U.S. Department of Education required educational institutions to collect and report racial and ethnic data in accordance with modified standards and aggregation categories. In particular, two new categories were added increasing the number of categories reported from five[26] to seven[27] for almost all WISE data. So, we used five racial categories for our 2005-06 – 2009-10 WISE data. For 2010-11 and 2011-12 data, we collapsed seven categories into six categories by combining Asian and Native Hawaiian or other Pacific Islander.

10. Unclear Variable Definitions

Because the definitions for some variables were not clear, we turned to DPI for clarification or to make a 'best guess' as to the appropriate definition. Although the variables with the problem of definition were not used in our analysis, we wanted to know the exact meaning of them for thorough understanding of the dataset. For example, we wanted to know exactly "when" the age variable was collected, and exactly when "end-of-school-year" denotes to get some sense about other anomalistic cases in the data. However, we could not obtain a definite answer from the DPI. Hence, we assumed that age was collected on the third Friday of September each year because the original age variable includes "CD" at the end of its variable name. Any records from DPI with "CD" at the end of the variable name denotes that those records are based on the specified count date (= third Friday of September). In addition, we presumed that "end-of-school-year" is approximately the end of June each year, because WI DPI defines school year as "the time commencing with July 1 and ending with the next succeeding June 30."

11. Anomalous Cases with No Clear Explanations

Because administrative data are gathered as a by-product of program administration and not for the purpose of research, we encountered a number of anomalous cases without clear explanations about them. For instance, in the secondary education data, some students' school attendance days were less than ten days or the same as 365. Neither of these appears to be possible value for normal attendance days. We suspected that those cases with less than ten attendance days were students who transferred to other school out of Wisconsin during the school year. Also, we suspected that some teachers or administrators entered 365 for students who attended school without missing a single day during the school year. We requested DPI to elucidate whether these assumptions are correct, but we were unable to obtain any clear answers to these puzzling cases. Therefore, we just retained those cases.

In addition, some students dropped out of school at 6th grade or below—an implausible outcome. According to DPI, this event may be related to data quality, and other times, it might be related to the inability of the school to find the student after he/she leaves the school. Also, for some cases, a test score existed for non-tested grade students; again, DPI was unable to resolve this discrepancy. Hence, we just kept those cases.

Lastly, some records are identical except for values on one variable. For instance, in the marital status data, some observations have the same IRPID, begin date, updated date, and

[26] American Indian or Alaska Native, Asian or Pacific Islander, Black Not Hispanic, Hispanic, and White Not Hispanic

[27] American Indian or Alaska Native, Asian, Black or African American, Hispanic/Latino, Native Hawaiian or Other Pacific Islander, White, Two or More Races

sequence number but different values on the marital status variable. For these cases, we retained the observation with the "lower marital status," where 1 = married; 2 = single, never married; 3 = annulled, divorced, legally separated, or separated; and 4 = widowed. This was done for 604 cases or .000017 percent of the sample.

12. Data Quality Issues

In the secondary education data, some values of a variable did not have labels. For instance, there was a variable "agency type" when we were working with school-level test score data. According to the information provided by the DPI website[28], this variable had only four values: 03 (districts), 4C (multi-district charter schools), 49 (nondistrict charter schools), and 04 (public schools within district, includes charter schools other than 4C and 49). However, while we were doing data inspection, we saw that we also had a value of 10 for this variable, but we could not find a label for this value. So, we contacted the DPI to clarify the meaning of this value and received the answer that this 10 denotes two schools run by the state education agency—the Wisconsin School of the Deaf and the Wisconsin Center for the Blind and Visually Impaired. In addition, a value 4C of this agency type variable was mistyped as "04C", which we corrected to be "4C."

In the IRP data, we observed some children to be associated with from one to four mothers and with one to eight fathers. In this case, we used all of the information on parents. We coded a child as living with his/her mother or father if he/she was living with any one of the individuals identified as mothers or fathers. And for individuals who had an age below "-1," we coded their age as missing and did not count these individuals toward the total number of children and adults in the household.

V. Conclusion

In this paper, we have documented the detailed steps that we have taken, challenges that we have faced, and decisions that we have made to secure an integrated data set that can serve as the basis for the analysis of important relationships. These are likely to be similar to those confronted by other researchers attempting to merge information from several administrative data files. In summary, these involved the following:

a) We linked information on the observations available in one dataset with information on the same observations in other datasets. This involved establishing a unique identifier for each observation across the several administrative datasets.
b) We linked individuals (and their characteristics) to households (and their characteristics) because means-tested benefits are distributed to households, rather than individuals. This involved a set of complex decisions based on best judgment to establish the required linkage.
c) We addressed duplicates cases so that a data file has only one observation per individual or case. Because administrative data is not initially collected for research,

[28] http://winss.dpi.wi.gov/winss_data_download

it usually has messy structure and duplicate cases. This step also required reasonable decisions based on the information we had.

d) We created variables describing aggregations of individuals – e.g., school-level characteristics, household characteristics – from information on the individuals that comprised the groups of individuals.
e) We ensured that all individual/information was measured for the same time frame and time period. Also, when we needed to decide when to capture the status and demographics of individual or case, we set the reference point for coding data.
f) We dealt with inconsistent/unclear/anomalous information on the same individual or household observation based on the information we had and advice from agency people.

The processes required involved a host of judgments leading to decision rules. Each decision, therefore, interjects some arbitrariness into the final integrated data set. The goal is to minimize the number of such decisions, and in the cases in which they are required, to clearly document the nature of the decision and the basis for it. Although this process of merging multiple data sources was complicated, we now have a strong dataset that we can use to answer our research questions – the effect of receiving a housing voucher on children's educational outcomes. Because administrative data sources that researchers need to evaluate social programs are usually not collected or managed by one agency (University of California-Berkeley Data Archive & Technical Assistance [UC DATA], 1999), linking the information regarding individual and case over time and over programs enables researchers to see the broader picture of the program participants' experiences (Brady et al., 2001, UC DATA, 1999).

We conclude by suggesting the important things to be considered by researchers who are planning to link multiple administrative data sources for their studies. First, it is necessary to invest enough time to thoroughly comprehend the contents and structure of the administrative dataset. Since administrative records are not intentionally collected for research purposes, it is difficult to understand the complex structure of the raw administrative records and in most cases we expect that will require cleaning and restructuring to make them useful for research purposes. Therefore, understanding the meaning of the records, the way the records are collected, and how they are structured will help researchers to ease the data cleaning and merging process and to get a high quality final dataset.

Second, it is important to have a close working relationship with people in the program agency who are responsible for administrative records. Prompt and direct conversation with agency people when help regarding administrative records is needed will smooth the data managing process. Moreover, responses from the agency people improve the researcher's level of understanding about administrative data and will also increase the precision of the data managing process.

Third, it is important to record the whole process of data construction, conversation among researchers and agency people, and every decision made by researchers to resolve issues faced when managing data. Because the data cleaning and construction process is long and complicated, it is easy for researchers to get confused about their past ideas and decisions. Therefore, recording the detailed process of data managing will help remind researchers of important previous decisions and get an objective and accurate final dataset for their study.

References

Berger, L., Cancian, M., Han, E., Noyes, J., & Rios-Salas, V. (2014). *Education outcomes for children in foster care: Administrative data analysis in support of the Wisconsin education collaboration for youth in foster care* (Final Report). Madison: University of Wisconsin-Madison, Institute for Research on Poverty.

Brady, H. E., Grand, S. A., Powell, M. A., & Schink, W. (2001). Access and confidentiality issues with administrative data. In C. F. Citro, R. A. Moffitt, & M. Ver Ploeg (Eds.), *Studies of welfare populations: Data collection and research issues* (pp. 220-274). Washington, DC: National Academy Press. Retrieved from http://aspe.hhs.gov/hsp/welf-res-data-issues02/pdf/08.pdf

Brown, P. R. (with Ross, D., Smith, J. A., Thornton, K., & Wimer, L.) (2014). *Technical report on lessons learned in the development of the Institute for Research on Poverty's Multi-Sample Person File (MSPF) data system*. Madison: University of Wisconsin-Madison, Institute for Research on Poverty.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). *Expanding access to administrative data for research in the United States* (SBE 2020 White Paper). Washington, DC: National Science Foundation. Retrieved from NSF website: http://www.nsf.gov/sbe/sbe_2020/2020_pdfs/Card_David_112.pdf

Einav, L., & Levin, J. D. (2013). *The data revolution and economic analysis* (Working Paper No. 19035). Cambridge, MA: National Bureau of Economic Research. doi: 10.3386/w19035

Folsom, J. S., Osborne-Lampkin, L., & Herrington, C. D. (2014). *Using administrative data for research: A companion guide to A descriptive analysis of the principal workforce in Florida schools* (Report No. REL 2015-049). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from IES website: http://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_2015049.pdf

Hotz, V. J., Goerge, R., Balzekas, J., & Margolin, F. (Eds.) (1998). *Administrative data for policy-relevant research: Assessment of current utility and recommendations for development.* Northwestern University/University of Chicago Joint Center for Poverty Research. Retrieved from http://public.econ.duke.edu/~vjh3/working_papers/adm_data.pdf

National Forum on Education Statistics. (2010). *Traveling through time: The forum guide to Longitudinal Data Systems. Book one of four: What is an LDS?* (Report No. NFES 2010-805). Washington, DC: National Center for Education Statistics. Retrieved from NCES website: http://nces.ed.gov/pubs2010/2010805.pdf

The Wisconsin Department of Public Instruction. (2006, April). *Questions and answers regarding the new 2005-06 WKCE-CRT tests, scales, and cut scores*. Madison, WI.

Retrieved from http://oeahist.dpi.wi.gov/sites/default/files/imce/oea/pdf/q&a-sscrswlk.pdf

The Wisconsin Department of Public Instruction. (2013). *Administrator's interpretive guide*. Monterey, CA: CTB/McGraw-Hill. Retrieved from WI DPI website: http://oea.dpi.wi.gov/sites/default/files/imce/oea/pdf/adminguide.pdf

The Wisconsin Department of Public Instruction. (n.d.). WISEdash glossary. Retrieved May, 15, 2015 from WI DPI website: http://wise.dpi.wi.gov/wisedash_glossary

U.S. Census Bureau. (2008, October). *A compass for understanding and using American Community Survey data: What general data users need to know*. Washington, DC: Government Printing Office. Retrieved from U.S. Census Bureau website: http://www.census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf

U.S. Census Bureau. (2013, August). *2013 TIGER/Line Shapefiles technical documentation*. Washington, DC: Government Printing Office. Retrieved from https://assets.nhgis.org/original-data/gis/TIGER_2013_TechDoc.pdf

University of California-Berkeley Data Archive & Technical Assistance (UC DATA). (1999). *An inventory of research uses of administrative data in social services programs in the United States 1998*. Berkeley, CA. Retrieved from UC DATA website: http://ucdata.berkeley.edu/pubs/inventory/entire.pdf

Welcome to WINSS! (n.d.). Retrieved from http://winss.dpi.wi.gov/

WISEdash (n.d.). Retrieved from http://wisedash.dpi.wi.gov/Dashboard/portalHome.jsp