**The NLSY Kinship Links: Using the NLSY79 and NLSY-Children Data to Conduct Genetically-Informed and Family-Oriented Research**

Joseph Lee Rodgers1 , William H. Beasley2, David E. Bard2, Kelly M. Meredith3, Michael Hunter2, Amber B. Johnson4, Maury Buster5, Chengchang Li6, Kim O. May7, S. Mason Garrison1, Warren B.Miller8, Edwin van den Oord9, & David C. Rowe10

Authors’ affiliations:

1 Department of Psychology and Human Development, Vanderbilt University, Nashville TN

2 University of Oklahoma Health Sciences Center, Oklahoma City, OK

3 Office of Institutional Research, Oklahoma City University, Oklahoma City, OK

4 Portland State University, Portland, OR

5 Alabama State Personnel Department, Pratville AL

6 HSBC, Schaumburg IL

7 Department of Psychology, College of Charleston, Charleston SC

8 Transnational Family Research Institute, Aptos CA

9 School of Pharmacy, Virginia Commonwealth University, Richmond VA

10 University of Arizona (deceased)

Contact Author: Joe Rodgers, 230 Appleton Pl, 552, Hobbs 202C, Vanderbilt University, Nashville TN 37203, 405-816-2941, [jrodgers@ou.edu](mailto:jrodgers@ou.edu); [joseph.l.rodgers@vanderbilt.edu](mailto:joseph.l.rodgers@vanderbilt.edu)

**The NLSY Kinship Links: Using the NLSY79 and NLSY-Children Data to Conduct Genetically-Informed and Family-Oriented Research**

**Abstract**

The National Longitudinal Survey of Youth datasets (NLSY79; NLSY-Children/Young Adults; NLSY97) have extensive family pedigree information contained within them. These data sources are based on probability sampling, a longitudinal design, and a cross-generational and within-family data structure, with hundreds of phenotypes relevant to behavior genetic (BG) researchers, as well as other developmental and family researchers. These datasets provide a unique and powerful source of information for BG researchers. But much of the information required for biometrical modeling has been hidden, and has required substantial programming effort to uncover -- until recently. Our research team has spent over 20 years developing kinship links to genetically inform biometrical modeling. In the most recent release of kinship links from two of the NLSY datasets, the direct kinship indicators included in the 2006 surveys allowed successful and unambiguous linking of over 94% of the potential pairs. In this paper, we provide details for research teams interested in using the NLSY data portfolio to conduct BG (and other family-oriented) research.

Keywords: NLSY79, NLSYC, NLSY97, behavior genetics, biometrical modeling, siblings, kinship

**The NLSY Kinship Links: Using the NLSY79 and NLSY-Children Data to Conduct Genetically-Informed and Family-Oriented Research**

**Introduction**

We describe a set of high-quality kinship links available for the National Longitudinal Survey of Youth 1979 (NLSY79) and NLSY-Children/Young Adult (NLSY-C/YA) datasets that support high-quality biometrical research using these broad and powerful longitudinal databases. Our online kinship link archive, located at [http://liveoak.github.io/NlsyLinks/](http://liveoak.github.io/NlsyLinks/A) contains a number of kinship link files, the original source code, and vignettes illustrating data management principles and procedures.

Chase-Landsdale et al. (1994) published an article promoting the value of the NLSY-Children dataset for developmental and other behavioral science researchers. *Harris, Halpern, Smoken, & Haberstick (2006) published a similar article describing the value of the Add Health dataset for researchers using twin samples.* The current article has a similar goal to these two articles. It is broader in the sense that we have developed research databases using two of the NLSY datasets (with a third in development), and links between them, all of which are available (either upon request or online) to academic researchers. But this article is also quite focused because it is specifically targeted to researchers in behavior genetics (with application as well to those working within related methodological and family-studies arenas).

Few population-based family data sources exist that are biometrically informed. Even rarer are biometrically-informed data with many decades of longitudinal structure, including cross-generational links. The set of three databases comprising the National Longitudinal Survey of Youth (NLSY) provides a valuable data source for behavior genetic -- and other demographic, aging, developmental, and family-based -- researchers. The immediate purpose of this article is to *document the history and development of the kinship links*, and to illustrate the value of the NLSY datasets for behavior genetic (BG) research. The ultimate goal of the article is to stimulate BG and other research teams to incorporate within BG *and family-based* research studies the kinship information that is now easily available and accessible to NLSY researchers.

**Background**

The NLSY Samples

The NLSY datasets are among the most popular and valuable of all social/behavioral science data sources. A part of the National Longitudinal Survey (NLS) portfolio, the NLSY contains three different data sources, two of which are linked cross-generationally (the NLSY79 and the NLSY-Children/Young Adult), and two of which are linked *through intentional replication, including approximately matching and comparable sampling designs* (the NLSY79 and NLSY97).

The NLSY79 originated as a household probability sample of 12,686 adolescents in the age range 14-21 on December 31, 1978, thus born between 1957 and 1964. The age range applies to the day the sample was drawn; interviews occurred several months later, so that there are a few 22-year-old respondents at first interview. The NLSY79 contained 6,111 respondents from the original household probability sample, 5,295 respondents who were part of the minority and poor white oversample, and 1,280 respondents who were part of the military sample. The 12,686 respondents represented 8,770 unique households, with 2,862 households that contained two or more NLSY79 respondents; of these, 2,448 were multiple-sibling households, with 414 households containing no siblings, but instead two or more young spouses, cousins, and other categories of relatedness. A total of 5,263 NLSY79 respondents had siblings who were also NLSY79 respondents. Importantly, until 2006 there was no effort to separately identify full, half, step, and adoptive siblings; all of these categories were identified in the household roster as "sibling." Therefore, the NLSY79 designation of “sibling” did not discern the various levels of genetic relatedness among siblings. However, several variables were informative of this distinction, and previous NLSY research using kinship links has been based on inferred level of kinship using implicit indicators of relatedness.

The kinship links identify sibling and other kinship relationships among the multi-respondent households. These links are approximately representative of those that occurred among 14 to 21 year olds in households in the U.S. on the last day of 1978, and include twins, different types of siblings, cousins, uncle-nephews, step-siblings, and a number of other more unusual types of kinship pairs. Respondents in the NLSY79 were interviewed yearly from 1979 to 1986, at which point interviews were conducted on a biennial basis until 2014 and continuing. The 1,643 respondents in the poor white oversample were dropped for budgetary reasons following the 1990 survey, and 1,079 respondents from the military sample were dropped following the 1986 survey (all but 201 randomly sampled respondents *(who were maintained in the sample*).

The NLSY-Children (NLSYC; we often refer to this dataset as “Generation 2,” or “Gen2”) consists of all biological children born to the 6,283 females in the NLSY79 (which we often refer to as *“Generation 1” or* “Gen1”). Interviews with the NLSYC Gen2 children began in 1986, and have occurred biennially until 2012 and continuing. By the 2012 interviews the NLSY79 Gen1 females were 47 to 55, and so childbearing can be considered complete for this cohort of women. By 2010, 11,504 children had been *identified as* born to the NLSY79 females, 86% of whom were interviewed in 2010; the NLSYC children ranged in age at the 2010 survey from 1 to 39. Obviously, because of the design of the survey, there are many full- and half-siblings within the NLSYC sample. As well, by linking mothers to their sisters, half-sisters, and cousins, we can also identify cousins, half cousins, and several other cousin categories of relatedness within the NLSY-Children data. In fact, the cousin data within Gen2 has seldom been used as an exclusive data source in academic research (see van den Oord and Rowe, 1999, and Goodnight et al., 2013, for exceptions), but has been used fairly often in combination with other kinship categories (including, in a few cases, half-cousins as well).

The cousin data within the NLSY is an underused and underappreciated source of information. R. A. Fisher (1918) noted the interesting status of cousin data in biometrical studies: “… the hypothesis of cumulative Mendelian factors seems to fit the facts very accurately. The only marked discrepancy from existing published work lies in the correlation for first cousins. … but until we have a record of complete cousinships measured accurately and without selection, it will not be possible to obtain satisfactory numerical evidence on this question” (p. 168). Within the current context, we note that our research team has observed several times that the cousin correlations obtained from the NLSY kinship links appear unusually high (or sometimes we have felt that the half-sibling correlations were surprisingly low, thus leading to the appearance of high cousin correlations). The NLSY data provide two different generations of cousin data, and the Gen2 cousins may be as close as can ever be obtained to a large cousin sample, emerging from probability sampling mechanisms, providing, in Fisher’s words, “a record of complete cousinships measured accurately and without selection.” Future NLSY researchers should pay particular attention to the NLSY cousins as the basis for a potentially fruitful research endeavor.

In 1994 and following, a separate survey has been administered to NLSYC respondents age 15 or older in a given survey year; this survey is referred to as the NLSY-Young Adult (NLSY-YA) survey. It contains different types of questions than the childhood surveys administered to the NLSYC respondents, including questions related to sexual development, dating, marriage, reproduction, and fertility; jobs and employment; delinquency, drinking, and smoking; and other adolescent and young adult behaviors. Though there are different names for the NLSYC and NLSY-YA surveys, it is important to emphasize that these are the same Gen2 respondents at different ages. In this paper, when the distinction between the two surveys does not matter, we refer to these respondents as NLSY-C/YA respondents.

The NLSY97 dataset was developed as an approximate replication of the NLSY79, for a cohort born almost two decades later. It consisted of a household probability sample that surveyed all adolescents between the ages of 12 and 17 on December 31, 1996 (therefore 12-18 at time of first interview, and born in 1979-1984 with an age range in 2012 of 27-33). The original sample size was 8,984; almost 83% were still being interviewed in 2012, which was round 15. Surveys have been annual, but occur biennially from 2012 on. The original sample included 6,748 respondents from a cross-sectional household probability sample, and another 2,236 in a minority oversample. Survey questions in NLSY97 are similar to those in the NLSY79, and in fact in many cases the original instrumentation is maintained.

There are a number of features of these three datasets that provide cross-dataset information. Of particular note is that the link between the NLSY79 Gen1 mothers and the NLSY-C/YA Gen2 children creates a two-generational data source. (*In addition,* some information is available about the parents of the NLSY79 respondents, Gen0, including parents’ educational outcomes, birthdates, and place of birth; thus, limited three-generational information is available.) Further, the timing of the sampling of the NLSY97 causes the NLSY97 mean age to be fairly close to that of the NLSY-C/YA respondents (though there is obviously much more age variability in the NLSY-C/YA data). Figure 1 shows a schematic of the timing and pace of births, fertility, and NLSY data collection for these three surveys.

The Kinship Links

**History.** In the early 1990's three independent research teams recognized the potential value of the NLSY-Children data for behavior genetic research: Joe Rodgers and David Rowe1 in the U.S.; Nazli Baydar and colleagues in Turkey and the U.S.; and Edwin van den Oord in The Netherlands. *Rodgers, Rowe, and Li (1994) documented* almost a thousand kinship pairs, *which naturally increased* as NLSY79 females (age 27-35 in the 1992 survey) continued childbearing. However, during the first 20 years of the NLSYC survey, and the first 27 years of the NLSY79 survey, no questions were included to explicitly distinguish between sibling categories.

Nevertheless, several questions in the NLSYC data were informative as to the level of kinship relatedness, in particular questions asked of a NLSYC respondent's biological mother (NLSY79 Gen1 female respondents) about the biological father. The three teams independently developed an algorithm that drew on several critical survey questions to distinguish between full and half siblings. Baydar (2001) produced research using her kinship links. Van den Oord produced a kinship linking algorithm, but began collaborating with Rowe and Rodgers, and ultimately used their algorithm and kinship links for most of his NLSY research. 2

The first set of research articles based on the kinship links were published in 1994 in two different articles using the NLSYC dataset. Rodgers, Rowe, and Li (1994) reported biometrical results using data on childhood problem behaviors for 822 Gen2 kinship pairs, and Rodgers, Rowe, and May (1994) reported biometrical results of a study of PIAT-Math, PIAT-Reading Recognition, and PIAT-Reading Comprehension using 855 Gen2 kinship pairs. Since then, over 30 additional research articles have been published using these NLSY-C/YA kinship links.

The NLSYC linking algorithm was successful enough *to justify development of* similar linking algorithms for the original NLSY79 dataset. *The NLSY79 provided* a much more difficult linking challenge, because there were many different types of relatedness in the original NLSY79 households; successful linking required three different programming efforts between 1992 and 1996 to produce useful NLSY79 kinship links. The first two linking efforts were tested in validation exercises similar to the original NLSYC validation analysis (in which biometrical models were fit to phenotypes -- height and weight -- with known heritability structure), and rejected as a useful set of kinship links. The algorithm was re-written using overlapping and different survey questions, and the third set of kinship links satisfied our validation analyses. Of the targeted 3,890 kinship pairs, 2,470 (63%) were successfully classified. The first NLSY79 kinship link publication was Rodgers, Rowe, and Buster (1999), reporting heritabilities and shared environmental variance associated with age at first intercourse. Since, around 20 research papers have been published using the NLSY79 kinship links.

Despite demonstrated concurrent validity with other research results, validation support, and a growing publication stream, a subset of behavior genetic researchers criticized the early kinship links (within reviewer comments, conference discussions, and direct e-mails). Some criticism was entirely justified, some arguably not. Critics tended to focus on the relatively low linking rates (63% of eligible links were specified in the NLSY79 data) and on lack of direct ascertainment to distinguish full and half siblings. One set of critics often used highly selected twin samples; we noted that for external validity, 63% of a probability sample has certain slight advantages over volunteer twin samples. Other critics included researchers using survey-based zygosity determination, a method often of lower reliability/validity than the NLSY algorithms, which used maternal information about the biological fathers of their children. A third consistent criticism involved a category of "ambiguous siblings," NLSY79 and NLSYC kinship pairs who were definitely full or half siblings but who could not be further distinguished. In early kinship pair research, the ambiguous sibling categories had phenotypic kinship correlations midway between half- and full-sibling correlations. Given hundreds of ambiguous pairs in each dataset, the category was empirically useful, especially for subgroup analyses by race and gender.

We agreed with many early criticisms; some of these features indeed were threats to internal and external validity. Just as obviously, we rejected the view that threats to validity automatically disqualify data as scientifically valid. We prefer to use the tenets of quasi-experimental design methodology (e.g., Shadish, Cook, & Campbell, 2002) to carefully address threats to both internal and external validity. Further, we have consistently argued the irony of some criticisms, given that behavior genetics has built its foundation and reputation on the use of creative and opportunistic kinship links (e.g., adoptive siblings, MZ/DZ twins raised apart). Starting in the mid-1990’s, we spent a considerable amount of research effort attempting to get direct ascertainment questions onto the NLSY surveys, but with little success. Economists at the Bureau of Labor Statistics (BLS, which has managed and financially supported the NLSY79 survey process) were concerned that asking about level of relatedness could potentially offend some of the respondent families, who might drop out of the survey. They recommended that we should obtain independent grant funding for new survey questions to do directed ascertainment (including one cost estimate of around $200,000 of direct cost funding in 1996 to add those questions to one of the surveys). However, we could not obtain permission from cautious BLS administrators to move forward with a grant application, which made the cost a moot issue. To support this effort, we drafted direct ascertainment questions for each survey, pilot tested those questions, and set them aside. Eventually, these questions became useful.

*By the early 2000's, the NLSY-C/YA kinship links needed to be updated.* Many thousands of new children had been born to the NLSY79 females (though childbearing was not completed), and many of the children in the original NLSYC kinship links had aged into the NLSY-YA category for NLSYC respondents aged 15 and older. Between 2003-2005 we updated the NLSY-C/YA links, which were immediately put to use by a number of research teams. Important publications that emerged included Rodgers, Bard, et al. (2008); D'Onofrio et al. (2008); Lahey et al. (2009); Mendle et al. (2009); Hart, Petrill, & Kamp Dush (2010); Salsberry & Reagan (2010); Goodnight et al. (2012); Beaver et al. (2013); Cheung et al. (2014); Rodgers et al. (2015); and a number of other articles. This linking update was completed in 2005, and plain text (CSV) and SAS files containing the links were e-mailed to 50+ researchers who had expressed earlier interest, and then over the next several years to others upon request.

In 2006, an important new development emerged. *Center for Human Resource Research* (CHRR), the Ohio State affiliate institute that manages and distributes the NLS datasets, decided (with support from BLS) to proceed with direct ascertainment. Our original questions were included in the 2006 NLSY79 and the NLSY-YA surveys, and have continued to be included on the NLSY-YA for respondents who passed age 15 in 2008, 2010, 2012, and continuing. We implemented a final three-year project from 2011-2014 to update both the NLSY79 and NLSYC kinship links. During this update, we used both direct explicit ascertainment – information from this type of ascertainment we refer to as “explicit indicators,” leading to “explicit links.” We also used the type of prior, indirect information that was used in earlier linking efforts. We refer to information obtained from that type of ascertainment as “implicit indicators,” leading to “implicit links.” (Note that some kinship links are developed using both types of indicators.) Instead of SAS algorithms, we implemented object-oriented programming using C#. We also began a user services support office, which has assisted researchers with kinship linking projects since 2011, and continuing through at least 2017. We distributed updated NLSY-C/YA kinship links in May, 2012, sent by e-mail to approximately 100 NLSY researchers, and with online availability through R, SAS, and CSV files. The NLSY79 kinship links were distributed in a beta version (using only new explicit indicators) in December, 2012. A completed update of all datasets was distributed in November, 2013, based on both explicit and implicit indicators.

The NLSY97 kinship links have not been developed (by fall, 2015). However, direct ascertainment was included in the original 1997 survey questions, reducing the time required to produce quality kinship links. We have a plan for defining NLSY97 links, implemented within the next few years, including integration of NLSY97 links into the online kinship link system.

**Modern Linking Mechanisms and Validity Studies.** During the 20+ years since we first began developing NLSY kinship linking algorithms, we have implemented a number of logical and data management innovations that we describe in this section. During this period, programming methods have transitioned from SAS-oriented linking algorithms to object oriented programming using C#, and from storage on tapes (originally) and diskettes to online storage (along with online documentation and associated vignettes); there are many advantages to this transition for linking purposes. We produce distribution files as SAS files; CSV files (read by Excel, etc.); and in R. We now post linking files on the web (in CRAN and GitHub), and also our linking algorithm, as an open-source code repository. Note that typical researchers will never need to read the C# code, but rather will only deal with the product of the code (i.e., the kinship links themselves). However, programming-oriented BG researchers or those with particular requirements, may readily adapt the code for special purposes when that flexibility is needed. *(Earlier linking algorithms were considered proprietary; only kinship links, and not the algorithms, were distributed).*

In addition to the linking information, we also post online a set of vignettes that serve two purposes. First, they perform basic data management necessary to use the kinship links, including doing extracts from the NLSY online files, producing double-entry data structure, and merging phenotype scores into the double-entry data structure. Second, they demonstrate and illustrate a number of basic biometrical methods, including fitting univariate ACE models, multivariate ACE models, and Cholesky models (see Neale & Cardon, 1992).

The kinship links themselves have been created using several basic principles. First, we consider information from the 2006 explicit links to be primary. However, both old implicit links and implicit links developed from new indirect questions are used extensively. In many cases the implicit links solve problems that cannot be addressed through the explicit links (because of missing data or other logical weaknesses). Further, in the thousands of redundant links that can be assigned using both explicit and implicit information, there is high agreement between the two methods in how they assign kinship links. However, it is also clear that the combination of the two types of information -- both explicit and implicit indicators of relatedness -- provides higher-quality kinship links than those using only one or the other.

A second principle that we implement is to resolve inconsistencies within families when possible. Because there are many thousands of kinship links to manage, virtually any possible inconsistency than can occur has occurred. As a simple example, if the older two of three brothers indicated in their explicit links that they are full-siblings to their other two brothers, but the third and youngest brother indicated a half-sibling relationship to the older two, this is a logically impossible inconsistency. By using longitudinal structure, the implicit indicators that used information from maternal reports, and additional implicit information from the respondents themselves, we have been able to resolve most such inconsistencies.

Third, in the latest round of grant support we have produced a master kinship linking file that includes all of the NLSY79 links, all of the NLSYC links, and also cross-generational links. Across the NLSY79 (Gen1) and NLSY-C/YA (Gen2) generations are links from mothers to daughters, aunts to nieces and nephews, and uncles to nieces and nephews (see Rodgers et al., 2008, for a paper based on cross-generational links3). It should be noted that, because only the children of NLSY79/Gen1 females are included in the NLSYC/Gen2 dataset, there are no father-child links. The uncle-niece/nephew links are available from the brothers of Gen1 females, whose children are in the Gen2 dataset. There are over 42,000 links within this master file, with variables indicating the type of link. Evaluating cross-dataset and cross-generational invariance properties within the NLSY data will become a new and important source of information for NLSY researchers; in some (perhaps many) cases, NLSYC and NLSY79 files can be combined. Fourth, we have produced both kinship pair structure, and also multi-level within-family data structure (consistent with methods developed by Guo & Wang, 2002; McArdle & Prescott, 2005; van den Oord, 2001). Fifth, we provide the sophisticated user with considerable flexibility by defining flags for certain data problems (e.g., inconsistencies that cannot be resolved). One of these flags notes potential for unreliability within the responses themselves. When respondent or maternal reports are simply in error (because of duplicity, confusion, or incorrect information), no amount of logical data management can correct this problem; however, certain data patterns imply incorrect responses, and we identify those for use by researchers.

Some of our modern validity studies are highly similar to the original methods used in the 1990's. We still use height, weight, body-mass-index (BMI), and other phenotypes in ACE models, and compare the results to meta-analyses and other research studies (e.g., Visschher, 2008). The current results for the updated NLSY79 and NLSYC data show high concurrent validity in relation to other studies. Rodgers et al. (2007, p. 349) listed a number of past kinship link findings along with results from other research teams and data sources that matched; phenotypes included age at first intercourse, birth weight and gestational length, delinquency, and education/IQ/age at first birth. To add to published validity studies from the past, we present validation results for NLSY79 and NLSY-C/YA height later in this paper.

Will’s writing on ROC –late Sept 2015

Many performance metrics informed the linking algorithm’s development. One type resembles receiver-operator characteristic (ROC) methodology, in which agreements and disagreements between the different classifications are monitored (for general discussions of ROC methodology, see Green & Swets, 1966; Lynn & Barrett, 2014). We were most interested in the agreement between the explicit and implicit assignments. In Figure 2, each point along the line indicates the performance of a successive version of the linking algorithm.

For instance, our 1st saved version of the algorithm had 6,101 Gen 2 sibling pairs whose implicit assignment perfectly agreed with their explicit assignment; this was balanced against 417 disagreements (a 15:1 ratio). By the 40th version, there were 7,492 agreements and 355 disagreements (21:1). Figure 2's progression is not always smooth, reflecting how we experimented with different adjustments in order to walk the multidimensional space.

The initial linking versions targeted the easily-classified pairs. Later versions targeted the more difficult pairs; these algorithmic decisions typically considered more NLSY items simultaneously. We required that all adjustments have face validity. We did not pursue the top-left corner at the expense of an algorithmic decision we could not justify.

Other ROC-like comparisons were monitored besides Figure 2. These include the (2) current links against previously-released links (e.g., Rodgers, 1996), (3) Gen1 implicit against the Gen1 explicit assignments, (4) Gen1 explicit assignment against the Gen1 1979 household roster, and the (5) and Gen1 implicit assignment against the Gen1 1979 household roster. Ideally the points in all these metrics/graphs would march to the top left corner (i.e., high agreement and low disagreement), but trade-off decisions were necessary between metrics. And some metrics could not be expressed in ROC terms, such as the total count of Gen 1 pairs linked (because many pairs had only the implicit items, and no definitive assignment from the explicit items or 1979 roster).

New types of validity checks have also recently been developed. One of those relies on receiver-operator characteristic (ROC) methodology, in which agreements and disagreements between the different classification algorithms are monitored (for general discussions of ROC methodology, see Green & Swets, 1966; Lynn & Barrett, 2014). We were most interested in the agreement between the explicit and implicit assignments (see Figure 2). However our algorithm’s development was also informed by other comparisons, including the explicit and implicit assignments against the 1979 household roster.

In summary, our goal within this section has been to provide background and support for the quality of both the NLSY79 and NLSYC kinship links. Besides the overall quality of these links -- which has improved substantially from the earlier links developed in the 1990's and early 2000's -- there are other features of the NLSY data to recommend to researchers. Among those are the broad array of outcomes that can be used to construct phenotype scores for many different variables of interest to BG researchers, the longitudinal structure for BG researchers who study development and aging, the broad age range that now exists to support research from infancy through the mid-50’s (and consistently increasing), and the fact that the original NLSY79 links were obtained within the context of a national household probability sample. The NLSY97 links, once developed, will have many of the same important features.

It is worth emphasizing that many researchers contributing to the corpus of BG literature have consistently sacrificed external validity for high internal validity (see Mook, 1983, for expansive justification of this decision). Especially in using recruited twin samples and many adoption designs, the focus of biometrical research has been on internal validity, and issues of generalizability have often been ignored. The NLSY as a national household probability sample (and whose data release includes complex sample weight calculation and availability) provides a relatively high level of external validity (in combination with potentially high internal validity). Only a few other such datasets (e.g., the Add Health data) support biometrical research with high levels of both internal and external validity and longitudinal structure. In addition, there also exist national registry-based twin samples (e.g., in the U.S., Denmark, Sweden, and Australia, among others) that provide large sample sizes and excellent levels of external validity.

**Publications.** A catalogue of journal articles, book chapters, technical reports, and dissertations/theses that have used the NLSY kinship links is posted online on our kinship link website: [http://liveoak.github.io/NlsyLinks/](https://exchange.ou.edu/owa/redir.aspx?C=HFRNPxKsGEOHOOHaF1l3k9K1a2hFyNEIIjF_Xe8PxuwnmV9TiwLszGk91cX9D0T_xnX6ueT4Sws.&URL=http%3a%2f%2fliveoak.github.io%2fNlsyLinks%2f). Many of these publications emerged from work by Rodgers, Rowe, and their collaborators; a number of other research teams have used the kinship links as well. Those for which we have records are also recorded in the online site above. As of April, 2015, the site above lists 44 research articles (41 refereed journal articles and 3 book chapters), 3 internal publications (including documentation of the NLSY kinship links), and 11 dissertations/theses that have used the NLSY kinship links. Of the 44 research articles, 13 used the NLSY79 kinship links, 24 used the NLSY-C/YA kinship links, and 7 used both the NLSY79 and NLSY-C/YA. The phenotypes that have been investigated include an extensive set of studies of fertility outcomes and precursors, including pubertal development, adolescent sexual behavior, and fertility outcomes; a number of studies of childhood and adolescent problem behaviors and conduct disorder; studies of smoking and drinking, both by NLSY respondents and their parents; several studies of intelligence and/or cognitive development; studies of educational outcomes; studies of health and physical development; a study of childhood maladjustment; a study of the quality of the child's home environment; and several studies that have used combinations of these phenotypes.

Two general methodologies are represented in past NLSY kinship link studies. First, a number of biometrical studies are represented. Some of these are standard heritability studies fitting ACE models to univariate and multivariate outcomes. Others are more complex biometrical designs using Cholesky models, competing pathway models, multivariate models, or other design structures developed for particular purposes (e.g., see the Mother-Daughter-Aunt-Niece, or MDAN, design presented in Rodgers et al., 2007). Prototypes of biometrical research on intelligence and cognitive functioning include Rodgers, Rowe, and May (1994), Rowe, Vaszonyi, & Flannery (1995), Rowe and Cleveland (1996), and van den Oord and Rowe (1997). Examples of biometrical studies of problem behaviors include Rodgers, Rowe, and Li (1994), Cleveland et al. (2000), Rodgers, Buster, and Rowe (2001), Van Hulle et al. (2009), and Connelly and Beaver (2014). Examples of biometrical studies of pubertal development, sexuality, and fertility include Rodgers, Rowe, and Buster (1999), Doughty and Rodgers (2000), Rodgers, Bard, and Miller (2007), and Miller et al. (2010).

The second general methodology has used the relatedness indicators from the kinship links within design structures to control for family background selection factors to assess causal influences on various child outcomes, including sexuality, cognitive development, and problem behaviors. Most of these designs are designated as "children-of-sibling" designs (or, sometimes, “discordant sibling designs”), slightly weaker but often more practical than children-of-twin designs. Some of those studies have also been biometrically informed (e.g., D'Onofrio et al., 2008; Rodgers et al., 2008), but the major innovation is the ability to control family background by linking sisters from the NLSY79 dataset to study child outcomes in the NLSYC dataset. Prototypes include Lahey et al. (2009), Mendle et al. (2009), Jaffee, Van Hulle, and Rodgers (2011), Goodnight et al. (2012), D’Onofrio et al. (2012), and Cheung et al. (2014).

In the sections that follow, we separately assess the NLSY79, the NLSYC, and cross-generational file structures in terms of their utility for fitting biometrical models. Included within these are validity analyses, which illustrate at a basic level some of the principals involved in using the NLSY kinship links. We then emphasize the potential for cross-dataset research, using the “super-file” of kinship links that combine over 42,000 links from two of the NLSY datasets. We conclude with general discussion of the kinship links, including suggestions for biometrical work with all three NLSY data sources.

**Using the NLSY79 Kinship Links**

In the updated NLSY79 kinship links *released in* November, 2013, we have classified 5,038 links out of a possible 5,302 links, a 95% classification rate (compare to the 60% classification rate from the earlier mid-1990's NLSY79 files used in previous studies using the NLSY79 kinship links). Table 1 presents the different kinship categories available for classification, with sample sizes for each category. It is notable that there are many more full siblings than the combination of all other categories, which may appear problematic at first glance. We note, though, that many past successful studies have been conducted using smaller NSLY79 sample sizes with the same unbalanced patterns. Further, these sample sizes reflect *the distribution of* those *kinship categories* in households in the December, 1978 U.S. population. These classifications use the 2006 explicit indicators as the primary classification level, and the earlier implicit logic to fill in missing links and to resolve inconsistencies.

The earlier 1996 NLSY79 kinship links had a much smaller denominator of kinship pairs to classify, because only sibling links involving the oldest NLSY79 sibling in the household were submitted to the linking algorithm. The 2012 NLSY79 links were not based on a larger sample -- the NLSY79 is the same dataset as the original (except for loss at later rounds due to attrition and design adjustments; see the overall summary above). However, all pairs of within-family and cousin pairs were submitted to the linking algorithm to create the most recent links. This adjustment explains the difference between the 3,890 potential links in 1996 and the 5,302 in 2012. Obviously, not only did the number of potential links increase substantially, but also the success rate increased (from 60% to 95%) because of the availability of the explicit indicators collected in the 2006 survey, used in combination with the earlier implicit links.

We used a similar validity analysis for the 2012 NLSY79 kinship links to the one we developed for the 1996 links (the original was documented in Rodgers, 1996). Similarly, the original validity study for the NLSY79 data followed a similar strategy to the one we used two years before in developing the 1994 NLSY-C/YA kinship links. Both relied on several mechanisms, the primary of which was ACE modeling (Neale & Cardon, 1992) of adult height and weight data, which have high and carefully documented heritabilities. We typically have focused on height, which has higher heritabilities than weight.

Johnson et al. (2010) specified height heritabilities between .70-.95, with adult height h2 generally around .80. Visscher et al. (2006) found adult height h2 = .80 in Australian Twin Registry data, and Silverntoinen, Kaprio, & Laheima (2000) found male h2=.78 and female h2 = .75 in a Finnish twin cohort. Furthermore, past research has noted race differences in height heritability, with lower h2 values for African-Americans and especially Hispanics (e.g., Lai & Mayer, 2006; Roberts, Billewicz, & McGregor, 1978). Estimated c2 values for adult height are consistently around zero.

We used gender-standardized and single-entered adult height measures contained within the NLSY79 files. We estimated an ACE model using OpenMx (Neale et al., 2015) with the NLSY79 sibling-based kinship pairs; h2 = .90, c2 = .00, and e2 = .10. These ACE results and kinship correlations are presented in Table 2.

**Using the NLSY-Children/Young Adult Kinship Links**

We released a complete NLSY-C/YA kinship link dataset in May, 2012. In November, 2013, we released updated and final versions of three datasets: NLSY79, NLSY-C/YA, and a “super-dataset” combining both within- and cross-generational links. The NLSY-C/YA links were slightly updated in the 2013 release compared to the 2012, and we report sample sizes from the most recent (2013) release.

The 2013 NLSY-C/YA kinship links file contained 16,083 kinship pairs with identified links (100% of the total); see Table 3. If we exclude the ambiguous sibling and ambiguous twin categories, the sibling classification rate (*not counting* cousins) is 10,466 links out of a possible 11,088 sibling links, a 94% classification rate (compared to the 75% classification rate from the 2005 NLSY-C/YA links). Within the 2013 release, there are 27 MZ twin pairs; 83 DZ twin pairs; 6,914 full sibling (non-DZ twin) pairs; and 3,442 half sibling pairs; these can be considered approximately representative of the distribution of siblings produced by the NLSY79 female cohort. Remaining are 610 sibling kinship pairs who our algorithm could not classify (but who are definitely either full or half siblings; this category was reduced from 2,636 ambiguous pairs in the 2005 kinship links to the current 610, and can be used in some research studies as their own category with R=.375), and 12 ambiguous twin pairs who are definitely *same-sex* twins, but cannot be distinguished by our algorithm as either MZ or DZ twins (and who can also be used in some studies by assigning R=.75). In addition, there are 4,995 cousin pairs.

There are many categories of cousins, because there are many categories of relatedness among the mothers/aunts of cousins. NLSY-C/YA cousin pairs whose mothers are MZ twins are genetically equivalent to half siblings (R=.25); there are 18 “half-sibling cousins.” Pairs whose mothers are full siblings are full cousins (R=.125); there are 3,941 full cousins identified in the 2013 dataset. Pairs whose mothers are R=.375 have an estimated cousin genetic coefficient of R=.09375; there are 12 such identified cousin pairs. Pairs whose mothers are half-siblings (R=.25) are half-cousins with a genetic coefficient of R=.0625; there are 309 such half-cousin pairs. There are several other cousin categories as well, including 131 cousin pairs whose mothers are of entirely unknown relatedness. The overall classification rate for cousins is 97%. In Table 4, we identify all of the possible cousin pair categories in the NLSY-C/YA dataset, and provide sample sizes for each. These large and diverse cousin categories support the kind of cousin studies suggested by Fisher (1919) and discussed earlier in this paper. There is potential to explore the “cousin anomaly,” often noted by NLSY researchers (including our own research, in a number of different studies). The anomaly is that cousin correlations are often too high on a number of different phenotypes to be consistent with other biometrical patterns.

The 3,276 mothers (females from the NLSY79) whose children contribute sibling pairs to the NLSY-C/YA dataset include 1,521 mothers with one kinship pair (i.e., mothers with two NLSY-C/YA children), 94 mothers with two kinship pairs (mothers with three or more children, but missing data for some of the kinship pairs), 858 mothers with three kinship pairs (almost all of whom have three children, a few of whom have more than three children but with missing data for some of their children), 120 mothers with four kinship pairs, 7 mothers with five kinship pairs, 366 mothers with six kinship pairs (almost all of whom had four children, resulting in six pairs of children), and 310 mothers with more than six kinship pairs (ranging from seven to 58 pairs).

We fit the NLSY-C/YA kinship pairs using ACE models to measures of adult height. Adult height was measured in inches in the NLSY-YA survey between ages 19-25. We used gender-standardized measures of adult height to adjust for gender differences. Data were used in single-entry form, ambiguous siblings and twins were deleted, and models were fit using both DF Analysis (Defries & Fulker, 1985; Rodgers & Kohler, 2005) and OpenMx (Neale et al., 2015); results were highly similar, and OpenMx results are presented here. As noted above, meta analyses show adult height h2 is generally around .80, with c2 around 0. ACE model results are presented in Table 5. These estimates are highly consistent with meta-analyses and studies cited above. For the overall sample, height heritability of h2=.82 compares to meta-analysis and other research study predictions from above of around .80. Heritabilities for African-American and Hispanic samples are smaller, as expected.

Though the height results are the ones we use as an example, and the ones to which we give greatest credence, we ran other validity studies using different phenotypes. Those included adult weight, BMI, female age at menarche, female age at first intercourse, and eye color, all used to estimate biometrical models. *These results will be presented in future substantive research, with papers devoted to the several separate topics (e.g., pubertal development; fertility/reproduction; physical development, etc.).* Results consistently matched those from past research, demonstrating concurrent validity in relation to other datasets. *As noted above, Rodgers et al (2007, p. 349) documented concurrent validity across many kinship link studies; it is not surprising that substantially improved versions of the NLSY kinship links also show the same approximate patterns of concurrent validity.*

**Using the NLSY79 and NLSY-Children/Young Adult “Super File”**

With the final release of the NLSY79 and NLSY-C/YA kinship links in November 2013, we also released a master file that has great potential to support interesting and innovative BG (and many other types of) research in the future. We combined 5,302 NLSY79 kinship pairs, 16,083 NLSY-C/YA kinship pairs, and 21,388 cross-generational kinship pairs to create a “super file” with 42,773 kinship pairs. The cross-generational kinship pairs consist of mother-daughter pairs, aunt-niece/aunt-nephew pairs, and uncle-niece/uncle-nephew pairs. 4

This dataset structure can be used for new and innovative research studies based on ideas such as the following. First, because in many cases instrumentation is identical across the two generations, analyses can be run with relatively little concern for instrumentation bias (except that the meaning of some words can change over time). Second, invariance studies can be conducted to determine if the status of the measurement process is fixed, or has changed, over the two different generations. In cases where it has not (and in those where it has with adjustments), the two generations can be combined in a remarkably large biometrically-informed dataset. Third, certain assumptions can be investigated within this dataset. For example, the equal environments assumption, which has been carefully studied within the BG community and the cause of much criticism from outside, can be studied from a unique perspective, because the nature of the shared environment is different across generations (who never shared a childhood rearing environment) compared to within-generation. If results converge between the cross-generational data and the within-generational data, this would *provide at least partial* support for the legitimacy of the equal environments assumption in relation to a given phenotype. Fourth, the NLSY97 kinship links can eventually be added to this super-dataset to support broader research across time and across all three NLSY datasets.

**Conclusion**

In conclusion, we recommend the three NLSY datasets to researchers doing biometrical research (and also to other broader developmental/family studies researchers). *We especially recommend the kinship links to graduate students planning thesis or dissertation research, and to early-career faculty. These recommendations are based on the ease and accessibility of the NLSY data and kinship links, the broad array of research topics that can be developed, and the potential for doing unique and methodologically sophisticated research using these data resources.* A publication history based on around 60 articles, book chapters, technical reports, and dissertations/theses supports the viability of the use of the NLSY kinship links for productive research. *Note that this earlier publication stream was based on earlier versions of the kinship links, which were of lower quality (reliability and validity) and reduced sample sizes compared to those recently released and described in detail within the current paper*. Further, extensive online support and a User Service research support function (available by contacting any of the first three authors) is provided by the current NLSY kinship research team to assist with software and modeling issues associated with using the kinship links. The NLSY datasets are remarkable resources, even without accounting for their biometrical structure. Combining the kinship link information with the long-term longitudinal structure, the cross-generational design, the thousands of phenotypes that can be constructed, and the probability sample origin of the original NLSY79 creates a highly unique and valuable data resource.

**References**

Baydar N, Greek, A (2001) Analysis of data from related individuals. Working paper, Battell Centers for Public Health Research and Evaluation, Seattle WA.

Beaver KM, Connolly EJ, Schwartz JA, Al-Ghamdi AS, Kobeisy AN (2013). [Genetic and environmental contributions to stability and change in levels of self-control](http://www.sciencedirect.com/science/article/pii/S004723521300069X).  Journal of Criminal Justice 41: 300-308.

Chase-Lansdale PL, Mott FL, Brooks-Gunn J, Phillips DA (1991). Children of the National Longitudinal Survey of Youth: A unique research opportunity. *Developmental Psychology,* 27: 918-931.

Cheung AK,, Harden KP, Tucker-Drob EM (2014). [Gene x Environment interactions in early externalizing behaviors: parental emotional support and socioeconomic context as moderators of genetic influences?](http://link.springer.com/article/10.1007/s10519-014-9664-8/fulltext.html) Behavior Genetics 44(5): 468-486.

Cleveland HH, Wiebe RP, van den Oord EJCG, Rowe DC (2000)  [Behavior problems among children from different family structures: The influence of genetic self-selection](http://onlinelibrary.wiley.com/doi/10.1111/1467-8624.00182/abstract). Child Development 71:733-751.

Connolly EJ, Beaver KM (2014) [Examining the genetic and environmental influences on self-control and delinquency: Results from a genetically informative analysis of sibling pairs.](http://www.ncbi.nlm.nih.gov/pubmed/24142448) Journal of Interpersonal Violence 29: 707-735.

DeFries JC, Fulker D (1985) Multiple regression analysis of twin data. *Behavior Genetics* 15: 467–473.

D’Onofrio BM, Van Hulle CA, Waldman ID, Rodgers JL, Harden KP, Rathouz PJ, Lahey BB (2008) Smoking during pregnancy and offspring externalizing problems: An exploration of genetic and environmental confounds. *Development and Psychopathology*  20: 139-164.

D’Onofrio BM, Rickert ME, Langstrom N, Donahue KL, Coyne CA, Larsson H, Ellingson JM, Van Hulle CA, Iliadou AN, Rathouz PJ, Lahey BB, Lichtenstein P (2012)  [Familial confounding of the association between maternal smoking during pregnancy and offspring substance use and problems: Converging evidence across samples and measures](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3622274/). Archives of General Psychiatry 69: 1140-1150.

Doughty D, Rodgers JL (2000) [Behavior genetic modeling of menarche in U.S. females](http://link.springer.com/chapter/10.1007/978-1-4615-4467-8_11). In: Rodgers JL, Rowe DC, Miller WB (eds.), Genetic influences on fertility and sexuality, Boston: Kluwer Academic Press.

Goodnight JA, D’Onofrio BM, Cherlin AJ, Emery RE, Van Hulle CA, Lahey BB (2013) Effects of multiple maternal relationship transitions on offspring antisocial behavior in childhood and adolescence: A cousin-comparison analysis. *Journal of Abnormal Child Psychology*  41: 185-198.

Goodnight JA, Lahey BB, Van Hulle CA, Rodgers JL, Rathouz PJ, Waldman ID, D'Onofrio BM (2012) [A quasi-experimental analysis of the influence of neighborhood disadvantage on child and adolescent conduct problems](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3287068/). Journal of Abnormal Psychology 121: 95-108.

Green DM, Swets JA (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Guo G, Wang JM (2002) The mixed or multilevel models for behavior genetic analysis. *Behavior Genetics* 32: 37-49.

*Harris, K. M., Halpern, C. T., Smolen, A., & Haberstick, B. (2006). The National Longitudinal Study of Adolescent Health (Add Health) Twin Data. Twin Research and Human Genetics 9: 988-997.*

Hart SA, Petrill SA, & Kamp Dush CM (2010) [Genetic influences on language, reading, and mathematics skills in a national sample: An analysis using the National Longitudinal Survey of Youth](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3148850/). Language, Speech, and Hearing Services in Schools 41: 118.

Jaffee S, Van Hulle C, Rodgers JL (2011)  [Effects of non-maternal care in the first three years on children’s academic skills and behavioral functioning in childhood and early adolescence: A sibling comparison study](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3640295/). Child Development 84: 1076-1081.

Johnson W, Turkheimer E, Gottesman I, Bouchard TJ (2010) Beyond heritability: Twin studies in behavioral research. *Current Directions in Psychological Science* 18: 217-220.

Lahey BB, Van Hulle CA, Rathouz PJ, Rodgers JL, D'Onofrio BM, Waldman ID. (2009) [Are oppositional-defiant and hyperactive-inattentive symptoms developmental precursors to conduct problems in late childhood?: Genetic and environmental links](http://link.springer.com/article/10.1007/s10802-008-9257-1).  Journal of Abnormal Child Psychology 37: 45-58.

Lai C, Mayer J (2006) How much of human height is genetic and much is due to nutrition? *Scientific American*, December 11, 2006.

Lynn SK, Barrett LF (2014) “Utilizing” signal detection theory. *Psychological Science* 25: 1663-1673.

McArdle JJ, Prescott CA (2005) Mixed-effects variance components models for biometrical family analyses. *Behavior Genetics* 35: 631-652.

Mendle J, Harden KP, Turkheimer E, Van Hulle CA, D’Onofrio BM, Brooks-Bunn, J, Rodgers JL, Emery RE, Lahey BB (2009) Associations between father absence and age of first sexual intercourse. *Child Development* 80: 1463-1480.

Miller WB, Bard DE, Pasta DJ, & Rodgers JL (2010). [Biodemographic modeling of the links between fertility motivation and fertility outcomes in the NLSY79](http://link.springer.com/article/10.1353/dem.0.0107). Demography 47: 393-414.

Mook DG (1983) In defense of external invalidity. *American Psychologist* 38: 379-387.

Neale MC, Cardon LR (1992) *Methodology for genetic studies of twins and families*.

New York: Springer.

Neale MC, Hunter MD, Pritikin JN, Zahery M, Brick TR, Kickpatrick RM, Estabrook, R, Bates TC, Maes HH, Boker SM (In press). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika* DOI 10.1007/s11336-014-9435-8.

Roberts DF, Billewicz WZ, McGregor IA (1978) Heritability of stature in a West African population. *Annals of Human Genetics* 42: 15-24.

Rodgers JL (1996). *NLSY Youth linking algorithm.* Unpublished manuscript.

Rodgers JL, Bard DE, Johnson AB., D’Onofrio BM, Miller WB (2008). [The cross-generational mother-daughter-aunt-niece design: Establishing validity of the MDAN design with NLSY fertility variables.](http://www.ncbi.nlm.nih.gov/pubmed/18825497) Behavior Genetics 38: 567-578.

Rodgers JL, Bard DE, Miller WB (2007)  [Multivariate Cholesky models of human female fertility patterns in the NLSY](http://link.springer.com/article/10.1007/s10519-006-9137-9). Behavior Genetics 37: 345-361.

Rodgers JL, Buster M, Rowe DC (2001). [Genetic and environmental influences on delinquency: DF analysis of NLSY kinship data](http://link.springer.com/article/10.1023/A:1011097929954). Journal of Quantitative Criminology 17: 145-168.

Rodgers JL, Kohler H-P (2005) Reformulating and simplifying the DF analysis model. *Behavior Genetics* 35: 211–217

Rodgers JL, Rowe DC, Buster M (1999). [Nature, nurture, and first sexual intercourse in the USA: Fitting behavioral genetic models to NLSY kinship data.](http://www.ncbi.nlm.nih.gov/pubmed/10081235) Journal of Biosocial Science 31: 29-41.

Rodgers JL, Rowe DC, Li C (1994). [Beyond nature versus nurture: DF analysis of nonshared influences on problem behaviors.](http://psycnet.apa.org/journals/dev/30/3/374/) Developmental Psychology 30: 374-384.

Rodgers JL, Rowe DC, May K. (1994). [DF analysis of NLSY IQ/achievement data: Nonshared environmental influences.](http://www.sciencedirect.com/science/article/pii/0160289694900116) Intelligence 19: 157-177.

Rodgers JL, Van Hulle C, D’Onofrio BM, Rathouz, PJ, Beasley WH, Johnson AB, Waldman ID, Lahey BB (2015) Behavior problems and timing of menarche: A developmental longitudinal biometrical analysis using the NLSY-Children data. *Behavior Genetics* 45: 51-70.

Rowe DC, Cleveland HH (1996) [Academic achievement in African -Americans and Whites: Are the developmental processes similar?](http://www.sciencedirect.com/science/article/pii/S0160289696900045) Intelligence 23: 205-228.

Rowe DC, Vaszonyi AT, Flannery DJ (1995). [Ethnic and racial similarity in developmental process: A study of academic achievement.](http://pss.sagepub.com/content/6/1/33.short) Psychological Science 6 33-38.

Salsberry PJ, Reagan PB (2010) [Effects of heritability, shared environment and non-shared intrauterine conditions on child and adolescent BMI](http://onlinelibrary.wiley.com/doi/10.1038/oby.2009.485/abstract). Obesity 18(9): 1775–1780.

Shadish WR, Cook TD, Campbell DT (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Silverntoinen K, Kaprio, J, Kahelma E (2000) Genetic and environmental contribgutions to the association etween body height and educational attainment: A study of adult Finnish twins. *Behavior Genetics* 30: 477-485.

van den Oord EJCG, Rowe DC (1997). [An examination of genotype-environment interactions for academic achievement in an U.S. national longitudinal survey](http://www.sciencedirect.com/science/article/pii/S016028969790043X). Intelligence 25: 205-228.

van den Oord EJCG, Rowe DC (1999)  [A cousin study of associations between family demographic characteristics and children's intellectual ability](http://www.sciencedirect.com/science/article/pii/S0160289699000227). Intelligence 27: 251-266.

van den Oord EJCG (2001) Estimating effects of latent and measured genotypes in multilevel models. *Statistical Methods in Medical Research* 10: 393-407.

Van Hulle CA, Waldman ID, D’Onofrio BM, Rodgers JL, Rathouz PJ, Lahey BB (2009)  [Developmental structure of genetic influences on antisocial behavior across childhood and adolescence.](http://psycnet.apa.org/journals/abn/118/4/711/) Journal of Abnormal Psychology 118: 711-721.

Visscher PM (2008) Sizing up human height variation. *Nature Genetics* 40: 489-490.

Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLOS Genetics* 2 e41 316-325.

Footnotes

1 David Rowe, Rodgers' original collaborator on the kinship linking projects, passed away in February, 2003.

2 *In September, 1996*, Baydar and Rodgers/Rowe compared kinship links, and found that their independent efforts produced *highly similar* linking results, *a rate of agreement of 94.3%.*

3 Rodgers et al. (2008) presented the Mother-Daughter-Aunt-Niece (MDAN) design, in which correlations from mother-daughter pairs are compared biometrically to those from aunt-niece pairs. Mothers and aunts are often either the same person (i.e., a mother of one or more NLSYC daughters also was the aunt of one or more NLSYC niece), or are related as sisters or half-sisters, providing substantial control over unobserved heterogeneity due to family background.

4 An interesting challenge involves the verbal identification of *certain* cross-generational pairs, because of implied gender. We specify the Parent-Child kinship links simply as “ParentChild” links. But to *separately* identify Aunt-Niece, Aunt-Nephew, Uncle-Niece, and Uncle-Nephew files would require gender identification of members of the kinship links (*creating* inconsistency *compared* to other kinship links such as Full Sibling, Half Sibling, Twin, and Parent-Children, identified without gender specification). We have searched the literature and the internet for resolutions of this problem; one interesting solution is to refer to non-gendered Nieces or Nephews as “Niblings,” the counterpart to non-gendered “Siblings.” We created our own new name for non-gendered *Uncle/Aunts of* “Niblings;” we use “AUncle” for *this category*. Thus, Uncle/Aunt in Gen1 paired with Niblings in Gen2 are designated as “AUncleNibling” pairs within our files.

**Acknowledgements**

Three grants from the National Institutes of Health (NICHD) have supported the work described within this paper: in the 1990’s, R01-HD21973; in 2003-2007 RO1-HD043265; in 2012-2014, R01-HD065865. The authors express appreciation to the Bureau of Labor Statistics and the National Opinion Research Center for longstanding expertise in conceptualizing, funding, and collecting the NLSY data. Particular appreciation is expressed to many individuals at the Center for Human Resource Research at Ohio State for expert management of the NLSY data, and more specifically for conceptual, statistical, and data management support of past NLSY kinship linking efforts. Those individuals at CHRR include Randy Olsen, Elizabeth Cooksey, Frank Mott, Paula Baker, Steven McLaskie, and Karima Nagy.

Table 1: Kinship link sample sizes (number of pairs) in the two past NLSY79 linking files

1996 2012

R=0(\*) -- 555

R=.0625 (\*)(\*\*) -- 47

R=.125 76 96

R=.25 43 297

R=.375 310 15

R=.50 (\*\*\*) 1,877 4,006

R=.75 32 11

R=1.0(\*) -- 11

----- -----

Total classified (\*\*\*\*) 2,338 5,038

Total pairs 3,890 5,302

% classified (\*\*\*\*\*) 60% 95%

\*We didn’t have the diagnostic information in 1996 to classify R=0, R=.0625, or R=1.0.

\*\* Pairs in the R=.0625 category are half-cousins, i.e., their mothers are half siblings

\*\*\* The R=.50 category, defining full siblings, includes a few DZ twins (who are genetically full siblings).

\*\*\*\*Though the NLSY79 sample didn’t increase any between 1996 and 2012, in the earlier linking effort we linked the first sibling to all other siblings, but did not classify the many other pairwise links within >2 child families. In 2012, all pairs were submitted to the classification algorithm.

\*\*\*\*\* In 1996 we created five different linking files, resulting in linking 53%, 58%, 58%, 60%, and 88% of the 3890 available links. The difference between the algorithms involved how conservative they were with respect to assumptions about the fathers living arrangements across time. After validity analyses, we used the 60% set (called the G3 coefficients in several publications) for all published research studies, and recommended that set of links to other research teams. The kinship sample sizes in this table are for that set of links.

Table 2: ACE results for Adult height, NLSY79 Sample

N h 2 c 2 e 2  N.25 N.50F N.50D N1.0 r.25 r.50F  r.50D r1.0

Total 4064 .90 .00 .10 88 259 3706 11 0.16 0.25 0.45 0.89

Table 3: Kinship link sample sizes (number of pairs) in the three past NLSY-C/YA linking files

1994 2005 2013

R=.125*,* etc. (\*) 165 1,980 4,995

R=.25 184 2,111 3,442

R=.375 (\*\*) 119 2,636 610

R=.50 (\*\*\*) 671 5,666 6,997

R=.75 24 16 12

R=1.0(\*\*\*\*) -- 22 27

-------- ------- -------

Total 1,163 12,431 16,083

\*The R=.125*,* etc. category defines cousin links, which are defined by identifying mothers who are sisters, and linking their children as cousins. The majority of mothers are full siblings, resulting in R=.125 cousin links. But there are many other potential categories. When mothers are half siblings, their children are R=.0625 half-cousins. When mothers are themselves half cousins, their children are R=.03125 quarter-cousins. In the 2005 kinship links, there were 1,926 full cousins and 54 half-cousins. In the 2013 links, there are many cousin categories. In this table we combine all of the various types of cousins within the R=.125, etc*.* category. The separate cousin categories for the 2013 links are separated in Table 4.

\*\* All NLSY-Children sibling pairs are twin, full or maternal half siblings, because all NLSYC respondents within the same family are biological offspring of the same NLSY79 mother. The R=.375 category, the “ambiguous siblings,” are composed of NLSYC children links who are either full or half siblings, but whom we cannot classify (due to missing data, etc.). We have used a .375 classification in past work for these pairs (see discussion in text).

\*\*\*The R=.50 category, defining full siblings, includes a few DZ twins (who are genetically full siblings). For example, there are 83 DZ twins among the 5,666 full siblings in the 2005 and the same 83 among the 7,036 full siblings in the 2013 kinship links

\*\*\*\* In 1994, there was no diagnostic information to classify identical twins; all same-sex twin pairs were assigned to an R=.75 category. In 2005 and 2013, the R=.75 twin category contained same-sex twin pairs who could not be classified.

Table 4: Kinship links for various types of cousins, categories and sample sizes (number of pairs) in the most recent (2013) NLSY-C/YA linking files

Cousins’ Sample Mothers’ Mother’s

R Coef Description Size Relatedness R Coef

R=0 Genetically 314 Adoptive Siblings 0 Unrelated Cousins

R=.015625 Eighth Cousins 61 Half Cousins .0625

R=.03125 Quarter Cousins 204 Full Cousins .125

R=.0625 Half Cousins 309 Half Siblings .25

R=.09375 Unknown Half/ 12 Ambiguous .375

Full Cousins Siblings

R=.125 Full Cousins 3,941 Full Siblings .50

R=.1875 Full Cousins, Mothers 5 Ambiguous .75

Ambiguous Twins Twins

R=.25 Full Cousins, Mothers 18 MZ Twins 1.0

MZ Twins

R=? Cousins, Mothers Unknown ??

Relatedness Unknown 131

-------

Total 4,995

Table 5: ACE results for Age 19-25 height standardized by gender, 2013 NLSY-C/YA Sample

Subgroup N h 2 c 2 e 2  N.25 N.50F N.50D N1.0 r.25 r.50F  r.50D r1.0

Total 6916 .82 .00 .18 2114 4728 48 26 0.25 0.39 0.50 0.95

FF 1744 .94 .00 .05 562 1158 14 10 0.32 0.45 0.23 0.95

MF 3444 .56 .09 .35 1062 2356 26 0 0.23 0.36 0.50

MM 1724 .81 .00 .19 490 1214 8 16 0.20 0.36 0.94

Hispanic 1742 .29 .26 .46 396 1336 10 0 0.33 0.40 0.04

Black 2818 .58 .02 .40 1324 1466 18 10 0.18 0.29 0.41 0.88

NBNH 2356 .74 .00 .26 394 1926 20 16 0.27 0.34 0.56 0.95

Hisp FF 408 .02 .45 .53 106 302 0 0 0.42 0.47

Hisp MF 832 .47 .14 .39 190 634 8 0 0.27 0.37

Hisp MM 502 .08 .35 .57 100 400 2 0 0.35 0.40

Black FF 768 .15 .24 .61 376 382 6 4 0.28 0.32

Black MF 1414 .47 .07 .46 664 742 8 0 0.18 0.30

Black MM 636 .53 .00 .47 284 342 4 6 0.07 0.24

NBNH FF 568 .98 .00 .02 80 474 8 6 0.20 0.46

NBNH MF 1198 .32 .15 .53 208 980 10 0 0.26 0.30 0.35

NBNH MM 590 .71 .00 .29 106 472 2 10 0.32 0.29 0.94

Notes: R=0 and all cousin categories omitted from analysis, see text on cousins for justification

N=# kinship pairs; .25=Half Siblings, .50F=Full Siblings; .50D = DZ twin pairs; 1.0= MZ twin pairs

Correlations based on sample sizes less than 10 kinship pairs are omitted

Figure 1: Time relationship among the three NLSY dataset birth periods, survey dates, and childbearing periods, as of 2010; black arrows indicated that the process is still ongoing; blue vertical line indicates year in which explicit kinship indicators were collected in NLSY79 and NLSYC. (DOB stands for “Date of Birth”)

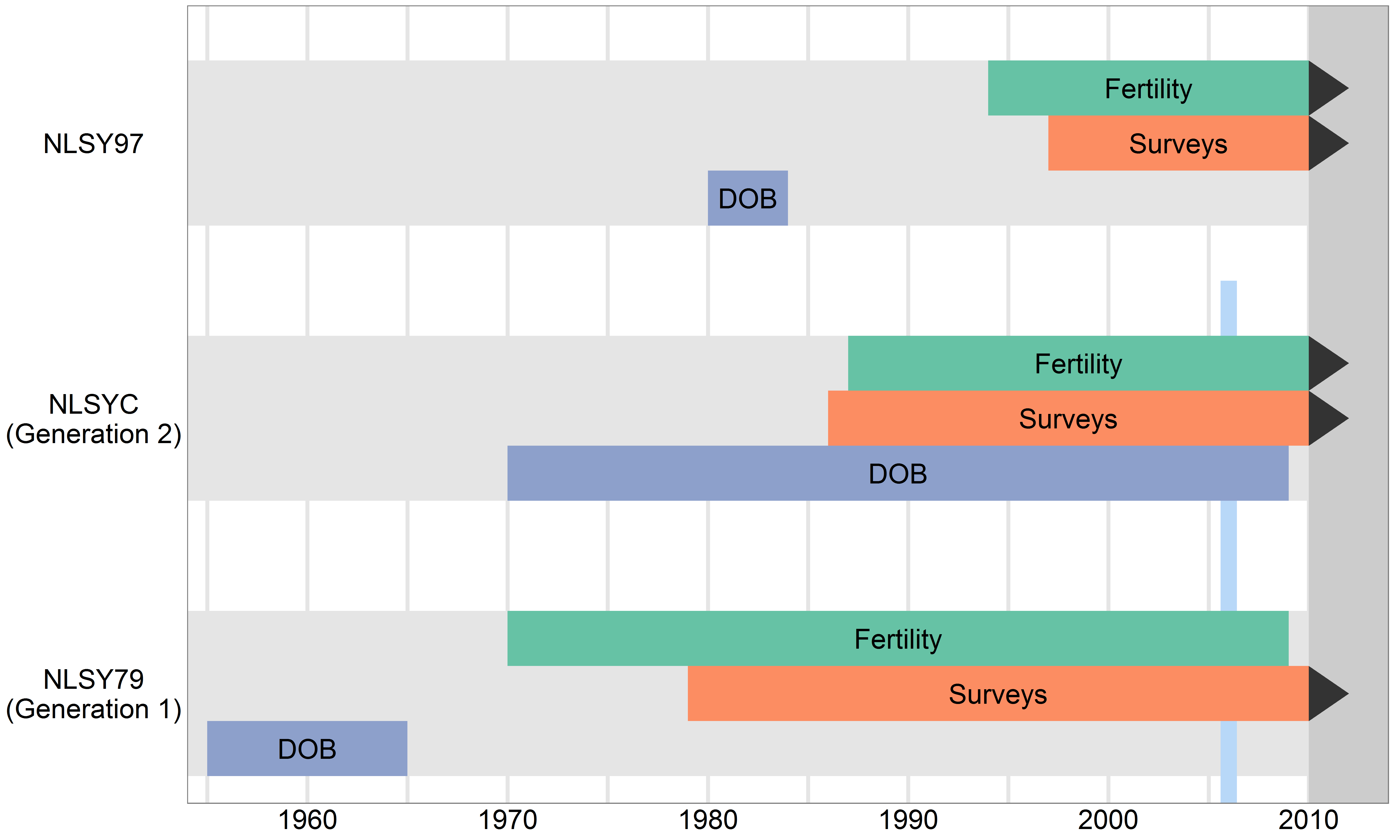


Figure 2: Using ROC curve logic, the performance of each algorithm version was monitored to inform future versions. The tension between the implicit and explicit assignments among Gen2 pairs eventually arrived at a 21:1 agreement-to-disagreement ratio. The axes represent the counts of agreement/disagreement, and the text indicates the algorithm version (which progresses from version 1 to version 40). The right panel zooms on the relevant movement.

