Journal of Cognition and Development

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hjcd20

Secondary Data Analysis: An Important Tool for Addressing Developmental Questions

Andrea Follmer Greenhoot a & Chantelle J. Dowsett a

a University of Kansas
Published online: 09 Feb 2012.


To link to this article: http://dx.doi.org/10.1080/15248372.2012.646613

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages
whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
TOOLS OF THE TRADE

Secondary Data Analysis: An Important Tool for Addressing Developmental Questions

Andrea Follmer Greenhoot and Chantelle J. Dowsett

University of Kansas

Existing data sets can be an efficient, powerful, and readily available resource for addressing questions about developmental science. Many of the available databases contain hundreds of variables of interest to developmental psychologists, track participants longitudinally, and have representative samples. In this article, the authors discuss the benefits and limitations of secondary data analysis so that researchers can make informed decisions about when it makes sense to use existing data and when it does not. They then provide an overview of best practices for handling the most common conceptual, methodological, and practical issues that arise when adopting this approach.

Readers of this journal are probably all too familiar with the many data collection challenges inherent in developmental research. Participants in developmental studies are difficult to recruit and retain in large numbers, and tracking participants’ development prospectively in a longitudinal study can be expensive and time consuming. Secondary analyses of existing data sets can be an efficient and powerful solution to some of these challenges. Rapid advances in 21st-century technology have enabled the construction of digital archives of vast amounts of data in all disciplines. Many of these databases...
contain hundreds of variables of interest to developmental scientists and track individual participants across extended periods of time, creating unprecedented opportunities to study developmental change in an era of shrinking grant funding and pressure to produce generalizable, high-impact findings.

Although secondary data analysis is becoming more common in developmental science, it is far less frequent than in many other disciplines, such as economics (Duncan, Engel, Claessens, & Dowsett, 2011). Many psychological scientists are hesitant to use existing data sets, in part because the approach is not included in standard methodological training in psychology and because it is undervalued within some circles. Therefore, our goals in this article are to 1) highlight the benefits and the limitations of secondary data analysis so that researchers can make informed decisions about when it makes sense to use existing data, and 2) describe best practices for handling common conceptual, methodological, and practical issues that arise when adopting this approach.

SECONDARY DATA ANALYSIS DEFINED

Secondary data analysis involves use of existing data by investigators who were not involved in the original data collection. These data can be analyzed to replicate or extend previously observed findings or to address new research questions that were not a part of originally published analyses of the data. Some secondary data sets are informally shared between researchers, whereas others are publicly available through clearinghouses (e.g., the University of Michigan’s Interuniversity Consortium for Political and Social Research [ICPSR]) or federal agencies (e.g., the U.S. Bureau of Labor Statistics’ National Longitudinal Surveys [NLS]). Although many publicly available data sets have open access, some are restricted to licensed users who have gone through the required application process to use the data. We provide details about how to access secondary data sets and provide a list of public-access data sets that may be especially useful for cognitive development research in later sections of this article. First, we turn to the issue of when and why developmental researchers might consider using existing data sets in their investigations.

WHY USE EXISTING DATA SETS

There are a number of reasons why a researcher might choose to carry out secondary analyses of an existing data set rather than design a new study to collect original data. The most obvious rationale is that the data have
already been collected, freeing the researcher to devote time and effort to other steps in the scientific process.

Several other advantages have to do with the fact that many shared data sets, especially public-use data sets, have very large samples, measures of many constructs, and longitudinal designs, which enable researchers to address questions that they may otherwise lack the time or resources to investigate. For developmental researchers, the potential to draw on longitudinal data to examine change and stability in individuals over time is especially attractive (Chase-Lansdale, Mott, Brooks-Gunn, & Phillips, 1991). Further, even greater longitudinal coverage of the life course can be afforded by aggregating multiple existing data sets (see McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). For early-career researchers, secondary data sets may provide their only opportunity to address longitudinal research questions.

Existing data sets can also enable researchers to study the precursors of atypical, infrequent developmental patterns (e.g., diagnosis of a disorder, peer rejection) or antecedents to unexpected or infrequent events (e.g., natural disasters, parent death). Researchers often resort to retrospective methods to explore precursors to unexpected outcomes (e.g., Greenhoot, 2011), but large publically available data sets or aggregated existing data sets present a viable option for prospective investigations. Another advantage is that some public-use data sets are collected using complex sampling procedures, resulting in representative samples and highly generalizable findings. Large-sample databases also provide unique opportunities to study specialized subpopulations, and some even oversample on specific characteristics. The large numbers of constructs measured in many secondary data sets allow researchers to test complex hypotheses involving multiple variables, and large samples facilitate the use of statistical techniques (e.g., structural equation modeling [SEM]) that can address such hypotheses.

On a broader level, there is an emerging consensus that the sharing and analysis of existing data sets is an essential strategy for advancing a cumulative, and often collaborative, science. This consensus is reflected in recent policy changes and open-source initiatives in science and scientific funding. Many of the funding agencies that support research on cognitive development now encourage or require that data collected through their funded studies are made publicly available. For example, the National Institutes of Health (NIH) now has a policy that requires applicants for substantial funding to include a plan for the “timely release and sharing” of data or to state why data sharing is not feasible (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html). The NIH position allows the original investigator to benefit from first and continuing use but not from prolonged exclusive use of data and defines “timely release” as no later than the acceptance for publication of the main findings from the final data set. Such policies recognize that large-scale
investigations that produce rich and complex data sets are unlikely to be fully used by a single set of investigators. Making such data sets available to other investigators maximizes the yield of grant dollars for the advancement of science and also increases the probability of replication. A related benefit is that an open-source approach encourages good scientific practices, including transparency, careful documentation, and thorough vetting of results before publication (Donnellan, Trzesniewski, & Lucas, 2011).

Analyses of secondary data sets have enabled developmental scientists to efficiently tackle significant and often-challenging research questions that replicate, reinterpret, or expand on key findings in the field. For instance, Zhai, Brooks-Gunn, and Waldfogel (2011) used a large existing data set to produce a compelling evaluation of the impact of Head Start on school readiness relative to other child care settings. In contrast to previous research, the large and diverse sample, drawn from hospitals serving disadvantaged communities, enabled the researchers to control for selection bias (through propensity score matching) and to establish clear reference groups defined by child care arrangements. In another example, Duncan and colleagues (2007) used six longitudinal data sets and a meta-analytic technique to test whether findings related to school readiness were robust to differences between samples, measures, or other design features. New techniques for integrative data analysis are also advancing our understanding of developmental phenomena by enabling researchers to pool data and integrate findings across multiple longitudinal data sets (e.g., Curran & Hussong, 2009). For instance, McArdle et al. (2009) combined three longitudinal studies of intelligence to examine cognitive change across a much longer period (2 to 72 years) than any previous study of cognitive development or aging.

New scientific insights have also emerged from opportunities to access and recode videotaped assessments, transcripts, narratives, or other raw data and apply newly developed models or methodologies to the data. Fivush, Marin, Crawford, Reynolds, and Brewin (2007), for example, applied a coding scheme to written narratives from an expressive writing intervention study with adolescents (Reynolds, Brewin, & Saxton, 2000). Their analyses revealed patterns that contradicted predictions from the adult literature and set the stage for new empirical and theoretical work in this area.

LIMITATIONS

Despite the many advantages, secondary data analysis is not suitable for all researchers or all research questions. Because the data are already collected, the researcher has no control over who was sampled, what constructs were measured, or how they were measured. Thus, the first step in evaluating the
appropriateness of this approach is to determine whether an existing data set provides a good match to an investigator’s research questions. Use of an existing data set is clearly inappropriate if the data set does not contain measures of the constructs of interest or does not draw from the population of interest.

Several additional limitations or pitfalls of secondary data analysis should be considered before adopting the approach. First, even if the data set measures the desired constructs, the measurement instruments selected by multidisciplinary teams that designed the project may not be those that the secondary researcher would have selected him or herself. Many large-scale surveys truncate standard measures or use less intensive data collection protocols to save time and money. Therefore, the secondary data analyst must carefully evaluate the reliability and validity of selected measures and be prepared to address measurement concerns that arise during the peer review process. The complex sampling design of some existing data sets also requires sophisticated analytic techniques and/or the use of software that can properly incorporate these elements in the analysis. Aggregating data sets may also necessitate statistical consultation as it introduces additional complications such as varied measurement of the same constructs.

Another pitfall to avoid is underestimating the time it takes to become familiar with a new data set. It can take a surprising amount of time to get to know a data set and select and capture the data on the constructs of interest. Existing data sets—both public and private—vary widely in the quality of data documentation and the amount of time needed to become familiar with the data set. Often, the problem is that there are too many data points!

Yet another consideration is that unbeknownst to the secondary data analyst, another researcher could be working on a similar or even identical idea. Thus, a very real concern is that someone else could publish a comparable article first using the same data set(s). Furthermore, when publicly available data are used, it is not uncommon for others to try to duplicate the findings. At a minimum, the secondary data analyst should be prepared for continued discussion of the findings as other researchers replicate and extend his or her work. Some researchers also explicitly investigate whether their findings are robust to variations in model specification and to different analytic techniques (e.g., Zhai et al., 2011).

**GETTING STARTED WITH SECONDARY DATA ANALYSIS**

**Gaining Access to Existing Data Sets**

Some data sets are informally disseminated between two or more investigators; with this type of data sharing, agreements regarding data access are
handled on a case-by-case basis. Several major data archives (see Table 1) such as the ICPSR, house and disseminate large numbers of data sets for a multidisciplinary scientific community. Researchers may both deposit and access data sets on the Web sites of these data archives. Publicly available federal data sets, such as the National Longitudinal Survey of Youth and the Early Childhood Longitudinal Studies, are available on the supporting agencies’ Web sites. Researchers can use open-access data sets by completing a request and downloading the data. As a rule of thumb, in open-access data sets, information that could be used to identify individual participants directly or indirectly has been recoded or removed. In contrast, restricted-use data sets include more detailed data, including individually identifiable information; therefore, they are only available to researchers who successfully apply for access and employ data security procedures. Table 1 lists publicly available data sets—both open and restricted access—that may be useful for cognitive developmental research.

For some data sets, users can perform online analyses on the Web site without downloading files. For example, ICPSR offers online analysis components for selected data collections using survey documentation analysis. Similarly, the National Center for Education Statistics [NCES] Web site allows users to analyze publicly available data using several different tools (e.g., the data analysis system), although there are limitations on the types of analyses that can be conducted and the types of variables that can be included.

Ethics and Institutional Review Board Approval

Whether institutional review board (IRB) approval is required for secondary data analysis depends on the characteristics of the data set and the specific human subjects policies of one’s institution. Federal regulations from the Department of Health and Humans Services do not require IRB approval if a secondary data set has been stripped of all information that would allow the participants to be identified. Further details can be found in the Code of Federal Regulation, Title 45 Public Welfare, Department of Health and Human Services, Part 46.101(b)(4): Protection of Human Subjects (http://ohsr.od.nih.gov/guidelines/45cfr46.html). Many publicly available data sets meet the federal criteria for exemption, but restricted-access data sets do not because they include individually identifying information. Nevertheless, practices vary across institutions, and investigations that might be exempt at some institutions might require full or expedited review at others. We recommend that an investigator always consult with his or her own institution’s IRB while designing a study involving secondary data analysis to determine whether their investigation would be exempt from IRB review.
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Links and data access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interuniversity Consortium for Political and Social Research (ICPSR)</td>
<td>Large-scale searchable data archive hosted by the University of Michigan with more than 700 consortium member institutions; largest nongovernmental social science archive in the world.</td>
<td><a href="http://www.icpsr.umich.edu/icpsrweb/ICPSR">http://www.icpsr.umich.edu/icpsrweb/ICPSR</a></td>
</tr>
<tr>
<td>Henry A. Murray Research Archive</td>
<td>Large-scale data archive housed at Harvard University. Collection includes numeric data as well as video data, audio data, and interview notes.</td>
<td><a href="http://www.murray.harvard.edu">http://www.murray.harvard.edu</a></td>
</tr>
<tr>
<td>The Howard W. Odum Institute for Research in Social Sciences</td>
<td>Archive housed at University of North Carolina Chapel-Hill; includes extensive U.S. Census data and National Center for Health Statistics data, among other sources.</td>
<td><a href="http://www.irss.unc.edu/odum/jsp/home.jsp">http://www.irss.unc.edu/odum/jsp/home.jsp</a> Most holdings are publicly available but some are only available to University of North Carolina affiliates.</td>
</tr>
<tr>
<td>UK Data Archive</td>
<td>Archive of more than 5,000 data sets, many from the United Kingdom, but also includes some cross-national data sets.</td>
<td><a href="http://www.data-archive.ac.uk">http://www.data-archive.ac.uk</a></td>
</tr>
<tr>
<td>National Archive of Computerized Data on Aging</td>
<td>Archive of data on aging funded by the National Institute on Aging and housed in the ICPSR.</td>
<td><a href="http://www.icpsr.umich.edu/icpsrweb/NACDA">http://www.icpsr.umich.edu/icpsrweb/NACDA</a></td>
</tr>
<tr>
<td>Integrated Analysis of Longitudinal Studies of Aging</td>
<td>Research network on aging intended to facilitate collaborative research on age-related changes in health, personality, and cognition. Includes both public-use and nonpublic data sets.</td>
<td><a href="http://www.ialsa.org">http://www.ialsa.org</a></td>
</tr>
<tr>
<td>Study Type</td>
<td>Description</td>
<td>Website</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>National Longitudinal Survey of Youth</td>
<td>Two broad longitudinal surveys of large, nationally representative samples of youth. One cohort was first surveyed in 1979 at 14 to 22 years of age and followed annually through 1994; other cohort was first surveyed in 1997 at 12 to 16 years and followed annually until present.</td>
<td><a href="http://www.bls.gov/nls">http://www.bls.gov/nls</a> Public-use files can be downloaded from the Web site after registering.</td>
</tr>
<tr>
<td>Early Childhood Longitudinal Study</td>
<td>A three-cohort, longitudinal, nationally representative study examining child development, school readiness, and early school experiences; birth cohort followed children born in 2001 through kindergarten; kindergarten class of 1998-1999 followed from kindergarten to eighth grade; and kindergarten class of 2010-2011 to be followed from kindergarten to fifth grade.</td>
<td><a href="http://nces.ed.gov/ecls/index.asp">http://nces.ed.gov/ecls/index.asp</a> After registering, a researcher can analyze some data sets on the Web site and download some data sets from the Web site. Some data sets are restricted access; the National Center for Education Statistics provides licenses to institutions or organizations, and qualified researchers can apply for access through their institutions. For more information, see <a href="http://nces.ed.gov/statprog/rudman/introduction.asp#red">http://nces.ed.gov/statprog/rudman/introduction.asp#red</a></td>
</tr>
<tr>
<td>National Education Longitudinal Study</td>
<td>A school-based longitudinal study of a nationally representative sample of participants surveyed in 1988 as eighth graders and assessed again in 1990, 1992, 1994, and 2000. Surveyed topics related to school, work, and home experiences.</td>
<td><a href="http://nces.ed.gov/surveys/nels88">http://nces.ed.gov/surveys/nels88</a> Some data sets can be analyzed (after registering) on the Web site; qualified researchers can apply for access to restricted-access data sets through their institutions.</td>
</tr>
<tr>
<td>Center for Longitudinal Studies, Institute of Education at the University of London</td>
<td>Archive of data from three birth cohort studies collecting information on education, family, parenting, physical and mental health, and employment:</td>
<td><a href="http://www.cls.ioe.ac.uk">http://www.cls.ioe.ac.uk</a> Data can be accessed by registering with the UK data archive. Restricted-access data sets may be accessed by applying for a license.</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Links and data access</th>
</tr>
</thead>
<tbody>
<tr>
<td>1958 National Child Development Study, 1970 British Cohort Study, and Millennium Cohort Study.</td>
<td>Longitudinal study of more than 1,000 children examining relations between child care and children's social, emotional, intellectual, and language development.</td>
<td><a href="http://www.nichd.nih.gov/research/supported/secy/care/overview.cfm">http://www.nichd.nih.gov/research/supported/secy/care/overview.cfm</a> Qualified researchers can apply for access to some data sets at the ICPSR Web site: <a href="http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21940/documentation">http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21940/documentation</a></td>
</tr>
<tr>
<td>National Children’s Study</td>
<td>Longitudinal study of environmental influences on child health and development in a sample of 100,000 children, followed from age 0 to 21 years, beginning in 2012.</td>
<td><a href="http://www.nationalchildrensstudy.gov">www.nationalchildrensstudy.gov</a> Not yet available</td>
</tr>
</tbody>
</table>
The use of informally shared data sets introduces additional ethical considerations. First, clear agreements regarding data use, acknowledgement, and authorship should be developed in advance. We also recommend that the secondary data analyst always inform primary investigators of planned submissions based on their data, regardless of authorship and acknowledgement agreements.

Choosing Existing Data Sets

It is of the utmost importance that the investigator selects data set(s) that are congruent with his or her research question(s). Major archives and public data sets typically provide study-level information (metadata) about the content areas assessed, the mode of data collection, the time period of data collection, and the sampling procedures, as well as codebooks with detailed documentation of variable-level information. Nevertheless, it is not always easy to find out whether a data set contains constructs of interest. Some archives allow researchers to search by topic. The metadata and codebooks do not usually include scale names, so researchers usually have to search for keywords or common phrases, or download and browse the codebooks. In the end, many researchers find several relevant data sets and must select one based on its strengths and weaknesses and the primary goals of the investigation. For example, one data set may have a longitudinal design and good measurement of key constructs but a small sample size, whereas another may have truncated measures on a nationally representative sample that includes key sub-populations. If examination of a subpopulation is a major goal, the researcher might select the latter data set at the expense of measurement depth.

BEST PRACTICES IN SECONDARY DATA ANALYSIS

Getting to Know Existing Data Sets

Once a data set is selected, the researcher must spend ample time becoming familiar with that data set by inspecting the codebooks and other supporting documentation about sampling design and procedures. Ideally, codebooks or the equivalent should accompany informally shared data sets, but the quality and completeness of the documentation will vary widely across projects. For each variable, the codebook should list information about the name of the variable, the wording of the item that corresponds to the variable, missing data codes, value ranges and frequency distributions, imputation or transformation details, and calculation details for constructed variables. Copies of all instruments may also be available; these should be inspected for
Preparing Data Sets for Analysis

As a researcher prepares a secondary data set for analysis, he or she should have a carefully thought-out theoretical or conceptual model and a clear idea of the types of variables needed to test the model. Most methodologists recommend that researchers avoid the temptation to pull nonfocal variables into the analysis and “fish” for interesting associations that are not theoretically driven because this approach can increase the probability of a Type I error (e.g., Magee, Lee, Guiliano, & Munro, 2006; McCall & Appelbaum, 1991; but see Donnellan et al., 2011). Once the variables have been identified, the researcher must construct a usable data set for secondary analysis. For some data sets and research questions, this process is as simple and straightforward as downloading a file, but when the desired variables are provided in multiple files or when combining multiple databases, the construction of a usable data set can require multiple steps that introduce the possibility of error. Common sources of inaccuracy include different variable scaling at different time points (e.g., age in months vs. years), variation in variable names, definitions, or coding across data sets (e.g., gender coded 0 to 1 vs. M or F), and different missing data codes across data sets (e.g., −999 vs. missing). Table 2 lists recommended resources for further information on preparing data sets for secondary analysis, as well as the other common analytic issues and pitfalls reviewed here.

Evaluating Measurement Issues

The abbreviated instruments that are common in secondary data sets tend to have poorer psychometric properties than the original forms. To address measurement concerns, researchers should evaluate the reliability and validity of all short forms from the actual data set being used in the research. Widaman, Little, Preacher and Salawani (2011) established that coefficient alpha underestimates homogeneity among scale items in short forms and instead recommend using coefficient omega to index internal consistency reliability (for calculation details, see Widaman et al., 2011, or McDonald, 1999). Researchers can assess test–retest reliability with Pearson correlations between two administrations of the measure at adjacent time points and can assess validity by examining whether correlations between the short form and other variables are similar to those observed with the long form in other
<table>
<thead>
<tr>
<th>Analytic issue</th>
<th>Description</th>
<th>Recommended resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluating Measurement Issues:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluating Measurement Issues:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Analytic issue</th>
<th>Description</th>
<th>Recommended resource</th>
</tr>
</thead>
</table>
published work. When examining associations among measures in a data set with short forms, Widaman et al. (2011) recommend a correction for attenuation to adjust the correlations for the short form’s lower reliability, or construction of latent variables using an SEM approach, which automatically adjusts for attenuation.

Another measurement issue in secondary data analysis is that the same constructs are often measured by different instruments, either across different data sets or at different time points within one data set. McArdle et al. (2009) summarize several techniques for handling lack of measurement equivalence, including SEM and linked item response theory. Even when the same scales are used, measurement invariance should be tested using confirmatory factor analysis (e.g., Horn & McArdle, 1992; T. D. Little, in press).

Dealing With Complex Sampling Designs

Many large secondary data sets are collected with complex sampling procedures and violate the assumptions required for common statistical techniques. In some data sets, the observations are not independent (e.g., children nested in classrooms nested in schools). Thus, the analysis model must take into account the hierarchical structure of the data and appropriately identify the standard errors, which can then be used for more precise confidence intervals and significance tests (Raudenbush & Bryk, 2002; Singer & Willett, 2003). This adjustment can be made in standard statistical packages by specifying the “cluster” variable (e.g., classroom) or by explicitly modeling the nested data structure using hierarchical linear modeling (HLM) or SEM.

Some studies use stratified random sampling, or disproportionate sampling, to increase coverage of underrepresented groups and/or ensure that the sample is representative of the designed population. But because different individuals in the population have unequal probabilities of being selected, standard statistical procedures that assume random sampling are inappropriate and will underestimate the true variance (Cohen, 1997; Shrout & Napier, 2011). Statistical analyses must include the sampling weights and stratification variables to reflect these unequal probabilities so that the effect of an oversampled group in the analyses more accurately reflects the effect in the population. Sampling weights, stratification, and/or cluster variables are usually included in publicly available data sets, and the data set documentation will help the analyst locate them. Some agencies and data archives also host workshops on how to identify and use the appropriate sampling weight in publicly available data sets. Some software packages (e.g., SUDAAN, SAS) enable the researcher to include these variables and adjust for survey sampling design. Other software packages, such as Stata and the Statistical
Package for the Social Sciences (SPSS), require replicate weights. See Levy and Lemeshow (2008) for a more detailed discussion of strategies for dealing with complex sampling designs.

Handling Missing Data

As the number of measures and assessment occasions in a data set increases, missing data are increasingly likely to be a problem. Many large-scale, longitudinal, public data sets have sampling weights that include a correction for sample attrition, but there are other sources of missing information, such as nonresponse within the survey. The quantitative literature suggests that the two most appropriate methods for handling missing data are: 1) full-information maximum likelihood (FIML) estimation, which can be invoked in standard software packages for HLM or SEM, and 2) multiple imputation. Multiple imputation uses associations among observed values in the data set to estimate missing data values before the analyses, whereas the FIML approach applies estimation methods during model fitting (see Enders, 2010; R. Little & Rubin, 1987; Schafer & Graham, 2002). Both methods are appropriate when the missingness is unrelated to the “true” value for the variable (i.e., either missing completely at random or missing at random; see Little & Rubin, 1987). When the data appear to be missing not at random (MNAR; missingness is systematically related to the “true” value), steps should be taken to recover the data. Enders (2010) provides a review of recently developed methods for MNAR conditions, although these may require specialized software.

RECOMMENDATIONS AND CONCLUSIONS

Overall, what is to be gained from analysis of existing data sets? Although some investigators feel that primary research is best, secondary data analysis holds tremendous potential for developmental researchers. There are existing data sets that are appropriate for all types of developmental questions, often with the sample size and longitudinal design that permits in-depth investigation of the mechanisms underlying developmental processes. Well-established statistical techniques are available to help researchers avoid or deal with most potential pitfalls. We believe that, relative to the time, effort, and resources spent collecting primary data, the challenges with secondary data use are actually quite minimal. If available existing data fit an investigator’s research question(s), the analysis of secondary data can be an efficient use of time and resources and can advance science by allowing researchers to address questions that they may be ill prepared to address with their own data.
ACKNOWLEDGMENT

The authors wish to thank Mijke Rhemtulla and two anonymous reviewers for their suggestions on earlier versions of this manuscript.

REFERENCES


