Large Scale Survey Data in Career Development Research

Matthew A. Diemer

Michigan State University


**Author’s note.** Correspondence regarding this paper should be directed to Matthew A. Diemer, Department of Counseling, Educational Psychology and Special Education, 441 Erickson Hall, College of Education, Michigan State University, East Lansing, MI 48824-1034; (517) 432-1524; email: diemerm@msu.edu.

Thank you to Qiu Wang and Adam Voight for their assistance with this manuscript.
Abstract

Large scale survey datasets have been underutilized but offer numerous advantages for career development scholars, as they contain a multitude of career development constructs with large and diverse samples that are followed longitudinally. Constructs such as work salience, vocational expectations, educational expectations, work satisfaction, and occupational attainment are readily available for study. However, with a few notable exceptions, studies of these datasets are infrequent in the career development literature. This paper reviews the strengths and weaknesses of these datasets for career development research, the technical aspects of complex sample design, and software options for analyses of these datasets. Career development scholars must clearly understand complex sample design and analysis strategies to avoid drawing inappropriate conclusions from analyses of large scale survey data. Through illuminating the potential of large scale survey datasets, providing a more user-friendly introduction to the features of complex sample design, and reviewing data analysis options, this paper hopes to increase the utilization of large scale survey datasets by career development scholars.

Keywords: Career Development; Complex Sample Data; Large Scale Surveys; Secondary Analysis
Large Scale Survey Data in Career Development Research

Large scale survey datasets (LSSDs) longitudinally examine phenomena among a broad group of individuals, such as youths’ labor market experiences. In the U.S., LSSDs are usually administered by federal agencies, such as the National Center for Education Statistics (NCES), or are housed at academic institutions. The National Educational Longitudinal Study (NELS) and the more recent Educational Longitudinal Study (ELS) are more well-known LSSDs administered by NCES that follow American high school students and their postsecondary transitions into college and/or work. LSSD participants are commonly surveyed about constructs that may be of interest to career development scholars, such as work salience, vocational expectations, educational expectations, job satisfaction, job-related income, and occupational attainment. Table 1 details variables available in LSSDs of particular interest to career development scholars.

---Insert Table 1 about here---

LSSDs offer a variety of strengths, as they contain large samples of underresearched groups (e.g. poor Youth of Color, rural adolescents, high school dropouts), follow the same participants longitudinally over a span of several years (such as from 8th grade until eight years after high school in the NELS study), and may survey participants’ parents, teachers, and school principals. For example, the NELS study would afford investigations of the roles of participants’ parents, teachers and school principals upon the career development of participants (Ingels, Scott, Taylor, Owings & Quinn, 1998). LSSDs offer advantages to the researcher in that the data have already been collected and compiled; further, the process of human subjects approval for LSSD analyses is very streamlined at many universities. For these and other reasons, a recent
A white paper published by the American Educational Research Association noted that over 10,000 articles, books, and dissertations have been published that analyze LSSDs (Schneider, Carnoy, Kilpatrick, Schmidt & Shavelson, 2006).

However, with a few notable exceptions (e.g. Author Citation A, in press; Author Citation B, 2007; Rojewski & Yang, 1997; Trusty, Ng & Plata, 2000; Trusty, Watts & Erdman, 1997), career development scholars have failed to take advantage of LSSDs. To remedy this underutilization and inform career development scholars of LSSDs’ untapped potential, this paper reviews the strengths and weaknesses of LSSDs, discusses the technical aspects of LSSD collection, and puts forward software options for analyses of LSSDs.

**Strengths of Large Scale Survey Data**

One of the most apparent and appealing aspects of LSSDs is that the data are readily available for analysis. In essence, the “front end” work of planning a study, identifying a setting for data collection, securing human subjects approval, collecting data, and entering the data into SPSS (or another software program) have essentially been completed. Career development scholars may identify LSSDs of interest via print or online directories. Print directories detail many of the available LSSDs in the social sciences (e.g. Young, Savola and Phelps, 1991), but are being supplanted by online resources. The Inter-University Consortium for Political and Social Research ([www.icpsr.umich.edu](http://www.icpsr.umich.edu)) holds a wide range of archival social science data and also provides technical assistance in identifying and analyzing LSSDs. The Murray Research Archive ([www.murray.harvard.edu](http://www.murray.harvard.edu)) is also a large repository of social science data. Although an exhaustive list of LSSD directories is beyond the scope of this paper, readers are directed toward Kiecolt and Nathan (1985) and Zaitzow and Fields (1996) for accessible discussions of strategies for locating, obtaining, and analyzing LSSDs.
Career development scholars can easily obtain public-use versions of many LSSDs (versus restricted-use versions, which require a thorough background check and ongoing data security procedures to be utilized). For example, the public-use version of the NELS dataset can be obtained by ordering a (free) CD-ROM via the NCES website. This CD-ROM can then be used to identify variables of interest to build a dataset that is exported into SPSS or SAS for data cleaning, recoding, and analyses. Using a series of filters, the population of NELS participants can be reduced to a subpopulation, such as non-dropout poor Youth of Color in the 12th grade, and data analyses can proceed using appropriate survey software. Similarly, the NLSY dataset noted in Table 1 is searchable via an online “Web Investigator” (www.nlsinfo.org/web-investigator) that lets users identify variables of interest, create their own dataset online, and extract this custom dataset for analysis. Further, the availability of LSSDs may facilitate the replication of findings across different sources of data, an underutilized practice that would bolster confidence in findings from previous career development scholarship (Zaitzow & Fields, 1996).

LSSDs offer additional advantages in that large samples of underresearched groups (e.g. rural adolescents, persons with disabilities) are also available. Because of the small population size of some subgroups (e.g. Asian American youth who reside in poverty and have dropped out of school) or potential difficulties in accessing other subgroups (such as rural adolescents), the possibility of examining aspects of career development with large samples of subgroups is a clear strength of LSSDs. Further, because LSSDs are collected via complex sample designs, they are nationally representative. Briefly, this entails that analyses of LSSDs generalize to the entire population of U.S. students, and, consequently, limitations germane to geographic region do not apply to analyses of LSSDs. However, complex sample designs also violate the assumptions of
traditional sampling, which must be accounted for in data analyses and will be discussed at length later in this paper.

LSSDs generally utilize a panel longitudinal design, where the same set of participants is followed over a span of several years (Ruspini, 2002). (However, due to participant attrition or nonresponse, new participants are often “freshened” into later waves of panel designs). Panel longitudinal designs add a degree of rigor to analyses, as the same career development constructs can be examined with the same set of participants over time. Being able to access a data set that follows the same set of participants over a decade is an enormous advantage that cannot be understated, as conducting a longitudinal study across a ten year period with a large sample size would be a very time-consuming and costly endeavor. Further, although longitudinal designs do not entail that causal inferences can be drawn, they do permit greater confidence in inferences drawn from analyses and in the direction of the relationships between constructs of interest (Schneider et al., 2006). Finally, because longitudinal designs are underutilized in career development scholarship (with some exceptions, such as Super’s Career Patterns Study), LSSDs could be used to examine career development constructs longitudinally.

Key aspects of participants’ social context are often surveyed in LSSDs, such as participants’ teachers, school principals, and parents. This permits examinations of key figures in youths’ lives and their impact upon career development. For example, the NELS dataset facilitates examinations of the availability of school-based resources for career development and parental support for career development upon the career development of participants (e.g. Author Citation B, 2007).

Finally, some funding opportunities are designated for analyses of LSSDs. For example, the American Educational Research Association’s “Research Grants” program supports studies
of LSSDs for dissertation work and faculty research projects. The Association for Institutional Research maintains a grant program reserved for studies of postsecondary transitions and experiences with LSSDs. The Institute for Education Sciences also maintains grant programs reserved for analyses of LSSDs. Many other sources of public and private funding, although not designated for analyses of LSSDs, have a history of supporting LSSD-based projects.

**Weaknesses of Large Scale Survey Data**

A particular set of research questions may not always match a particular LSSD, which may entail modifications to optimize their fit. In some cases, one’s research questions cannot be examined in a dataset and a more appropriate dataset identified. In other cases, one’s research questions may be revised to capitalize upon the advantages of a particular dataset (e.g. a large sample of understudied individuals). Indeed, scholars’ creativity guides the fluid process of matching theoretical models and LSSDs (Zaitzow & Fields, 1996). For example, career development scholars may apply new theoretical lenses to LSSDs primarily collected for other purposes, reconceptualizing extant LSSD variables to advance career development scholarship.

A second weakness is the manner in which career development constructs are operationalized in LSSDs. LSSDs generally contain fewer variables to operationalize a construct than a more traditional scale would contain; also, the wording of questions may differ from the researcher’s perspective of how items should be written. Often, several items can be combined to best represent a construct (Kiecolt & Nathan, 1985). For example, “work salience,” a construct defined by Greenhaus (1971) as the relative importance of work and career in an individual’s life, was operationalized in the NELS and ELS surveys by three questions assessing the (1) importance of success in work, (2) the importance of money in work, and (3) the importance of being able to find steady work (Ingels et al., 1998). This operationalization of work salience does
not measure all dimensions of Greenhaus’s (1971) and Super’s (1980) conceptualization of this construct and contains fewer items than the 27-item Work Role Salience Scale (Greenhaus, 1971). This limitation should be considered by career development scholars. However, it is largely offset by the capacity to follow work salience and other career development constructs longitudinally among a diverse set of participants and by the careful matching of research questions to a particular dataset, as discussed above.

LSSDs also require an investment of time in understanding the nature of complex sample design, managing large scale datasets, and the complexities of LSSD data analysis. (The primer on technical aspects of LSSDs in this paper provides an introduction to stratification, clustering, and weighting, as well as survey software). Variables from LSSDs also tend to violate assumptions of normality required for many forms of data analyses. Although these issues can often be addressed by a series of data transformations (e.g. logarithmic or Box-Cox transformations; Johnson & Wichern, 1998), career development scholars unfamiliar with these procedures must also invest time into understanding the perils of variable nonnormality and identifying the correct transformations to attenuate variable nonnormality. Finally, effectively managing datasets of this size (e.g. the NLSY dataset contains over 9,000 participants and hundreds of variables) with statistical software can be challenging.

Relatedly, missing data is a particular problem in longitudinal designs, as the opportunity for subject attrition is great over a study that spans over a decade (Ruspini, 2002). There are also many “opportunities for missingness” in a large scale survey that queries participants, their teachers, their parents, and their school principals, for example. However, significant advances in addressing missing data have occurred in recent years. The expectation-maximization likelihood algorithm imputes “expected” values for missing data based on the values for non-missing data
Large Scale Survey Data in Career Research

implemented in SPSS version 14.0; Allison, 2003). In multiple imputation methods, more than one imputation of missing data is conducted, and the average values of these imputations replace missing data (available in the NORM freeware program and other packages; Schafer, 1999). The Full Information Maximum Likelihood method makes use of all existing data in analyses (rather than deleting cases pairwise or listwise) with minimal statistical bias (available in MPlus and other software packages; Muthén & Muthén, 2006). The reader is directed to Allison (2003) and Schafer and Graham (2002) for particularly accessible discussions of missing data strategies.

Although a variety of issues related to analyses of LSSDs are reviewed here, several resources may be of assistance to scholars interested in conducting research using LSSDs. First, Lee and Forthofer (2006), Thomas and Heck (2001) and the present paper provide more user-friendly introductions to analyses of LSSDs. Lohr (1999) and Kish (1965) provide more of the methodological and theoretical basis behind complex sampling strategies. Secondly, many universities offer low-cost workshops that provide training in analyses of LSSDs. The National Center for Education Statistics also conducts a summer workshop series for scholars and students interested in learning to analyze LSSDs. These summer workshops are provided free of charge; however, the selection process is competitive. Finally, although understanding and analysis of LSSDs does require an investment of time, this investment may result in the acquisition of methodological skills (e.g. addressing variable non-normality and missing data) that can also be applied to analyses of simple random sampling data.

Technical Aspects of Complex Sample Design

LSSDs are created via complex sampling designs, which entail the collection of nationally representative data from a smaller sample of participants (Kish, 1965). Because of the financial and logistical problems in surveying all current high school students in the U.S., for
example, large-scale surveys carefully select a sample that represents some population of interest, such as current high school students in the U.S. The characteristics of this sample are proportionate to some external benchmark, such as recent census data, so that the smaller sample of the LSSD reflects the characteristics of this larger population of interest (Lohr, 1999).

In a common complex sample design, stratification, clustering, and weighting are used to efficiently collect nationally representative data. (It would be even more costly and time-consuming to collect data from all current U.S. high school students, for example). Generally, the country is divided into geographic regions (stratification), schools within each stratum are identified (clustering), and some individuals within each school are selected into the sample, with an unequal probability of being sampled (weighting). In this design, individuals are nested within clusters, and clusters nested within strata. This non-independence of observations violates the assumptions of the more traditional simple random sampling design (Thomas & Heck, 2001).

These features of complex sample design must be accounted for to avoid inappropriate conclusions being drawn from data analyses (Brogan, 1998; Muthén & Satorra, 1995). Understanding how LSSDs are constructed will assist career development scholars in correctly analyzing complex sample data. Therefore, these complex sample design features are discussed in further detail below and are depicted in Table 2.

---Insert Table 2 about here---

*Stratification*

As a first step in many complex sample designs, the country is divided into geographic regions, called strata. Stratification breaks a larger group, such as the entire population of U.S. high school students, into more manageable groups of strata (such as all students in a geographic
Statistically, stratification has the effect of decreasing standard errors and yielding more precise parameter estimates (Chambers & Skinner, 2003).

**Clustering**

Clusters, or primary sampling units, are used as targets for data collection within strata. Rather than considering all the individuals within a geographic region for data collection, schools (or some other type of cluster/primary sampling unit) are used to identify sites for data collection within a strata, which is a more efficient method of collecting data than collecting data from all individuals within a strata (Lohr, 1999). Accordingly, clusters are nested within strata (e.g., schools are nested within particular geographic regions).

In a complex sample design, the characteristics of clusters within each strata reflect the “real” characteristics of schools within a particular geographic region. For example, if the “real” proportion of Catholic schools in a geographic region is 6%, and 100 schools will serve as primary sampling units within this strata, six Catholic schools would be selected into the sample as primary sampling units. The proportion of Catholic schools in the survey sample would equal the proportion of Catholic schools in the “real” population. Similarly, the proportion of public schools serving as primary sampling units in each strata would also be proportionate to the “real” proportion of public schools in each geographic region (Kish, 1965).

This nesting of clusters within strata, along with the sampling of individuals from the sample school, violates the simple random sampling assumptions of homogeneity and independence of observations (Thomas & Heck, 2001). Individuals from the same cluster, or school, are more likely to share similarities than individuals from different schools. For example, we would expect two students selected from the same private suburban school (cluster) to share more characteristics than a student selected from a private suburban school and a student from a...
public urban school. This consequence of clustering reduces the precision of parameter estimates and increases standard errors (Lohr, 1999). The nonindependence of observations, along with the increase in standard errors germane to clustering and the decrease in standard errors germane to stratification, must be addressed in analyses of LSSDs (Brogan, 1998).

**Weighting**

In a (more common) simple random sampling design, all individuals have an equal probability of being selected into the sample. By contrast, in a complex sample design, individuals have an unequal probability of being selected into the sample. This is due to the need to select individuals based on their characteristics. For example, African American/Black students from private schools and African American/Black students from public schools will be selected at different rates for inclusion into the survey sample. The selection probability will correspond to the need to select individuals in proportion to their representativeness of their cluster (school) and strata (geographic region), as well as the need to construct an overall survey sample that represents the population of interest (e.g. U.S. high school students) (Lohr, 1999).

*Weights* are constructed to reflect the probability of an individual being selected into the sample. Generally, these initial weights represent the inverse of the probability of an individual being selected into the sample (Pfefferman, 1993).

Next, weights are often adjusted for the effects of participant non-response (Stapleton, 2002). In a LSSD, particularly in a survey that follows participants over a span of several years, participant attrition becomes a more salient issue, as individuals may die, change addresses, or not respond to a survey for other reasons. This likelihood of response for individuals is often incorporated into survey weights by adjusting the initial set of weights by the inverse of the response rate for an individual’s subgroup (Lee & Forthofer, 2006). (Response rates for
subgroups can be calculated in various ways, such as by geographic region, racial/ethnic group membership, or gender). This weighting for non-response addresses the differential response rates to a survey by different subgroups and creates more accurate parameter estimates (Pfefferman, 1993; Stapleton, 2002).

The third way that weights are often computed involves the process of post-stratification raking, which involves adjusting weights (after data have been collected) so that the weighted survey sample is representative of some external population of interest (Lee & Forthofer, 2006). For example, in the NELS survey, post-stratification weighting was employed so that the weighted survey corresponds to the U.S. Census Data in terms of racial/ethnic membership, socioeconomic status, gender, and other factors (Ingels et al., 1998).

In sum, many types of weights can be constructed; however, weights are generally computed to address (1) unequal probability of selection, (2) non-response, and (3) the need for the survey sample to match some external benchmark. There are also less common forms of weighting not addressed here (the reader is encouraged to consult Kish, 1965 or Lohr, 1999 for further information). Regardless of how the weights have been constructed, the weighting used in LSSDs must be attended to in data analyses to avoid making inappropriate conclusions from data analyses (Pfefferman, 1993; Thomas & Heck, 2001).

Analyses of Large Scale Survey Data

The stratification, clustering, and weighting germane to complex sample designs violate the assumptions of simple random sampling. If treated as simple random sample data in data analyses, complex sample data will bias standard error, variance, and sample size estimates and therefore must be addressed in data analyses (Brogan, 1998). Analysts of LSSDs must consult the technical documentation that accompanies LSSDs to identify the appropriate stratum,
cluster/primary sampling unit, and weight variable for analyses. The stratum and cluster variable can then be incorporated into data analyses by specialized software packages (discussed below) to obtain accurate variance estimates and standard errors in analyses of LSSDs. Similarly, the appropriate weight variable can then be incorporated into data analyses by specialized software to obtain accurate estimates of sample size and address unequal selection probabilities, non-independence of observations, and non-response in analyses of LSSDs. (Readers interested in a more technical discussion of these issues are directed toward Chambers & Skinner, 2003, Pfefferman, 1993, and Stapleton, 2002, 2006). Combined with careful scrutiny of technical documentation, the use of selected software packages ensures that LSSDs are correctly analyzed and are discussed in further detail next. In particular, Thomas and Heck (2001) and Lee and Forthofer (2006) provide user-friendly discussions of the multiple analytic methods that can be brought to bear upon LSSDs.

*Software for Analyses of Large Scale Survey Data*

Career development scholars interested in analyses of LSSDs may either (1) use traditional software packages and make several technical adjustments or (2) make use of software specifically designed for LSSDs. Software packages designed for analyses of simple random survey data, such as the SPSS base package, can be used if several steps are followed. One could scrutinize the dataset’s supporting documentation and the software’s user guide to manually correct for the impact of complex sample design features upon analysis variables and obtain unbiased parameter estimates, standard errors, and significance tests. This would entail the calculation of normalized weights, intraclass correlation coefficients, and design effect (DEFF) adjustments, more technical issues not reviewed here but accessibly covered by Thomas and Heck (2001) and Lee and Forthofer (2006) for the interested reader.
On the other hand, career development scholars could utilize one of the many software packages designed for LSSDs. In recent years, the availability and user-friendliness of these software packages has increased exponentially. These packages allow the researcher to set the appropriate weight, cluster, and stratification variables in data analysis. These software packages then calculate the appropriate weights (if not already included by the LSSD) and make use of advanced variance estimation methods (e.g. Taylor series linearization) to accurately analyze large scale survey data (Siller & Tompkins, 2006). Although users of these more specialized software packages must also study supporting documentation and technical manuals, these specialized software packages obviate the need for the user to make the technical adjustments required for analyses of LSSDs. If used properly, these specialized software packages take on the onerous computations of normalized weights, variance estimation, and standard error adjustments, eliminating several potential sources of “user error” in analyses of LSSDs.

Generally, specialized software packages require that the user identify the appropriate weight, cluster, and stratification variables from the LSSD, and “set” these variables via command syntax or pull-down windows. SPSS has recently introduced a complex sample add-on module (available for additional cost) that permits the user to make use of the general analysis features of SPSS (descriptive and regression analyses) with LSSDs. SAS version 8.0 implemented a complex survey feature, which permits descriptive, frequency, and regression (linear and logistic) analyses of LSSDs. The American Institutes for Research offers a freeware program, called AM, designed for analyses of LSSDs that is available for download from http://am.air.org/default.asp. The AM program has more limited data cleaning and data analysis options, but does allow the user to import SPSS and/or SAS datasets for descriptive, frequency, and regression analyses.
SUDAAN and Stata were specifically designed for analyses of LSSDs; both packages use advanced variance estimation methods and incorporate sample weights to obtain unbiased estimates. SUDAAN can be used as a stand-alone program, or “called” for use within SAS, as SUDAAN uses syntax compatible with SAS. This feature enhances SUDAAN’s stand-alone analysis options (descriptive and regression analyses). Stata has more options for dataset management than other specialized software, commands which allow users to create new variables and manage datasets effectively. (Other specialized software packages generally require cleaning and recoding data in SPSS or SAS, and then importing the data for analysis into the specialized software package.) Stata allows users to conduct descriptive and linear regression analyses with complex sample data, as well as Poisson and probit regression. Stata can also analyze LSSD using Heckman selection models and propensity score matching to approximate causal inferences from observational data (Schneider et al., 2006).

The selection of a software package for descriptive and regression analyses may be guided by familiarity, cost, and sophistication, as there appear to be minimal differences in the precision of SPSS, SAS, SUDAAN, and STATA for basic analyses of LSSDs (Siller & Tompkins, 2006). Career development scholars familiar with SPSS and SAS may find the recent complex sample modules added to these programs appealing; with AM serving as an underutilized and no-cost alternative. SPSS, SAS, or AM would be a good choice for the career development scholar planning more basic or a more circumscribed number of LSSD analyses. On the other hand, SUDAAN and STATA have a broad range of analytic capabilities that career development scholars planning more sophisticated or a series of LSSD analyses should utilize.

The MPlus program appears to be the best choice for structural equation modeling of LSSDs, according to recent simulation studies (Stapleton, 2006). According to its technical
Large Scale Survey Data in Career Research 17
documentation (Muthén & Muthén, 2006), MPlus is also well-equipped to address the analytic complexities that LSSDs pose, such as variables that violate assumptions of normality, variables that are dichotomous or categorical, and variables with a higher proportion of missingness (surveys that follow the same participants across several waves over a decade have several opportunities for missing data). MPlus has incorporated several recent technical developments, such as estimators (e.g. MLR) robust against violations of nonnormality, estimators (e.g. WLS, WLSMV) developed for analyses of categorical/binary variables, and the Full Information Maximum Likelihood method (which makes use of all existing data in analyses), that address these analytic complexities (Flora & Curran, 2004; Muthén & Muthén, 2006).

HLM (Hierarchical Linear Modeling, Raudenbush, Bryk, Cheong, Congdon & Du Toit, 2004) was developed to analyze hierarchically structured data collected through a multistage complex sampling design. “Hierarchical structure” refers to the nesting of students within schools and schools within geographic regions, for example. Hierarchical linear modeling involves the linear modeling of growth trajectories at the individual level, as well as the effects of a second level, such as the school (cluster), on these growth trajectories. For example, the mean value for a career development construct of interest for each school could be examined over time, as well as the mean value for a career development construct of interest for individuals within those schools. In this way, both the effects of schools and the trajectories of individual’s growth can be delineated via HLM. This is an underutilized analysis method in career development scholarship, with promise to illuminate growth trajectories and the delineation of school effects in the processes of career development. However, HLM does require that datasets be built and cleaned in a software package such SPSS or SAS before being imported to HLM for analyses.
Summary & Conclusions

In summary, LSSDs have been underutilized but offer numerous advantages for career development scholars, as they contain a multitude of career development constructs with large and diverse samples that are followed longitudinally. The public-use versions of LSSDs are readily available, and career development constructs contained in LSSDs are identified in this paper. However, the strengths of LSSDs are offset by limitations in how career development constructs are operationalized, the need to carefully match research questions to a dataset, and in the analytic complexities that LSSDs pose.

These analytic complexities must be addressed by data analysts to avoid inappropriately analyzing complex sample data under the more traditional assumptions of simple random sampling. Further, understanding the nature of complex sample designs and the analytic complexities of LSSDs requires an investment of time and energy. However, this investment is borne out by the many strengths that LSSDs offer to career development scholars. Accordingly, a review of the features of complex sample design and these analytic complexities was provided in this paper. Through highlighting the many strengths of LSSDs, introducing readers to complex sample design, and detailing software packages designed for analysis of LSSDs, this paper hopes to remediate the underutilization of LSSDs by career development scholars.

Author Citation A. (in press). Sociopolitical development and vocational expectations among lower-SES Adolescents of Color. *Career Development Quarterly.*


## Table 1: Selected Large Scale Survey Datasets & Career Development Variables

<table>
<thead>
<tr>
<th>Dataset Acronym &amp; Full Name</th>
<th>Brief Dataset Description</th>
<th>Selected Career Development Variables</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>NELS: National Education Longitudinal Study</td>
<td>Cohort of eighth graders surveyed in 1988 and again in 1990, 1992, 1994 and 2000. Participants’ parents, teachers and school principals also surveyed.</td>
<td>Work salience; Vocational and educational expectations; Work satisfaction (job conditions, security, opportunities for advancement, income); Workplace training; Adult occupational attainment; Academic experiences (school-based learning programs, high school graduation status, postsecondary experiences).</td>
<td><a href="http://nces.ed.gov/surveys/nels88">http://nces.ed.gov/surveys/nels88</a></td>
</tr>
<tr>
<td>ELS: Educational Longitudinal Study</td>
<td>Cohort of tenth graders surveyed in 2002 and again in 2004, 2006 (future waves planned). Participants’ parents, teachers and school principals also surveyed.</td>
<td>Work salience; Vocational and educational expectations; Work satisfaction (job conditions, security, opportunities for advancement, income); Workplace training; Academic experiences (school-based learning programs, high school graduation status, postsecondary experiences).</td>
<td><a href="http://nces.ed.gov/surveys/els2000">http://nces.ed.gov/surveys/els2000</a></td>
</tr>
<tr>
<td>NLSY: National Longitudinal Survey of Youth</td>
<td>Youth (ages 12 to 16) and one parent surveyed in 1997; follow-up waves annually. Participants’ schools surveyed in 2000.</td>
<td>Work experiences (e.g. hours per week, pay, industry type); Academic experiences (achievement test scores, school-based work programs, postsecondary experiences); Educational and vocational expectations; Job satisfaction; Adult occupational attainment.</td>
<td><a href="http://www.bls.gov/nls">http://www.bls.gov/nls</a></td>
</tr>
<tr>
<td>HS&amp;B: High School &amp; Beyond</td>
<td>Cohorts of high school sophomores and seniors followed every other year from 1980 to 1986 (sophomore cohort surveyed again in 1992). Surveys of graduating college seniors</td>
<td>Work salience; Work satisfaction; Work experiences (workplace training offered, racial/ethnic diversity in current occupation); Job income (data for each of nine years after highest degree attained); Educational expectations; High school and postsecondary experiences.</td>
<td><a href="http://nces.ed.gov/surveys/hsb">http://nces.ed.gov/surveys/hsb</a></td>
</tr>
</tbody>
</table>

**Note:** The URL links are placeholders and may not lead to actual websites.
<table>
<thead>
<tr>
<th>Study</th>
<th>Beginnings</th>
<th>Follow-up Waves</th>
<th>Data Collection Areas</th>
</tr>
</thead>
</table>
Table 2: Summary of Complex Sample Design Features, Effects, and Analysis Options

<table>
<thead>
<tr>
<th>CSD Feature</th>
<th>CSD Strategy</th>
<th>CSD Effect</th>
<th>Analysis Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dividing country into regions</td>
<td>Stratification</td>
<td>Reduces SEs and improves precision of parameter estimates</td>
<td>Setting appropriate stratification variable in analysis</td>
</tr>
<tr>
<td>Non-independence of observations</td>
<td>Clustering</td>
<td>Increases SEs and decreases precision of parameter estimates</td>
<td>Setting appropriate cluster variable in analysis</td>
</tr>
<tr>
<td>Unequal selection probability</td>
<td>Weighting</td>
<td>Corrects for unequal selection probability</td>
<td>Setting appropriate weight variable in analysis</td>
</tr>
<tr>
<td>Participant non-response</td>
<td>Weighting</td>
<td>Weights adjusted so that impact of non-response upon parameter estimates corrected</td>
<td>Setting appropriate weight variable in analysis</td>
</tr>
<tr>
<td>Matching sample to population of interest</td>
<td>Weighting</td>
<td>Weights adjusted so that survey sample corresponds to external benchmark</td>
<td>Setting appropriate weight variable in analysis</td>
</tr>
</tbody>
</table>

Note. Software that incorporates all features of complex sample design - weighting, clustering and stratification – will yield the most accurate parameter estimates and reduce risks of Type I and Type II error.