7

Accessing and Analyzing National Databases

Terrell Lamont Strayhorn
University of Tennessee

Introduction

There has been a plethora of published studies in education. Yet, despite the advancements of virtually every single line of inquiry in educational research from student achievement to access, faculty promotion to institutional finance, each study has a number of limitations and each design brings with it both virtues and dangers. While some studies are limited due to the lack of statistical controls for potentially confounding factors, others are limited by the number of respondents to the study's survey or the extent to which the study's findings are generalizable to other populations. To address some of these issues, federal, state, and other research organizations such as the National Center for Education Statistics (NCES) and National Science Foundation (NSF) have developed a number of large-scale databases, consisting of nationally representative samples, which can be used in secondary data analysis.

In this chapter, the use of large national databases in educational research at the K-12 and postsecondary level is discussed. Drawing on the existing literature, studies that employ national databases to examine a myriad of experiences and outcomes across all constituent groups including students and faculty/teachers are examined. Special attention is given to the opportunities that secondary analysis provides, the challenges associated with analyzing large-scale databases, and the supports available to secondary data analysts. Before readers can mine the idea of secondary data analysis for what it is worth—that is, the opportunities, challenges, and supports associated with this form of research—a general understanding of national databases and secondary analysis is warranted.

Secondary Data Analysis in Education

Secondary data analysis is a widely accepted form of educational research and K-12 and higher education analysts have used it extensively. For example, Crosnoe (2005) explored the experiences and outcomes of children from Mexican immigrant families using a national sample of 14,912 American kindergarteners who participated in the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) sponsored by the National Center for Education Statistics within the U.S. Department of Education. He conducted multilevel analyses and found that "children from Mexican immigrant families were overrepresented in schools with a wide variety of problematic characteristics, even when family background differences were taken into account" (p. 269). Other K-12 researchers have employed nationally representative samples drawn from the *National Education Longitudinal Study of 1988* (NELS:88/2000) to estimate the effects of working part-time in high school (Marsh & Kleitman, 2005; Singh & Mehmet, 2000; Steinberg, Fegley, & Dornbusch, 1993); opportunity to learn (Braun, Wang, Jenkins, & Weinbaum, 2006); and school and family contexts (Strayhorn, in press-b) on important educational outcomes such as student achievement.

Using large-scale national databases to study issues germane to postsecondary education has a long history as well. For instance, Carter (2001) used *Beginning Postsecondary Students Longitudinal Study* (BPS:90/92) data to examine the degree aspirations of Black and White college students. Zhang (2005) analyzed data drawn from the *Baccalaureate and Beyond Longitudinal Study* (B&B:93/97) to "examine the effect of college quality, among other academic and non-academic factors, on educational continuation for college graduates" (p. 314). But, even earlier researchers used large national databases to study dropouts of higher education (Munro, 1981) and attrition at 2- and 4-year colleges (Williamson & Creamer, 1988).

Focusing on another constituent group in higher education, Perna (2001) conducted hierarchical multinomial logit analyses on data from the *National Study of Postsecondary Faculty* (NSOPF:93) to determine if family responsibilities are related to the employment status of women and men junior faculty. NSOPF is the largest national survey of faculty members; it has been used extensively in prior research on postsecondary instructors (Bellas & Toutkoushian, 1999; Kirshstein, Matheson, & Jing, 1997; Rosser, 2004; Strayhorn & Saddler, 2007).

Even sociologists of education whose principal interests center on socioeconomic disparities, the academic profession, and educational organizations as a system characterized by an institutional division of labor (Clark, 1973; Clark & Trow, 1966) have analyzed data from secondary databases (Constantine, 1995; Mau & Kopischke, 2001). In a previous study data from the *Baccalaureate and Beyond Longitudinal Study of 1993* (B&B:93/97) were used to estimate the influence of "attending a historically Black college or university (HBCU)" on three labor market outcomes of recent African American college graduates: annual earnings, occupational status, and job satisfaction (Strayhorn, in press-a). Drawing on data from the same source, analyses were conducted on 11,192 individuals, representing approximately 1 million students nationwide, to estimate the effect of background, pre-college, and college experiences on the achievement disparities between first- and continuing-generation college students (Strayhorn, 2006a).

While our knowledge of national databases in education has burgeoned in the past few decades, researchers have focused their attention on presenting the *results* of secondary data analyses rather than the *mechanisms* through which national databases are accessed and analyzed. Such mechanisms are sometimes implied but rarely made explicit. Some book chapters point to the plethora of secondary data sources that is available in the public domain (Carter, 2003). Other researchers provide highly technical guidance about how to analyze large-scale national databases appropriately using adjusted sample sizes, panel weights, and design effects (Galloway, 2004; Thomas & Heck, 2001), to name a few. Yet, far less is known about how to access and analyze national databases, particularly those provided by federal organizations like the U.S. Department of Education, and what national databases provide in terms of opportunities and challenges. It is out of this context that the need for the present chapter grew.

Conceptual Framework

When I was in seminary at the Samuel Dewitt Proctor School of Theology at Virginia Union University, I studied with the late Rev. Dr. Myles Jones who taught homiletics or the art of preaching. He introduced me to the concept of an "organizing observation," which he generally referred to as "the OO." An organizing observation generally fit the following template, "There are times when . . . [something occurs]; when such conditions exist, we must . . . [do something]." When preparing to write this chapter, I reflected on the masterful teaching of Dr. Jones and found his idea of "the OO" particularly relevant to my approach in this chapter. Indeed, *there are times when* we, as educational researchers, want to enhance the generalizability of our findings; when such conditions exist, we must consider the important role that national databases play in advancing lines of inquiry within the social sciences (Carter, 2003). Indeed, there are other "conditions" under which the use of a national database is justified if not absolutely necessary. Based on this theoretical construct, part of the present chapter is devoted to identifying opportunities for using large-scale, nationally representative data in research.

To organize my thoughts about accessing and analyzing data drawn from nationally representative samples, I settled upon Sanford's (Sanford, 1966; Sanford & Adelson, 1962) notion of *challenge and support.* I found this conceptual framework useful as it provided constructs for talking about the issues involved in the process of securing and using secondary data in educational research. Also, Sanford's ideas are related, at least in part, to the "OO" described above. That is, I considered the antithesis of "challenge" which is generally believed to be "opportunity."

Drawing on these underlying concepts, the present chapter is organized into three main sections: opportunities, challenges, and supports. The final section presents an extended example of how researchers might employ a national database in higher education research. The balance of the chapter turns to potential research questions that can be investigated using large-scale national data.

Opportunities

This section focuses on the seemingly limitless opportunities provided by using national databases in educational research. In a way, it serves as an introduction to the sections that follow, which emphasize the challenges and supports associated with secondary analysis of large-scale survey data in social sciences and education.

Variety of Databases

First, the sheer number of national databases available represents a state of affairs that is opulent with opportunity for research purposes. Not only are there many providers of national databases such as the National Center for Education Statistics (NCES) within the U.S. Department of Education, National Science Foundation (NSF), Educational Testing Service (ETS), and Higher Education Research Institute (HERI), but each of these organizations provides access to a wide range of national databases that can support different sorts of research questions. For instance, quite often researchers want to measure the impact of educational policies on student outcomes; one might use National Postsecondary Student Aid Survey (NPSAS) data which are perfectly suited for exploring questions like: What percent of African American, low-income students received Pell grants before and after recent reauthorization of the Higher Education Act? Galloway (2004) proffered a useful five-category typology of research questions that are supported by large-scale national surveys—those measuring the relationship among variables, those testing the significance of group differences, those involving the prediction of group membership, those exploring the underlying structure of data, and those involving the time course of events. Researchers are encouraged to match their techniques to their questions, their questions to their database.

Yet, not all databases are made equal; there are important characteristics that distinguish national databases. For instance, some databases are appropriately titled population studies while others consist of nationally representative samples. Population, "census," or universe studies include information for every element (i.e., student, institution, etc.) in the larger group. IPEDS and Survey of Earned Doctorates (SED) are good examples of population studies. The SED is an annual census of all research doctorates awarded by U.S. degree-granting institutions. The database includes information on individuals in all fields, not just science and engineering. On the other hand, most national databases consist of nationally representative samples drawn from the population using complex, stratified sampling designs. *Baccalaureate and Beyond Longitudinal Study* (B&B), *High School and Beyond* (HSB), *Education Longitudinal Study* (ELS), *National Education Longitudinal Study* (NELS), *National Study of Postsecondary Faculty* (NSOPF), *Survey of Doctorate Recipients* (SDR) and *Cooperative Institutional Research Program* (CIRP) are good examples of nationally representative samples.

National databases may also differ in their purpose and intended use. For example, the NCES provides access to a wide range of databases that serve significantly different purposes, contain remarkably different samples, and can be used to answer qualitatively different research questions. The following list describes the nature and purpose of several databases available through federal agencies and national research centers:

1. *National Education Longitudinal Study* (NELS): The NELS:1988/2000 tracks 24,599 individuals, who represent the national population of eighth graders in 1988, up to eight years after high school graduation. NELS consists of over 6,000 variables and includes surveys from students, teachers, parents, and administrators in a series of data collection waves (Curtin, Ingels, Wu, & Heuer, 2002). NELS is one of the most widely used databases in educational research. Researchers have studied the effects of motivation and academic engagement on math and science achievement (Singh, Granville, & Dika, 2002), race and academic disidentification (Osborne, 1997), and the effects of social capital on students' transition to selective colleges (Kim & Schneider, 2005) using NELS data.
2. *Baccalaureate and Beyond Longitudinal Study* (B&B): The B&B study follows baccalaureate degree completers over time to provide information on work experiences after college and post-BA outcomes such as postgraduate educational experiences, earnings, occupation, and job satisfaction to name a few. This database is particularly useful for studying the long-term effects of college on student outcomes. In prior studies, B&B data have been used to examine the decision to enroll in college (Perna, 2004), the impact of undergraduate college selectivity on enrollment in graduate school (Zhang, 2005), and graduate student persistence (Strayhorn, 2005).
3. *Education Longitudinal Study of 2002* (ELS): According to the NCES website, ELS is "designed to monitor the transition of a national sample of young people as they progress from tenth grade through high school and on to postsecondary education and/or the world of work" (National Center for Education Statistics, n.d.). The database is ideal for studying critical transitions from high school to college, pathways to postsecondary education, the role parents play in their child's success, and the influence of course-taking patterns on subsequent achievement. Given its relative newness, the ELS database has not been used extensively in education research; researchers have used ELS data to study the initial college experiences of high school seniors (Bozick & Lauff, 2007).
4. *College Student Experiences Questionnaire* (CSEQ): The CSEQ consists of 191 items designed to measure the quality and quantity of students' involvement in college activities and their use of college facilities. Other items measure students' sociodemographic characteristics, educational aspirations, and perceived development across an array of learning domains including critical thinking, self-awareness, and appreciation of diversity (Pace, 1990). More than 500 colleges and universities in the United States have used the questionnaire. In previous studies, CSEQ data have been used to understand the effect of involvement for employed students (Lundberg & Schreiner, 2004); faculty–student interactions and college grades (Anaya & Cole, 2001); and the impact of collegiate experiences on self-assessed skills (Grayson, 1999). There is an adapted version of the CSEQ primarily designed for students at two-year community colleges called the Community College Student Experiences Questionnaire (CCSEQ). For more information, consult the primary author (Friedlander, Murrell, & MacDougall, 1993; Friedlander, Pace, & Lehman, 1990).
5. *Cooperative Institutional Research Program* (CIRP): The CIRP is the oldest and largest empirical study in higher education focusing on entering college students. Initiated in 1966 by Alexander Astin, the survey provides useful

information on entering students' academic preparation for college, high school activities, academic major, values, and demographic characteristics, to name a few. This database is particularly useful for studying the predisposition of high school students to college, self-reported competencies, and how college affects student learning and development. Previous researchers have used CIRP data to study student leadership development (Kezar & Moriarty, 2000), faculty–student interactions (Cole, 1999), and citizenship and spirituality (Sax, 2004).

Table 7.1 presents a list of several national databases by level of education; while this list does not include all databases that are available to secondary analysts, it includes those that are most commonly used.

Table 7.1 National databases in education, by level of schooling.

Early childhood/elementary	*Secondary*	*Postsecondary and beyond*
Early Childhood Longitudinal Study (ECLS) +	Education Longitudinal Study of 2002 (ELS) +	Baccalaureate and Beyond Longitudinal Study (B&B) +
National Household Education Survey +	National Household Education Survey +	Beginning Postsecondary Students Longitudinal Study (BPS) +
Crime and Safety Surveys +	Crime and Safety Surveys +	Integrated Postsecondary Education System (IPEDS) +
	National Education Longitudinal Study of 1988 (NELS) +	National Education Longitudinal Study of 1988 (NELS) +
	High School and Beyond Longitudinal Study (HSB) +	High School and Beyond Longitudinal Study (HSB) +
	National Longitudinal Study of the High School Class of 1972 (NLS: 72) +	National Longitudinal Study of the High School Class of 1972 (NLS: 72) +
		National Postsecondary Student Aid Study (NPSAS) +
		National Study of Postsecondary Faculty (NSOPF) +
		National Survey of Student Engagement (NSSE) *
		College Student Experiences Questionnaire (CSEQ) *
		Survey of Earned Doctorates (SED) #
		Recent College Graduates Survey (RCG) #
		Cooperative Institutional Research Program (CIRP) *

(+) provided by the National Center for Education Statistics.
(#) provided by the National Science Foundation.
(*) provided by an independent research center.

Maximizing Generalizability

Results produced by large-scale secondary data analyses are inferentially robust. And, more often than not, national databases are used for the expressed purpose of drawing generalizations about behaviors, trends, or patterns in the broader population of "units of analysis" (i.e., institutions, students, faculty/staff) in higher education. This is accomplished in a number of ways. First, national databases such as those provided by NCES and NSF consist of entire populations or large samples. For example, the Integrated Postsecondary Education Data System (IPEDS) is a population study that contains *every* single element in the broader group (i.e., postsecondary institutions). IPEDS, also known as the postsecondary institution universe survey, contains information on 6,600 institutions plus 80 administrative units. As another example, *High School and Beyond* (HSB) is a longitudinal study of 28,000 high school seniors in the United States. Studies based on such samples are more representative of the broader population than any sample that individual researchers can generate without enormous resources and time.

Nevertheless, gains in generalizability come at the expense of specificity to some degree. In other words, findings based on nationally representative samples are more representative of the broader population (i.e., students, teachers, women) but may not reflect subtle nuances that exist for a particular school, a specific classroom, "exceptional" individuals, or a campus that has established a new program or service that aids students' adjustment to college. In such instances, locally designed surveys may be a more appropriate strategy for assessing the impact of educational practices on specific outcomes (Strayhorn, 2006b).

There is at least one other way in which nationally representative analyses require analysts to give up a degree of specificity. When using primary data sources (e.g., locally designed surveys, commercial questionnaires administered to individuals directly), researchers have more control over the way in which variables are measured. For instance, in a study of first-year college student experiences at a single institution, items were developed for measuring students' satisfaction with college and sense of belonging (Strayhorn & Blakewood, 2007, 2008). Satisfaction was assessed in multiple areas including students' satisfaction with campus life, individual support services (e.g., academic advising, cultural centers, etc.), and their relationships with faculty members to name a few. Also, satisfaction with college was assessed generally using multiple items; an example of this scale is, "I am satisfied now with my college experience at [said institution]." This allowed my collaborators and myself to tap various facets of student satisfaction in a highly specific way at a single institution.

To test whether the determinants of first-year student satisfaction at "said university" are similar to those found in the broader population of college-going students, multivariate analyses were conducted on a nationally representative sample of students who participated in the 2004–2005 administration of the College Student Experiences Questionnaire (CSEQ) sponsored by the Center for Postsecondary Research at Indiana University. To access data, a research proposal was sent to the Center for Postsecondary Research along with the paperwork required to request a slice of CSEQ data (for more information, see the Center's website). Specifically, a large, random sample of students who matched certain criteria (e.g., race, institutional type, etc.) was obtained. To analyze these data, measures were identified in the database that were appropriate for operationalizing student satisfaction and sense of belonging. However, my choice of "proxies" for these two constructs was constrained by decisions made by those who created the CSEQ (Pace, 1990). According to the codebook and related sources (Gonyea, Kish, Kuh, Muthiah, & Thomas, 2003), several items on the CSEQ have psychometric qualities that are consistent with student satisfaction with college; an example of this scale is, "how well do you like college?" While there is considerable agreement among researchers about the use of these items in satisfaction studies (Hollins, 2003; Kuh & Hu, 2001; Strayhorn & Terrell, 2007) and the *generalizability* of findings based on national samples (Thomas & Heck, 2001), it is plausible that CSEQ items have a marginal relationship (lack *specificity*) with the constructs they are purported to measure. Using CSEQ data to increase the reliability of my estimates also required me to surrender a degree of precision in measuring constructs like satisfaction; while important to this discussion, such trade-offs are inevitable in secondary data analyses and do not necessarily limit the usefulness of studies based on large-scale survey samples.

Combining Data Sources

The development and use of national databases in education afford researchers the opportunity to combine information from multiple sources to create powerfully robust datasets for secondary analysis. To be sure, this presents an unparalleled occasion to calculate highly reliable parameter estimates that might otherwise require decades of data collection, inordinate amounts of time, millions of dollars in survey design and storage, and an unfathomable number of hours for participants in responding to surveys, questionnaires, and interviews which most institutional review boards would consider unnecessarily stressful, if not inhumane.

To avoid such "costs," many national databases either (a) draw information from previously existing sources or (b) include an identifier that can be used to merge information from multiple sources into a single dataset. To illustrate the former, National Postsecondary Student Aid Study (NPSAS) serves as the base-year on which several other NCES databases are built. Using NPSAS:93 as the base year, the B&B:1993/2003 follows baccalaureate (BA) degree completers up to 10 years after college graduation (Wine, Cominole, Wheeless, Dudley, & Franklin, 2006). That is, from the NPSAS:93 sampling criteria, 16,316 baccalaureate degree recipients were identified and surveyed in the base year (1993), the first-year follow-up (B&B:93/94), the second follow-up (B&B:93/97), and the third follow-up (B&B:93/03). The resulting database provides information on how students and their families pay for college (NPSAS) as well as their post-BA experiences (B&B). Similarly, the *Beginning Postsecondary Students* (BPS) *Longitudinal Study* collects information on the undergraduate experiences of students who responded to the NPSAS initially (Wine et al., 2002).

Finally, developers of national databases may often include an "identifying" variable that can be used to combine multiple datasets and to connect various units of analysis. For instance, several NCES databases consisting of stratified samples (i.e., students within schools) include identifiers for the individual and the institution that allow analysts to combine student- and institution-level data for multilevel analyses. The *Baccalaureate and Beyond Longitudinal Study* (B&B:1993/2003) includes information at the student level ranging from demographic variables such as age, race, and gender to academic factors such as undergraduate grade point average (GPA), academic major, and degree goals (Green et al., 1996). Institution-level data can be merged into the B&B from extant data sources like IPEDS and Barron's *Profiles of American Colleges* using the institution's name or IPEDS code (usually labeled "unitID"). For example, Zhang (2005) created an integrated B&B database by extracting information on institutional control from IPEDS and college selectivity from various editions of the *Barron's Guide*. In a previous paper, measures were derived of institutional control, selectivity, and campus type (i.e., historically Black or not) from IPEDS data and merged with B&B:1993/1997 information to create an expanded panel study (Strayhorn, in press-a).

It is important to note, however, that identifying information is included in "restricted datasets" only and not those available in the public domain. Analysts must secure a license to access restricted databases; this issue is addressed in the next section. However, research centers like the Higher Education Research Institute (HERI) at the University of California, Los Angeles, and the Center for Postsecondary Research at Indiana University have their own application process for analysts who desire to use their data. Readers are encouraged to contact individual agencies for information about their application process.

Challenges

Indeed, secondary analysis of data from nationally representative samples provides a number of unique opportunities for advancing lines of scholarly inquiry. However, secondary analysis of national data is rife with a number of challenges as well. Generally, these challenges can be organized into two areas: challenges associated with accessing and analyzing databases.

Accessing Databases

Restricted Use Licenses As previously mentioned, identifying information (i.e., student- and institution-level identifiers, continuous data on earnings, etc.) is included in restricted datasets only and not those available in the public domain. For this reason, analysts must secure a license to access restricted use databases. For NCES and NSF databases, the process can be simplified into five steps:

(1) access license application forms (NCES now uses an online application process);
(2) submit a research proposal that identifies the database requested, its intended purpose, and the researcher's intended use including research questions, hypotheses, analytic plan, and outcomes or products that might emanate from such analyses;

(3) submit a license agreement signed by the principal investigator (PI) or project officer and a senior official with authority to bind the organization to the terms of the license;
(4) submit signed affidavits of nondisclosure for everyone involved in the research project (e.g., PI, research assistants, collaborators), each bearing the seal of a notary public; and
(5) a detailed security plan signed by the individuals listed in #3 plus a system security officer.

It is important to note that federal organizations like NCES and NSF grant restricted use licenses to organizations (i.e., research centers, educational institutions, etc.) not individuals. In other words, individuals apply on behalf of their institution. This is clarified in the license agreement statement described above.

Cost Another challenge associated with gaining access to national databases is cost. Cost is fees that must be paid either to receive a copy of the database on compact disc or for the time and effort of staff at the "data granting" organization to prepare the dataset for secondary use. In some instances, staff not only prepare the dataset but also analyze the data and provide external researchers with the necessary tables, graphs, and output. Rates for compiling data on specific schools, individuals, or scales can range from $100 per hour and up. For instance, a simple random sample of 8,000 individuals was purchased, in a previous study, for approximately $1,500. Those interested in conducting secondary data analysis would do well to secure grant funds to subsidize such costs. The American Education Research Association (AERA) awards up to $20,000 to individual researchers who conduct studies using large-scale survey data provided by the NCES and NSF. For more information, go to AERA's grants program online.

Analyzing Databases

Missing Data Secondary analysis of national databases is often complicated by the amount of missing cases or missing data (Graham & Hoffer, 2000; Little & Rubin, 1987). As a general rule, cases, for which data are missing completely at random, tend to be dropped from analyses. However, if patterns are observed in the "missingness" of data (e.g., family income and student aid amounts are missing for many more minorities than majority students), analysts must take steps to account for missing information. There are at least three commonly used solutions to this problem:

(1) dropping missing observations for analyses,
(2) imputing the sample mean in place of the missing information known as the *zero-order correction* procedure, and
(3) predicting or forecasting the missing information on Y_1 (e.g., family income) by estimating other models based on non-missing variables (e.g., gender, race, parent's education) that are correlated with Y_1.

If missing information on the dependent variable, analysts have no option but to drop the case from all analyses.

Complex Sampling Designs In most cases, nationally representative data were collected using complex sampling designs. That is, most large-scale survey data were collected using stratified, multistage, cluster sampling techniques or "sample designs that involve nesting observations at one level within higher-order units at another level" (e.g., students within schools, families within neighborhoods, faculty within institutions, etc.). Nested strategies are further complicated by the oversampling of *certain* individuals in *certain* situations that need to be included in sufficient numbers for the purposes of analysis. A good example of this is the B&B database that employed a stratified multistage sampling design and oversampled for education majors. Without proper adjustments, estimates of variances and standard errors derived from such data would be at least biased and at worst incorrect, leading to inaccurate statistical conclusions or Type I errors (Muthen & Satorra, 1995).

Most datasets, especially those provided by the NCES, provide a set of sample weights to adjust for unequal probabilities of selection due to the sampling design. A detailed discussion of various types of weights goes beyond the limits of this discussion (Thomas & Heck, 2001), but it is important to note that different weights (e.g., raw versus relative) affect differently standard errors and sample sizes, upon which most statistical conclusions are drawn. Thus, applying the appropriate relative weight to national databases using complex sampling designs is a critical adjustment.

It is important for readers to note, however, that considerable debate exists (and persists) about how to remedy potential problems associated with secondary data analysis (Thomas, Heck, & Bauer, 2005). Most scholars agree that decisions about dealing with complex samples should be based on the conceptualization of one's study. For instance, researchers must apply appropriate sample weights to account for the unequal probability of selection when using National Study of Postsecondary Faculty (NSOPF) data to study gender equity in terms of salary among tenured full-time professors. However, if the study was designed to measure differences in compensation among faculty members employed at two-year and four-year institutions, researchers would do well to (a) estimate the model using specialized software (e.g., SUDAAN, WesVar, AM) that takes into account the complex sampling design, (b) apply appropriate sample weights to adjust for unequal selection rates, and (c) adjust estimated standard errors upward using the design effect (DEFF) to account for the intracluster homogeneity of variance. Again, the decision should be based on the study's design and purpose, and researchers are encouraged to link databases to questions, questions to techniques, and designs to adjustment decisions.

Supports

Despite the challenges listed above, a number of supports are available to assist secondary data analysts in their work with large-scale national databases.

Codebooks and Surveys To analyze secondary databases, one must know which questions to ask and why, how to interpret or make meaning of the data addressing those questions, the limitations of the instruments and methods selected, how data were collected including sampling designs and design effects, and how to best report findings to those concerned. To this end, most developers of large-scale secondary data sources produce copies of the actual instrument and codebooks that explain how items were worded, response options for each item, and the meaning of each response category (e.g., 1 = *female*; 5 = *very often*, etc.). Technical reports also provide information on data collection procedures (i.e., paper survey, computer-assisted telephone interview) and design effects that should be considered when analyzing and reporting findings based on national data. NCES provides electronic copies of its methodological reports and codebooks on the center's website (http://nces.ed.gov).

Data Analysis Software Technological advances in terms of data analysis software also support the work of secondary analysts. Standard statistical packages (e.g., SPSS and SAS) can be used to conduct single-level studies on national data. Advanced software packages also exist that specialize in adjusting for complex sampling designs, accounting for multi-level clustering (i.e., faculty within institutions), and applying pre-determined design effects (DEFF). These packages range from expensive, sophisticated programs like SUDAAN, WesVar, PCCARP, and HLM (Raudenbush & Bryk, 2002) to free programs like *AM* v.0.06 provided by the American Institutes for Research. All of these programs make it relatively easy to apply sample weights (e.g., WEIGHTBY function in SPSS) and impute sample means for missing cases. Finally, users may also use NCES' *Data Analysis System Online* (DAS) to create descriptive tables or calculate correlation matrices.

Theory Analyzing national databases often requires the researcher to operationalize various constructs, or complex abstractions, using multiple items from the dataset. Theory is a powerful support or tool for making sense of such "complex abstractions." In other words, theory, by definition, presents a plausible explanation for observed phenomena and may provide clues for measuring variables that are difficult to observe or quantify otherwise such as motivation, self-concept, or self-esteem to name a few.

Indeed, there are a number of theories that can prove useful when measuring independent and dependent factors in educational research. For instance, research studies, based on nationally representative samples, might be grounded in theoretical underpinnings related to cognitive-structural, psychosocial, attribution, and college impact theories. This list is not exhaustive; rather it provides a starting place for those who need assistance with identifying theories that might be employed in large-scale national analyses. Examples of published research studies, grounded in theory, that use national databases abound (Perna, 2001, 2004; Stage & Rushin, 1993; Strayhorn, 2006a, 2006c).

In this way, theory serves as a guide to support secondary data analysts in their work with large-scale surveys that employ complex sampling designs and yield information on thousands of items that, in isolation, may hold little meaning (e.g., "I feel good about myself at times") but taken together may measure important theoretical constructions (e.g., self-concept, self-esteem, identity, etc.).

Workshops/Training Seminars There are workshops and training seminars available for those who are concerned about analyzing data from nationally representative samples. Several of these are provided by the U.S. Department of Education such as "Using ELS:2002 and NELS:88 for Research and Policy" (http://iesed.gov/whatsnew/conferences).

Professional associations in education sponsor a number of training seminars as well. The Association of Institutional Research sponsors a summer data policy institute, with support from NCES and NSF, to train researchers on the use of federal databases in educational research. Also, the American Educational Research Association provides pre-conference workshops on analyzing secondary data, in advance of its annual meeting, for a fee (ranging from $85–$150). In 2007, I served as director of such a pre-conference course that enrolled approximately 50 participants. Participants represented a range of experiences from doctoral students and faculty members in K-12 and higher education, to institutional researchers and researchers from national organizations such as the American Council on Education. Readers are encouraged to consider these professional development options as many are designed to offer hands-on experience in analyzing secondary data under the guidance of expert users of large-scale national data.

Relevant Readings Finally, in consonance with my professional identity as a graduate faculty member, the set of readings, handbooks, and articles found in this chapter's list of references are recommended for interested persons. This list is not designed to be exhaustive but rather a mere illustration of the kinds of supports available to novice and advanced secondary analysts. Indeed, there is more to be read, including the present chapter, and most good readings end with more stuff to read. Thus, readers are encouraged to "go where the path leads" until you feel equipped with enough information to design a study carefully and to analyze secondary data appropriately.

An Extended Illustration

A Chinese proverb reads, "Example carries more weight than preaching." This wisdom is relevant to using national databases for research purposes. However, a modifier is warranted—a "good" example carries much weight. The chapter's author created the anecdote presented below with this goal in mind—to provide a good, albeit brief, illustration of how large-scale national datasets can be employed in research studies.

> Fred Panera is the advisor of Latino-American students at the University of Knox. For several years, he has expressed concerns about the relatively low representation of Latino students on campus and the limited number of opportunities available in which Latino students can be involved. After reading through a recent article in the *Chronicle* about low co-curricular involvement rates among Latino males and the important role that summer bridge programs play in the socialization and success of minorities in college, he logged three questions in the back of his day planner:

(1) What affect do summer bridge programs have on Latino student success in college?
(2) What affect does campus involvement have on Latino student success in college?
(3) Do these effects differ between men and women?

After reading published works on student involvement, summer bridge programs, and Latino students in college, Panera learned that first-year Latino students faced unique challenges related to student success. He decides to explore his questions by way of research. With few staff to assign to this project and limited resources to collect data, Panera decides to use pre-existing data to probe his concerns. Using his reflection as a guide, the following questions were adopted for the study:

(1) What affect do precollege programs have on *first-year* Latino student success in college?
(2) What affect does campus involvement have on *first-year* Latino student success in college?
(3) Do these effects differ between men and women?

After talking with Dr. Edwards, a faculty member in the college of education about his interests, Fred is encouraged to consider the *Beginning Postsecondary Students Longitudinal Study* (BPS:1996/2001) which tracks first-time entering collegians up to six years after enrolling initially. He applies for a restricted use license through the NCES website, identifying himself and Dr. Edwards as principal investigators. After receiving the data, Panera struggles to identify suitable proxies for his variables of interest. Returning to the literature that he read earlier, he decided to draw upon extant theories to develop "working" or operational definitions for success outcomes (i.e., institutional selectivity, first-year grades, first-year retention) and campus involvement (i.e., individual items from the social integration subscale in BPS). Since BPS does not include information on actual participation in a TRIO program, the team modified their focus on *being eligible* to participate in TRIO programs; this reflects the issue of specificity to which I alluded earlier in the chapter.

Given the stratified, multi-stage complex sampling design employed in BPS, Dr. Edwards chooses the appropriate sample weight from the database. As a prior user of federal databases, he is comfortable calculating the relative weight by dividing the raw weight (provided in BPS) by its mean: $w/\bar{w}$. Using the survey design effect (DEFF) published in the methodology report, Dr. Edwards calculates the design effect adjusted weight by dividing the relative weight by the DEFF. Using the "WEIGHTBY" function in SPSS, they conduct regression analyses on a nationally representative sample of Latino students—regressing first-year GPA (SECPAY1) on measures of involvement in clubs and organizations (CMCLUBS, CMINTRAM) and whether the respondent was eligible to participate in a pre-college outreach or TRIO program (e.g., Upward Bound).

To test for differences within group, they created interaction terms between "gender" and their independent variables. These cross-product variables were entered in the next step of the equation. Increased variance explained at the $p < 0.01$ level suggests the presence of conditional effects. The researchers decided to use a more rigorous threshold of statistical significance (0.01 versus 0.05) in accordance with recommendations provided by Thomas, Heck, and Bauer (2005); using a more conservative critical alpha value is one of their "corrective strategies" for adjusting

for design effects. Reports were presented in tabular and written form to administrators on campus. For instance, using information found in Table 7.2, they presented evidence to student affairs staff members of the benefits that campus involvement provides to students, especially Latino students. In their study, Latino students who were involved in campus clubs and activities earned higher first-year GPAs than those who were not involved in clubs and organizations. That is, the predicted first-year GPA of uninvolved students was 2.22 while the predicted GPA of involved students ranged from 2.26 to 2.29. With this information in mind, the group considered barriers to Latino students' involvement in campus clubs and organizing including the limited number of ethnic and multicultural student organizations on campus (e.g., Latino Student Union) and the virtual absence of Latino faculty and staff members who might serve as advisors and mentors to such students.

Future Research Using National Databases

Future research based on national databases might pursue a number of promising directions. Similar to the example above, future researchers might use BPS data to study the college experiences of students who qualify for federal TRIO programs (e.g., low-income, first-generation minorities). Data from the Survey of Earned Doctorates can be used to study comparatively the educational pathways of Asian scientists in engineering and the social sciences. Recent data from ELS:2002 can be used to study science course taking in high school, opportunity to learn math skills, and even the college choices of students grounded in prevailing theory about the college decision-making process (Hossler & Gallagher, 1987).

Darling-Hammond (2007) noted that "for students of color, the [educational] pipeline leaks more profusely at every juncture" (p. 318). To test the validity of this statement or perhaps more appropriately to provide evidence of the "leaking pipeline," future research might use longitudinal databases like the ELS, NELS, or BPS to measure the proportion of minority students who move, or fail to move, across critical junctures between middle school and high school, high school and college,

Table 7.2 First-year GPA regressed on campus involvement variables, BPS:1996/2001.

Variable	*B*	*SE*	*t*
Intercept	222.20	11.84	18.77
Clubs			
Sometimes	6.70	3.94	1.70*
Often	3.45	6.63	0.52*
Never (reference)			
Intramural sports			
Sometimes	4.31	6.38	0.68
Often	−18.08	4.70	−3.85*
Never (reference)			

$R = 0.18$, $R^2 = 0.03$.
* $p < 0.01$.

college and graduate school. To further this line of inquiry, future studies may also be designed to model the way in which various factors at multiple levels (i.e., individual, family, school) coalesce to affect future educational opportunities. An example of this kind of study might employ NELS data and Bronfenbrenner's (1979) theory to explore the question: Are parental expectations, teachers' perceptions, and students' aspirations related to degree attainment for African American collegians after controlling for differences in human capital and prior academic achievement?

Conclusion

This chapter extends our understanding about the nature of national databases in education, the opportunities that secondary data analysis provides, and the challenges and supports associated with accessing and analyzing national databases. While many research questions linger, research in education has benefited greatly from the availability of relatively high-quality data through large-scale national databases. Using this chapter as a guide, researchers, educators, and policy makers can easily access and appropriately analyze national databases for decision making in education.

References

Anaya, G., & Cole, D. (2001). Latina/o student achievement: Exploring the influence of student–faculty interactions on college grades. *Journal of College Student Development, 42*(1), 3–14.

Bellas, M. L., & Toutkoushian, R. K. (1999). Faculty time allocations and research productivity: Gender, race and family effects. *Review of Higher Education, 22*, 367–390.

Bozick, R., & Lauff, E. (2007). *Education Longitudinal Study of 2002 (ELS:2002): A first look at the initial postsecondary experiences of the high school sophomore class of 2002*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black–White achievement gap: Do state policies matter? [Electronic version]. *Educational Policy Analysis Archives, 14*(8).

Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.

Carter, D. F. (2001). *A dream deferred? Examining the degree aspirations of African American and White college students*. New York: RoutledgeFalmer.

Carter, D. F. (2003). Secondary analysis of data. In F. K. Stage & K. Manning (Eds.), *Research in the college context: Approaches and methods* (pp. 153–167). New York: Brunner-Routledge.

Clark, B. R. (1973). Development of the sociology of higher education. *Sociology of Education, 46*, 2–14.

Clark, R., & Trow, M. (1966). The organizational context. In T. M. Newcomb & E. K. Wilson (Eds.), *College peer groups: Problems and prospects for research*. Hawthorne, NY: Aldine de Gruyter.

Cole, D. G. (1999). *Faculty–student interactions of African American and White college students at predominantly White institutions*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Constantine, J. M. (1995). The effects of attending historically Black colleges and universities on future wages of Black students. *Industrial and Labor Relations Review, 48*(3), 531–546.

Crosnoe, R. (2005). Double disadvantage or signs of resilience? The elementary school contexts of children from Mexican immigrant families. *American Educational Research Journal, 42*(2), 269–303.

Curtin, T., Ingels, S., Wu, S., & Heuer, R. E. (2002). *NELS 1988/2000: Base year to fourth follow-up data user's manual*. Washington, DC: National Center for Education Statistics.

Darling-Hammond, L. (2007). The flat earth and education: How America's commitment to equity will determine our future. *Educational Researcher, 36*(6), 318–334.

Friedlander, J., Murrell, P. H., & MacDougall, P. R. (1993). The community college student experiences questionnaire. In T. W. Banta (Ed.), *Making a difference: Outcomes of a decade of assessment in higher education* (pp. 196–210). San Francisco: Jossey-Bass.

Friedlander, J., Pace, C. R., & Lehman, P. W. (1990). *Community college student experiences questionnaire*. Los Angeles: University of California, Center for the Study of Evaluation.

Galloway, F. J. (2004). *A methodological primer for conducting quantitative research in postsecondary education at Lumina Foundation for Education*. Retrieved November 27, 2004, from www.luminafoundation.org/research/researchersgalloway.pdf

Gonyea, R. M., Kish, K. A., Kuh, G. D., Muthiah, R. N., & Thomas, A. D. (2003). *College student experiences questionnaire: Norms for the fourth edition*. Bloomington, IN: Indiana University Center for Postsecondary Research, Policy, and Planning.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201–218). Mahwah, NJ: Erlbaum.

Grayson, J. P. (1999). The impact of university experiences on self-assessed skills. *Journal of College Student Development, 40*, 687–699.

Green, P. J., Meyers, S. L., Giese, P., Law, J., Speizer, H. M., & Tardino, V. S. (1996). *Baccalaureate and beyond longitudinal study: 1993/94 First follow-up methodology report* (NCES Report No. 96–149). Washington, DC: U.S. Government Printing Office.

Hollins, T. N. (2003). *Participation in an extended orientation course and its relationship with student involvement, student satisfaction, academic performance, and student retention*. Unpublished doctoral dissertation, Florida State University, Tallahassee.

Hossler, D., & Gallagher, K. S. (1987). Studying student college choice: A three-phase model and the implications for policymakers. *College and University, 62*, 207–221.

Kezar, A., & Moriarty, D. (2000). Expanding our understanding of student leadership development: A study exploring gender and ethnic identity. *Journal of College Student Development, 41*(1), 55–69.

Kim, D. H., & Schneider, B. (2005). Social capital in action: Alignment of parental support in adolescents' transition to postsecondary education. *Social Forces, 84*(2), 1181–1206.

Kirshstein, R. J., Matheson, N., & Jing, Z. (1997). *Instructional faculty and staff in higher education institutions: Fall 1987 and Fall 1992* (NCES Report No. 97–447). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kuh, G. D., & Hu, S. (2001). The effects of student–faculty interaction in the 1990s. *Review of Higher Education, 24*, 309–332.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: J. Wiley & Sons.

Lundberg, C. A., & Schreiner, L. A. (2004). Quality and frequency of faculty–student interaction as predictors of learning: An analysis by student race/ethnicity. *Journal of College Student Development, 45*(5), 549–565.

Marsh, H. W., & Kleitman, S. (2005). Consequences of employment during high school: Character building, subversion of academic goals, or a threshold? *American Educational Research Journal, 42*(2), 331–369.

Mau, W., & Kopischke, A. (2001). Job search methods, job search outcomes, and job satisfaction of college graduates: A comparison of race and sex. *Journal of Employment Counseling, 38*, 141–149.

Munro, B. (1981). Dropouts from higher education: Path analysis of a national sample. *American Educational Research Journal, 18*(2), 133–141.

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association.

National Center for Education Statistics (n.d.). *Education longitudinal study of 2002*. Retrieved January 2, 2008, from http://nces.ed.gov/surveys/els2002

Osborne, J. W. (1997). Race and academic disidentification. *Journal of Educational Psychology, 89*(4), 728–735.

Pace, C. R. (1990). *College student experiences questionnaire* (3rd ed.). Los Angeles: University of California, Center for the Study of Evaluation, Graduate School of Education.

Perna, L. W. (2001). The relationship between family responsibilities and employment status among college and university faculty. *Journal of Higher Education, 72*(5), 584–611.

Perna, L. W. (2004). Understanding the decision to enroll in graduate school: Sex and racial/ethnic group differences. *Journal of Higher Education, 75*(5), 487–527.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rosser, V. J. (2004). Faculty members' intentions to leave: A national study on their worklife and satisfaction. *Research in Higher Education, 45*(3), 285–309.

Sanford, N. (1966). *Self and society: Social change and individual development*. New York: Atherton.

Sanford, N., & Adelson, J. (1962). *The American college: A psychological and social interpretation of higher learning*. New York: Wiley.

Sax, L. (2004). Citizenship and spirituality among college students: What have we learned and where are we headed? *Journal of College and Character* 2. Retrieved December 29, 2004, from http://www.collegevalues.org/articles.cfm?a=1&id=1023

Singh, K., & Mehmet, O. (2000). Effect of part-time work on high school mathematics and science course taking. *Journal of Educational Research, 94*, 67–74.

Singh, K., Granville, M., & Dika, S. (2002). Mathematics and science achievement: Effects of motivation, interest, and academic engagement. *Journal of Educational Research, 95*(6), 323–332.

Stage, F. K., & Rushin, P. W. (1993). A combined model of student predisposition to college and persistence in college. *Journal of College Student Development, 34*, 276–281.

Steinberg, L., Fegley, S., & Dornbusch, S. M. (1993). Negative impacts of part-time work on adolescent adjustment: Evidence from a longitudinal study. *Developmental Psychology, 29*, 171–180.

Strayhorn, T. L. (2005). More than money matters: An integrated model of graduate student persistence. *Dissertation Abstracts International, 66*(2), 519A. (AAT No. 3164184)

Strayhorn, T. L. (2006a). Factors influencing the academic achievement of first-generation college students. *NASPA Journal, 43*(4), 82–111.

Strayhorn, T. L. (2006b). *Frameworks for assessing learning and development outcomes*. Washington, DC: Council for the Advancement of Standards in Higher Education (CAS).

Strayhorn, T. L. (2006c). Influence of gender, race, and socioeconomic status on college choice: A National Longitudinal Survey of Freshmen (NLSF) investigation. *NASAP Journal, 9*(1), 100–117.

Strayhorn, T. L. (in press-a). Influences on labor market outcomes of African American college graduates: A national study. *Journal of Higher Education*.

Strayhorn, T. L. (in press-b). The invisible man: Factors affecting the retention of low-income African American males. *NASAP Journal*.

Strayhorn, T. L., & Blakewood, A. M. (2007, June). *Studying the wonder year: A university-wide first year assessment*. Paper presented at the National Association of Student Personnel Administrators International Assessment and Retention Conference, St. Louis, MO.

Strayhorn, T. L., & Blakewood, A. M. (2008, February). *Using empirical data to improve first-year seminars*. Paper presented at the 27th Annual Conference on The First-Year Experience & Students in Transition, San Francisco, CA.

Strayhorn, T. L., & Saddler, T. N. (2007, April). *Factors influencing employment satisfaction for African American faculty members*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Strayhorn, T. L., & Terrell, M. C. (2007). Mentoring and satisfaction with college for Black students. *Negro Educational Review, 58*(1–2), 69–83.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*(5), 517–540.

Thomas, S. L., Heck, R. H., & Bauer, K. W. (2005). Weighting and adjusting for design effects in secondary data analyses. In P. D. Umbach (Ed.), *Survey research: Emerging issues* (pp. 51–72). San Francisco: Jossey-Bass.

Williamson, D. R., & Creamer, D. G. (1988). Student attrition in 2- and 4-year colleges: Application of a theoretical model. *Journal of College Student Development, 29*, 210–217.

Wine, J. S., Cominole, M. B., Wheeless, S. C., Dudley, K. M., & Franklin, J. W. (2006). *1993/03 Baccalaureate and beyond longitudinal study (B&B:93/03) methodology report* (NCES Report No. 2006–166). Washington, DC: National Center for Education Statistics.

Wine, J. S., Heuer, R. E., Wheeless, S. C., Francis, T. L., Franklin, J. W., & Dudley, K. M. (2002). *Beginning postsecondary students longitudinal study: 1996–2001 (BPS:96/01) methodology report* (NCES Report No. 2002–171). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Zhang, L. (2005). Advance to graduate education: The effect of college quality and undergraduate majors. *Review of Higher Education, 28*(3), 313–338.