Published in final edited form as:

Dev Rev. 2013 December ; 33(4): 357–370. doi:10.1016/j.dr.2013.08.003.

Sampling in Developmental Science: Situations, Shortcomings, Solutions, and Standards

Marc H. Bornstein, Justin Jager, and Diane L. Putnick

Child and Family Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Public Health Service

Abstract

Sampling is a key feature of every study in developmental science. Although sampling has far-reaching implications, too little attention is paid to sampling. Here, we describe, discuss, and evaluate four prominent sampling strategies in developmental science: population-based probability sampling, convenience sampling, quota sampling, and homogeneous sampling. We then judge these sampling strategies by five criteria: whether they yield representative and generalizable estimates of a study's target population, whether they yield representative and generalizable estimates of subsamples within a study's target population, the recruitment efforts and costs they entail, whether they yield sufficient power to detect subsample differences, and whether they introduce "noise" related to variation in subsamples and whether that "noise" can be accounted for statistically. We use sample composition of gender, ethnicity, and socioeconomic status to illustrate and assess the four sampling strategies. Finally, we tally the use of the four sampling strategies in five prominent developmental science journals and make recommendations about best practices for sample selection and reporting.

When we undertake to study some phenomenon, we wish to know something about that phenomenon in a population, but in practice we study the phenomenon in a group of individuals who purportedly represent the target or reference population to whom we wish our results to generalize. That is, we *sample* the population. We sample because we normally do not command the resources (time, money, or personnel) to assess the entire population of interest. Sampling is therefore a key feature of every study in developmental science, and sampling has far-reaching implications in all studies. This article is concerned with sampling in developmental science. As we point out, different sampling strategies exist, and each has its implications. Employing sub-optimal sampling strategies is far too common in developmental research, compromises the validity and utility of the research, renders replication and cross-study comparisons difficult, and most generally impedes progress in the field of developmental science.

Address correspondence to: Dr. Marc H. Bornstein, Child and Family Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Suite 8030, 6705 Rockledge Drive, Bethesda MD 20892-7971 USA, TEL: 301-496-6832, FAX: 301-496-2766, Marc_H_Bornstein@nih.gov.

Authorship is alphabetical; all authors contributed equally to this paper.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

In this article, we briefly describe and illustrate four prominent strategies that answer the sampling challenge, and we evaluate each in terms of some fundamental, meaningful, and practical criteria. The four strategies include (a) population-based probability sampling as well as nonprobability sampling strategies such as (b) convenience sampling, (c) quota sampling, and (d) homogeneous sampling. The five criteria by which we appraise these sampling strategies include (a) whether they yield representative and generalizable estimates of a study's target population (e.g., estimates of intelligence among the population when all sociodemographic groups are collapsed), (b) whether they yield representative and generalizable estimates of sociodemographic group differences within a study's target population (e.g., how estimates of intelligence vary across a population's ethnic groups), (c) the recruitment efforts and costs they entail, (d) whether they provide sufficient power to detect sociodemographic group differences, and (e) whether they introduce noise related to variation in sociodemographic factors and whether that noise can be accounted for statistically. After overviewing the four sampling strategies, we examine how the sociodemographic composition of a sample in terms of gender, ethnicity, and SES can compromise a study's findings – regardless of the study goals. We then recount the use of each prominent sampling strategy in five high-profile journals in contemporary developmental science. On these bases, we arrive at conclusions and recommendations about best practices and practical considerations, including ethical issues, and discuss the importance of weighing the research question when considering the merits of various sampling strategies.

This article is not comprehensive, and we have not assumed some related burdens. By now demographers, sociologists, and others in many disciplines have weighed the pros and cons of different sampling strategies (Davis-Kean & Jager, 2011; Henry, 1990; Onwuegbuzie & Collins, 2007; Sue, 1999; Watters & Biernacki, 1989). This article does not provide a tutorial on sampling (see <http://stattrek.com/statistics/data-collection-methods.aspx?Tutorial=Stat>). We also eschew technical details in favor of highlighting “big picture” issues of design and practicality in an accessible way. Although our examples and arguments are applicable to any single sociodemographic factor or set of sociodemographic factors, here we limit our focus to gender, ethnicity, and SES. Also, although we fully recognize that gender, ethnicity, and SES are non-independent (ethnicity and SES in particular) but interact in myriad complex ways, when discussing the implications of these three sociodemographic factors we typically limit our examples to a single factor for the sake of conceptual clarity. Finally, for the purposes of this exposition about sampling, we combine “race” and “ethnicity” as used by the U.S. Government and its agencies and define six ethnic categories (see Table 1). We acknowledge that these ethnic groups are also heterogeneous in that each group contains people who originated from many different countries with different cultural practices.

Common Sampling Strategies in Developmental Science

Here we describe four of the most used sampling strategies, and we assess their advantages, disadvantages, and limitations. How each of the four sampling strategies fares on the five criteria is summarized in Table 2.

Population-Based Probability Sampling

A principal set of sampling strategies falls under the category of “population-based probability sampling.” These strategies include simple random sampling as well as more complex sampling designs such as stratified sampling and cluster sampling (and its variants such as probability proportional to size sampling). Because a detailed exposition of these strategies is beyond this scope of this paper, here we only provide basic descriptions (for more thorough reviews see Cochran, 1977, or Levy & Lameshown, 2011). In simple random sampling, a random subset (n) of the target population (N) is selected, with each member of N having an equal probability of selection. In stratified sampling, the population is divided into separate groups called “strata” (such as ethnic groups), and then a probability sample (often a simple random sample) is drawn from each stratum. With cluster sampling, the target population is divided into separate geographic groups called “clusters” (such as schools, neighborhoods, businesses), a simple random sample of clusters is selected from the population, and data collection is limited to those who fall within these randomly selected clusters. Within each selected cluster, data collection can be probability based (i.e., based on a simple random sample or a stratified design) or complete (i.e., every individual within a given cluster is eligible to participate in the study). Although these population-based probability sampling strategies differ from one another in important ways, they all, when carried out properly, yield an unbiased sample that is representative of the target population (i.e., the sociodemographic characteristics of the sample faithfully reflect the sociodemographic characteristics of the target population). For example, the National Longitudinal Study of Adolescent Health (Add Health; Bearman, Jones, & Udry, 1997), which used a clustered sampling design that applied stratified sampling within cluster (school), is a population-based probability sample of 7th-12th graders in the United States during the 1994–1995 school year. The sample has been re-interviewed three times, the most recent being in 2008, when the sample was ages 24 to 32. Because Add Health is a population-based probability sample with a clear target population, researchers and practitioners can be confident that findings from studies utilizing Add Health data generalize to the U.S. population of adolescents and young adults as a whole.

Based on our five criteria, population-based probability sampling appears to have more advantages than disadvantages. Focusing first on its advantages, in terms of representativeness and generalizability, when carried out properly, these sorts of samples yield generalizable estimates of the target population (i.e., all sociodemographic groups combined) and of sociodemographic group differences (e.g., gender, ethnic, or SES differences within a target population). Additionally, assuming the subsamples for a given sociodemographic factor are sufficiently large (say $n = 45$; based on .80 power to detect a medium effect of $f = .25$, $\alpha = .05$, in an ANOVA design with four groups¹; Faul, Erdfelder, Lang, & Buchner, 2007), this sampling strategy yields sufficient power to detect differences among sociodemographic subgroups within the target population. Finally, for researchers not interested in subgroup differences, probability samples also allow accounting for noise introduced by variation in sociodemographic factors. Taking advantage of the substantial

¹This is only one example for illustrative purposes. The sample size needed for different analyses should always be estimated using an a priori power analysis prior to selecting a sample.

sociodemographic variation in these studies, researchers can take steps to control for sociodemographic group differences, or, better, researchers can simply examine their research question separately for each sociodemographic subgroup (i.e., examine questions separately by gender using multiple-group analyses) and compare the findings.

Regarding its disadvantages, when done properly the recruitment costs and efforts for population-based probability sampling are high. Regardless of the population-based sampling strategy used, researchers need to carefully define the target population and clarify its sociodemographic composition. Depending on the target population, doing so can be straightforward (e.g., the ethnic composition of the U.S. population is tracked by the U.S. Census). However, in many cases the sociodemographic compositions of other, smaller target populations are not fully known and may require great efforts to accurately determine. In addition, when researchers sample from geographic areas smaller than their country, geographical decisions can be arbitrary. Suppose one were to sample from the geographic home of the National Institutes of Health (NIH). What would define that sampling area? The main campus of the NIH? The town of Bethesda? Montgomery County? The state of Maryland? The Middle Atlantic States? Second, population-based probability-sample sizes need to be quite large, often coming at great costs in terms of money, time, and effort. For example, consider a research group that wishes to collect ethnicity data that is representative of a given target population via simple random sampling. Unless the target population is a single ethnic group (see Homogeneous Sampling below), any target population will consist of a set of ethnic groups that are not distributed equally (i.e., some comprise a far lesser proportion of the target population than others). Therefore, to yield a subsample of each ethnic group that is statistically useful², including those that comprise a small proportion of the target population, requires collecting a substantial amount of data. For example, using the U.S. population as the target population, to yield a Hispanic n of 45 (a rather modest n for making generalizations to the entire U.S. population of Hispanics) calls for a total representative sample N of 275. To yield n s of 45 for non-Hispanic Blacks, Asians, American Indians, and Hawaiian/Other Pacific Islanders would call for total representative sample N s of 369, 959, 6,164, and 28,125, respectively. Of course the required sample size balloons even higher when multiple sociodemographic factors are considered (e.g., to yield an n of 45 for Hawaiian/Other Pacific Islander females would call for a total sample N of 56,250).

One of the benefits of clustered and stratified sampling designs is that, relative to simple random sampling, the recruitment costs and efforts are often lower (Groves, 1989; Heeringa et al., 2004), although still considerable. Within stratified designs, strata that are underrepresented within the population (e.g., the American Indian stratum among ethnicity strata or the highly affluent stratum among SES strata) can be oversampled (termed “disproportionate allocation”) reducing the overall N required to yield a statistically useful subsample of each subgroup. Sample weights, which “down weight” the oversampled strata, can then be applied to the data to yield estimates that are generalizable to the total population. Additionally, although cluster designs do not necessarily reduce the number of

²A sample large enough to yield sufficient power and to reasonably account for other confounds via random sampling.

individuals that need to be recruited, they can reduce recruitment costs. In terms of effort and time, it is far easier to sample, for example, 100 individuals within a single sampling cluster (i.e., neighborhood or school) than it is to sample 100 individuals scattered across a number of different sampling clusters. Stratified and clustered sampling designs also have their disadvantages relative to simple random sampling. Complex sampling strategies are less straightforward to implement and require the use of specialized analytical techniques to obtain accurate variance estimates (Davis-Kean & Jager, 2011).

In summary, population-based probability sampling strategies allow for clear generalizability to both the target population and its sociodemographic subpopulations, and they enable researchers to account for the noise introduced by variation in sociodemographic factors. Despite these important advantages, population-based sampling strategies are often prohibitively costly and labor-intensive.

Due to the substantial costs of population-based sampling strategies, the use of nonprobability samples, which are typically less expensive in all ways, is much more common in developmental research. Next, we review and critique three specific types of nonprobability samples: convenience sampling, quota sampling, and homogenous sampling.

Convenience Sampling

Unlike population-based sampling strategies, convenience sampling is a nonprobability sampling strategy where participants are selected based on their accessibility and/or proximity to the research. One of the most common examples of convenience sampling within developmental research is the use of student volunteers as study participants. This strategy entails recruiting a sample that has some ad hoc sociodemographic composition, which is tailored neither to the United States nor to any other identifiable target population but rather settles for whatever convenience sample the researcher recruits (presumably) on a first come-first recruited basis.

This strategy's clear advantage is that, of all the sampling strategies, convenience sampling is the easiest, least time-intensive, and least expensive to implement, perhaps accounting for its popularity in developmental research. Regarding its disadvantages, results that derive from convenience sampling have known generalizability only to the sample studied. Thus, any research question addressed by this strategy is limited to the sample itself. The same limitation holds true for estimates of differences between sociodemographic subgroups. As another disadvantage, convenience samples typically include small numbers of underrepresented sociodemographic subgroups (e.g., ethnic minorities) resulting in insufficient power to detect subgroup differences within a sociodemographic factor or factors. Moreover, although small in number, these underrepresented sociodemographic subgroups introduce modest amounts of variation into the sample, enough variation to produce statistical noise in the analyses but not enough variation to harness or control statistically. Indeed, the widespread use of convenience sampling may be partly responsible for the host of small and inconsistent effects that pervade developmental science, why sizes of effects often vary depending on the variables considered, and why research shows links between particular setting conditions and outcomes for some, but not other, groups (Bornstein, 2013).

In summary, convenience sampling is a common strategy, but its scientific disadvantages appear to outweigh its practical advantages. Relative to population-based probability sampling, convenience sampling is far easier and less expensive to implement. However, unlike population-based probability sampling, convenience sampling produces estimates that lack generalizability to any identifiable target population or subpopulations (except for the sample studied), provides insufficient power to detect differences among sociodemographic subgroups, and includes noise due to sociodemographic variation that cannot be controlled or accounted for.

Quota Sampling

Because of the well-intentioned movement to improve the representation of underrepresented groups in developmental research, there has developed a sampling strategy of recruiting fixed numbers of participants from different sociodemographic groups (e.g., sample *ns* of 45 for each ethnic group within the target population). Like convenience sampling, quota samples are typically nonprobability samples. Although quota sampling (also referred to as “equal sampling”) bears some resemblance to stratified sampling using disproportionate allocation (i.e., oversampling underrepresented groups or strata), it is distinct from stratified sampling in two important ways. First, stratified sampling using disproportionate allocation draws a probability sample from each stratum under investigation, but quota sampling typically draws a nonprobability sample from each group under investigation (i.e., a non-random sample of each ethnic group within the target population). Second, stratified sampling using disproportionate allocation is typically accompanied by the use of sample weights, which render estimates that are generalizable to the target population, but quota sampling rarely involves the calculation or application of sample weights.

Based on our five criteria, quota sampling has more disadvantages than advantages. One advantage of quota sampling is that, because it often entails oversampling of underrepresented groups, quota sampling typically provides sufficient statistical power to detect group differences. Another advantage of quota sampling is that, for researchers not interested in differences across subgroups of a sociodemographic factor, quota sampling permits accounting statistically for noise introduced by sociodemographic variation. However, because quota sampling typically draws nonprobability samples from each group under investigation, it does not yield generalizable estimates of the target population or of subgroup differences within the target population. Indeed, that sample weights are rarely applied renders estimates of the target population all the more biased. A second disadvantage of quota sampling is that, when done properly, the recruitment costs and efforts are intermediate. Although not as large as population-based samples, the size of the overall sample still needs to be fairly large. That is, the sample still needs to include a statistically useful sub-sample of each subgroup of a sociodemographic factor, which at the low end amounts to an $n = 45$ (for a 4-group ANOVA). For example, for studies focused on ethnicity, based on the six-category definition of ethnicity, a minimum total sample N of 270 would be required (i.e., a subgroup n of 45, by 6 subgroups, equals a total N of 270). Locating high numbers of those in underrepresented groups to participate in a study can be challenging.

In summary, relative to convenience samples, quota samples are equally disadvantaged when it comes to producing estimates of the target population and of subgroup differences that are generalizable and more disadvantaged with respect to recruitment costs. However, unlike convenience samples, quota sampling typically provides the statistical power necessary to detect subgroup differences among sociodemographic factor(s) of interest (i.e., the sociodemographic factor(s) whose subgroups are quota sampled) and permits accounting statistically for noise introduced by variation in the sociodemographic factor(s) of interest. Therefore, relative to convenience samples, quota samples are better suited for examining subgroup differences within a sociodemographic factor or factors.

Homogeneous Sampling

In homogenous sampling, researchers undertake to study a sociodemographically homogeneous population (e.g., the overall sample is comprised of just females). Homogenous samples can vary in their degree of sociodemographic homogeneity. For example, samples that restrict variation on multiple sociodemographic factors (e.g., a sample limited to female, European American adults) are more homogeneous than samples that restrict variation on a single sociodemographic factor (e.g., a sample limited to females that includes all ethnic groups). Like convenience and quota samples, homogenous samples are typically nonprobability samples.

Based on our five criteria, homogeneous sampling has important advantages and disadvantages. Focusing first on its advantages, the recruitment costs and efforts are generally low for homogeneous samples, although relative to homogenous samples of overrepresented sociodemographic groups (e.g. European Americans in the United States), homogeneous samples of underrepresented sociodemographic groups (e.g., Native Americans) can be more costly and time consuming. An additional advantage is, because the homogeneous sampling design eliminates all variation associated with one or more sociodemographic factors, it adds no noise associated with those sociodemographic factors to the overall results. That is, by including only one ethnic group, for example, the research intentionally avoids the noise associated with ethnic confounds that can cloud the findings if different ethnic groups are combined, thereby improving the accuracy and quality of the resultant data. Provided homogeneous samples are nonprobability samples, which in practice is typically the case, a key shortcoming is that they yield estimates that are not generalizable to the target population (e.g., a non-probability, homogenous sample of Native Americans would not yield estimates that are generalizable to the population of Native Americans). However, this limitation does not hold for probability homogeneous samples. Additionally, because the target population is defined as a particular sociodemographic group, homogeneous samples are often ill-suited to examine sociodemographic differences. For example, homogeneous samples limited to females cannot be used to examine gender differences, and homogeneous samples limited to African American females cannot be used to examine ethnic differences, gender differences, or ethnic differences within gender.

In summary, because homogeneous samples are typically nonprobability samples, they are equally disadvantaged relative to convenience samples and quota samples when it comes to producing estimates of the target population that are generalizable. Like convenience

samples, recruiting costs for homogeneous samples are typically low, and lower than the costs associated with population-based probability and quota samples. Unlike convenience samples and quota samples, homogeneous samples eliminate noise due to variation in one or more sociodemographic factors. Therefore, homogeneous samples provide key advantages over both convenience samples (i.e., noise due to variation in one or more sociodemographic factors is eliminated in homogeneous samples) and quota samples (i.e., recruitment costs are typically lower for homogeneous samples).

Why the Sociodemographic Composition of Samples Matters

Given the advantages and disadvantages of the four sampling strategies, it is important to note how sociodemographic characteristics can affect study outcomes and the interpretation of study results. Gender, ethnicity, and SES variation in many characteristics—physical and mental health (Adkins, Wang, Dupre, van den Oord, & Elder, 2009; American Public Health Association, 2004; Breslau et al., 2005; Crimmins & Saito, 2001; National Center for Health Statistics, 2011), beliefs and cognitions (Burke et al., 1992; Courtenay, McCreary, & Merighi, 2002; Diala et al., 2000), and behaviors and practices (Blum et al., 2000; Burke et al., 1992; Courtenay et al., 2002; National Center for Health Statistics 2011; Snowden & Yamada, 2005)—is ubiquitous. Beyond mean-level differences in these constructs, cross-sectional associations as well as developmental linkages among them and other salient constructs also vary by gender (Card, Stucky, Sawalani, & Little, 2008; Griffin, Botvin, Scheier, Diaz, & Miller, 2000), ethnicity (Amey, Albrecht, & Miller, 1996; Burke et al., 1992; Jager, 2011), and SES (Geoffroy et al., 2007; Sachs-Ericsson et al., 2007). These broad and entrenched sociodemographic differences in both the levels and the correlates of significant developmental characteristics—constructs, structures, functions, or processes—place demands on developmental researchers regarding the sociodemographic composition of their analytic samples (Bornstein, 2010; Davis-Kean & Jager, 2011; Henry, 1990; Onwuegbuzie & Collins, 2007; Sue, 1999; Watters & Biernacki, 1989). Moreover, the practical implications of a sample's sociodemographic composition vary depending on whether a study is (a) focused on one or more sociodemographic factors as a source of heterogeneity or (b) focused on broad developmental patterns (not focused on sources of heterogeneity of any kind).

Sampling Implications for Research Focused on Sociodemographic Factors as a Source of Heterogeneity

The roots of sociodemographic differences – including gender, ethnicity, and SES differences - in developmental characteristics are complex and likely the product of layered interactions among biological, behavioral, and sociocultural factors (Betencourt & Lopez, 1993; Courtenay et al., 2002; Crimmins & Saito, 2001; Phinney, 1996). Nonetheless, they are important to unpackage because without a scientific base of knowledge regarding human health and behavior that takes into account the sociodemographic diversity of the population, health care delivery, planning, and policy making would be compromised by inadequate information and potentially misleading generalizations (Betencourt & Lopez, 1993; Hahn & Stroup, 1994; Mays, Ponce, Washington, & Cochran, 2003). Indeed, a sizable amount of developmental research has been devoted to examining sociodemographic

differences in health and development. However, if research does not include a broad, representative sample of the sociodemographic groups under examination, then a distorted view of sociodemographic differences likely results. Additionally, if research does not include adequately large samples of each sociodemographic group under examination, then sociodemographic differences that exist in the target population may go undetected (i.e., Type II errors) due to a lack of power. Thus, a proper examination of socioeconomic differences in a given developmental characteristic demands an analytical sample whose sociodemographic composition is (a) varied enough to represent the diversity of the sociodemographic groups under examination and (b) large enough to yield sufficient power to detect differences among the sociodemographic groups under examination.

Studies whose sociodemographic compositions do not meet these criteria are problematic because (a) when considered individually it is unclear whether a study's findings generalize to the intended target population and (b) when considered collectively, the findings from these studies are difficult to synthesize. For purposes of illustration, consider the following three hypothetical studies exploring ethnic differences in depression within the United States, all of which poorly sample ethnicity (either in terms of representation, number, or both). Study A utilized a sample of 1,000 middle-aged, married adults from an affluent East Coast suburban community and found that European Americans reported higher rates of major depressive disorders than all other ethnic groups; here, the sample is large but not representative of the ethnic distribution of the United States. Study B utilized a diverse sample of 200 adults and found that all ethnic groups reported equivalent rates of major depressive disorders; here the sample is representative but not large. Study C, whose sample is neither large nor representative, utilized an impoverished, rural sample of 150 adults and found one group difference: African Americans reported higher rates of major depressive disorders than did European Americans. When considering any one of these three studies individually, can one confidently conclude that its findings regarding ethnic differences in depression are a true reflection of the United States population? Although Study A's sample is adequately large to yield sufficient power to detect group differences, its estimates of group difference may be biased due to the characteristics of the sample (i.e., affluent, married, suburban adults). Although Study B's sample is more diverse, it is also much smaller and likely insufficiently powered to detect ethnic group differences. As a result, its finding that the ethnic groups did not differ from one another could simply reflect lack of power (i.e., a Type II error). Finally, the one group difference found by Study C could be an artifact of its biased sample, and its failure to detect any other group differences could be ascribable to insufficient power on account of the small sample size. Additionally, when considered collectively, how does one go about integrating the findings from these three studies? Their findings regarding ethnic differences in depression are inconsistent; therefore, they cannot all be correct. But is one more correct (or less incorrect)? Unfortunately, determining the answers to these questions is difficult because each study's sample is deficient in representation and/or sample size. Although this example involved a set of hypothetical studies, substantial variation in the size and sociodemographic composition of samples is all too common across studies examining sociodemographic differences in a given developmental characteristic in an equivalent target population. These variations make

it difficult to determine whether inconsistencies across studies represent true population differences or instead are artifacts of differences in sample composition.

Sampling Implications for Research Not Focused on Sociodemographic Factors as a Source of Heterogeneity

Even for developmental research not focused on sociodemographic factors as a source of heterogeneity, issues pertaining to the sociodemographic composition of the sample still warrant consideration (LaVeist, 2005; Phinney, 1996; Williams, 1999). For example, much developmental research focuses on developmental patterns in a particular target population (e.g., mental health trajectories among children, romantic relationship formation among young adults, and so forth). Even though this sort of research is concerned with population norms and not sources of heterogeneity (such as gender, ethnicity, and SES), the nature of the sociodemographic composition of the sample still influences the patterns found. In studies where heterogeneity due to sociodemographic factors is not the focus, variation introduced by sociodemographic factors can be thought of as “noise” or nuisance variance that, although tangential to the topic of study, should ideally be accounted for. When left unaccounted for, this noise at the very least introduces additional variance resulting in unnecessarily inflated standard errors and increases the likelihood of Type II errors. In addition to inflating the standard errors of parameter estimates, this noise can also bias or alter parameter estimates themselves. For example, consider two studies designed to examine developmental linkages between alcohol abuse and major depressive disorders in a target population. The two studies’ samples are the same except for their gender composition; Study A’s sample is 60% male, and Study B’s sample is 60% female. Because the association between alcohol abuse and major depressive disorders is markedly higher among females (Kessler et al., 1997; Poulin, Hand, Boudreau, & Santor, 2005; Zilberman, Tavares, Blume, & el-Guebaly, 2003), the association between alcohol abuse and major depressive disorders will vary across the two studies. Specifically, the association is likely higher for Study B because its sample included a higher proportion of females. Thus, two studies focusing on population norms and examining the same question (in this case the association between alcohol abuse and major depressive disorders) could reach different conclusions due to differences in sample composition (in this case gender).

A popular option for dealing with the noise introduced by sociodemographic factors is to control for it statistically, but in many cases attempts to do so prove unsatisfactory. To be able to adequately control for a sociodemographic factor, there must be sufficient variance in it. Thus, if a study has only a small number of individuals in one or more ethnic groups, for example, then the ability of the study design to effectively control for ethnicity is reduced. Because gender, ethnicity, and (in some instances) SES are categorical variables, a common method for controlling for the noise they introduce is to use a series of dummy variables. However, this method of controlling for sociodemographic “noise” is problematic for two reasons. First, this technique only controls for sociodemographic differences in levels of variables; it does not account for sociodemographic differences in associations among variables (i.e., the technique assumes that the slope of the regression line in each sociodemographic group is the same). Thus, it only partially accounts for the noise attributed to sociodemographic differences. Second, this technique successfully controls for level

differences across sociodemographic groups; however, if dummy-coding (as opposed to effects-coding) is used, it yields findings that generalize only to the reference group. To account properly for the noise introduced by sociodemographic variation, we suggest using one of the two following options. One option is to conduct preliminary analyses testing for sociodemographic differences (in both means and covariances) and if differences are found to conduct primary analyses separately for each sociodemographic group (i.e., multiple-group analyses). Option two is to recruit a sample that contains no variation in a sociodemographic factor (i.e., a data set limited to a single SES group; see Homogeneous Sampling above).

How Sampling is Used in Developmental Science

To determine the frequency with which the four different sampling methods are used in contemporary developmental science, we surveyed five years (2007–2011) of five high-profile developmental journals. The five journals included two that generally focus on abnormal development, (*Journal of the American Academy of Child and Adolescent Psychiatry* (JAACAP) and *Development and Psychopathology*), two that generally focus on normative development (*Developmental Psychology* and *Child Development*), and one that focuses on experimental studies related to development (*Developmental Science*). As we have done previously, for purposes of illustration and the sake of simplicity we limit our focus to just one of the three sociodemographic factors – ethnicity. For all published articles that included sample descriptions, we recorded the country of data collection, the sample size, and the percentages of participants who were European American or White, Hispanic American or Latino, African American or Black, Asian American or Asian, and another ethnicity. We also noted if the study did not report ethnicity and if the study mentioned that the sample was nationally representative. If ethnicity was reported on multiple groups or in multiple studies within an article, each sample was documented separately. Meta-analyses, reviews, commentaries, and editorials were ignored unless they presented new data. Because we were only interested in documenting the sampling distributions of U.S. samples for this illustration, we excluded any non-U.S. samples from our analysis. Any article that did not report the sample distribution of ethnicity, or that did not provide enough information to evaluate the ethnic distribution (e.g., “most participants were European American”), was coded as “not reported.”

Population-based probability sampling was coded if the authors reported that the study sample was nationally representative or if the distribution of ethnicities in the sample was ordered as it is in the population of the United States (with European Americans represented in higher proportions than Hispanic Americans, Hispanic Americans higher than African Americans, and African Americans higher than Asian Americans; Table 1). This decision constitutes a generous way to code population-based probability sampling because some studies that approximate the distribution of ethnicities in the United States were not truly representative. Quota sampling was coded when the authors mentioned that a particular ethnic group was oversampled (but not part of a probability sample), or when the sample included only 2 ethnic groups, each representing 45–55% of the sample, or when the sample included only 3 ethnic groups, each representing 28–38% of the sample. Homogeneous

sampling was coded when the entire sample was comprised of a single ethnic group. Finally, convenience sampling was coded for the remainder of the samples.

Table 3 presents the categorization of samples in each of the five journals. The number of samples drawn from the United States ranged from 226 to 647 across the five journals. The range of sample sizes for each journal was wide, and the median sample sizes varied across journals, Kruskal-Wallis $\chi^2(4, N=2104) = 293.96, p < .001$. Post-hoc tests indicated that the median sample size for *Developmental Science* was smaller than all other journals, and the median sample size for *Development and Psychopathology* was larger than all other journals. Notably, depending on the journal, ethnicity was not adequately reported for one-quarter to more than two-thirds of the samples. As expected, the most prevalent type of sample in every journal was the convenience sample (78–88% of codable samples). Homogeneous sampling was the next most prevalent (5–19% of codable samples), followed by population-based probability samples (3–7% of codable samples) and quota samples (0–2% of codable samples). It could be that most samples for which ethnicity was not reported were convenience samples. If we assumed that the “not reported” samples were all convenience samples, then 88–98% (overall 91%) of samples would be convenience.

Overall, the five journals were remarkably consistent in their representation of different types of samples. *Developmental Science* (the experimental journal) had a much larger percentage of samples where ethnicity was not reported adequately, $\chi^2(8, N=2122) = 330.66, p < .001$, but otherwise the patterns were very similar across journals. We explored whether the type of journal (abnormal, normative, or experimental) was associated with the prevalence of different sampling strategies. When samples where ethnicity was not reported were excluded, the distribution of sampling strategies was similar across journal types, $\chi^2(6, N=1243) = 9.85, p = .13$ (compare percentages after the slash in Table 3).

The median sample sizes for the four sampling strategies differed, Kruskal-Wallis $\chi^2(4, N=2104) = 326.45, p < .001$. Post-hoc tests indicated that the median sample size for population-based probability sampling (Median = 2,741) was larger than the median of all other sampling strategies except quota sampling (Median = 486). The median sample size for quota sampling was larger than the median sample size of homogeneous (Median = 152), convenience (Median = 153), and “not reported” samples (Median = 52). Finally, the median sample sizes for homogeneous and convenience samples were both larger than the median sample size for “not reported” samples, but the median sample sizes for homogeneous and convenience samples did not significantly differ. These results support the proposition that population-based probability sampling requires a relatively large sample size and that there is little difference in sample size for convenience and homogeneous samples.

Implications and Recommendations for Reporting Sample Characteristics

The analysis of published articles reveals several inconsistencies and inadequacies in basic reporting of demographic characteristics of study samples in developmental science. Overall, 41.4% of articles published in five high-profile developmental science journals failed to report ethnicity at all or only reported that the sample was “predominantly White”

or “about half minority”, which is not detailed enough to draw conclusions about the representativeness of the sample to a particular population. Many studies also grouped different prominent ethnicities into an “other” category. We recommend that published studies report the percentages of their sample that fall into (at least) the four major ethnic groups (White/European American, Hispanic/Latino American, Black/African American, and Asian American), but more detailed accounting of groups and subgroups is always preferred. Of course, if the study includes an oversampling of other ethnicities or ethnic subgroups, they should also be reported. Similarly, any other sociodemographic characteristics that would help to evaluate the study sample’s representativeness or generalizability (e.g., proportions of females and males, SES, education levels, and ages of participants) should be reported in detail.

A second related recommendation is that each published study explicitly report the population to which the study generalizes. Not all studies intend to represent the entire population of the United States, but in most cases authors do not explain what population their study sample is intended to represent. For example, knowing whether a study generalizes to the entire population of 8- to 10-year-old children in the United States or only to 8- to 10-year-olds in five public schools in Maine frames the interpretation and application of the study’s findings considerably. Furthermore, a study that was designed to represent public school children in Maine should have very different sociodemographics from a study designed to represent 8- to 10-year-olds in the entire United States.

Summary and Conclusions

Here we have recounted four common sampling strategies (population-based probability sampling, convenience sampling, quota sampling, and homogenous sampling) and evaluated each by five meaningful criteria. We find that by far the most common sampling strategy (convenience sampling) is the least desirable in terms of representativeness, generalizability, and noise. Convenience samples require less practical investment in recruitment costs and efforts, but this advantage does not offset its aforementioned consequential scientific disadvantages. We find population-based probability sampling the most desirable strategy, but potentially cost and resource prohibitive for many researchers. The sample size required to adequately represent all sociodemographic groups would be large and the recruitment costs and efforts for this method are consequently considerable. Furthermore, large samples often involve a trade-off with detailed measurement. For example, a small sample may allow a researcher to investigate a topic in greater depth, such as with open-ended questions or more extensive measurement. Larger samples may only allow for more blunt instruments or require even greater allocation of resources for detailed measurement. Whether homogeneous sampling or quota sampling is a better nonprobability option depends on the research question. For research questions focused on sociodemographic factors as a source of heterogeneity (e.g., the sociodemographic factor is a focal variable of interest in the study), quota sampling is the better choice. In this case, quota sampling allows the researcher to recruit sociodemographic groups (e.g., ethnic or SES groups) that are sizeable enough to address the research questions. For research questions not focused on sociodemographic factors as a source of heterogeneity (e.g., the sociodemographic factor is not a focal variable and would likely be used as a covariate), homogeneous samples are

preferable because they eliminate sociodemographic noise in the sample that could cloud the study results. Collecting a homogeneous sample might marginally increase recruitment costs and efforts compared to convenience samples (especially if the targeted group is infrequently occurring in the population), but pays off in terms of reducing noise.

One concern many researchers have about homogeneous samples is the success in obtaining protocol approval by internal review boards or to secure funding. Many funding agencies require (or strongly recommend) inclusion of all major sociodemographic groups. For example, the NIH released guidelines about including women and minorities in clinical research in 1994 (revised in 2001; Hohmann & Parron, 1996; NIH Office of Extramural Research, 2001) which indicate that all grant applications are evaluated for the inclusion of sociodemographic groups, and if groups are omitted, a strong justification is required. As we have outlined, there are circumstances when a homogenous sample is the best available option (e.g., probability samples are too expensive and sociodemographic differences are of little interest). Including sociodemographic variation that cannot be properly addressed statistically (as in a small convenience sample) should not be considered preferable to a homogeneous sample. Researchers are encouraged to make principled theoretical and statistical arguments to support their choice of a better sampling strategy, even if that strategy excludes one or more sociodemographic groups.

Sampling is a necessary evil of developmental science. It is practically and feasibly impossible to conduct a true population study, and so scientific researchers are forced to resort to sampling. Insofar as scientists must sample, they are confronted with the question of defining the profile of their sample. There are different types of samples, and with respect to different scientific questions different sampling strategies have different advantages and disadvantages. Although they are clearly preferable in terms of their generalizability and power, probability samples are infrequently used because of the cost and effort required to plan and execute them. Instead, nonprobability samples are much more common, with faulty convenience samples making up the largest share. However, with only a small amount of extra effort and planning, researchers could recruit a quota or homogeneous sample, thereby improving the characteristics of their sample and study.

Acknowledgments

We thank A. Bradley, O. M. Haynes, P. Horn, A. Mahler, C. Padilla, and C. Yuen. Supported by the Intramural Research Program of the NIH, NICHD.

References

- Adkins DE, Wang V, Dupre ME, van den Oord E, Elder GH. Structure and stress: Trajectories of depressive symptoms across adolescence and young adulthood. *Social Forces*. 2009; 88:31–60.10.1353/sof.0.0238
- American Public Health Association. *Eliminating Health Disparities: Toolkit*. Washington, DC: American Public Health Association; 2004.
- Amey CH, Albrecht SL, Miller MK. Racial differences in adolescent drug use: The impact of religion. *Substance use and Misuse*. 1996; 31:1311–1332.10.3109/10826089609063979 [PubMed: 8879076]
- Bearman, P.; Jones, J.; Udry, J. *The National Longitudinal Study of Adolescent Health: Research design*. Chapel Hill, NC: Carolina Population Center; 1997. Available at www.cpc.unc.edu/projects/addhealth/design.html

- Bettencourt H, Lopez SR. The study of culture, ethnicity, and race in American Psychology. *American Psychologist*. 1993; 48:629–637.10.1037/0003-066X.48.6.629
- Blum RW, Beuhring T, Shew ML, Bearinger LH, Sieving RE, Resnick MD. The effects of race/ethnicity, income, and family structure on adolescent risk behaviors. *American Journal of Public Health*. 2000; 90:1879–1884.10.2105/AJPH.90.12.1879 [PubMed: 11111260]
- Bornstein, MH., editor. *The handbook of cultural developmental science*. New York, NY: Psychology Press; 2010.
- Bornstein, MH. Eunice Kennedy Shriver. National Institute of Child Health and Human Development; 2013. The specificity principle in parenting and child development. Unpublished manuscript
- Breslau J, Aguilar-Gaxiola S, Kendler K, Su M, Williams D, Kessler R. Specifying race-ethnic differences in risk for psychiatric disorder in a USA national sample. *Psychological Medicine*. 2005; 35:1–12.10.1017/S0033291705006161
- Burke GL, Savage PJ, Manolio TA, Sprafka JM, Wagenknecht LE, Sidney S, Jacobs DR. Correlates of obesity in young Black and White women: The CARDIA study. *American Journal of Public Health*. 1992; 82:1621–1625.10.2105/AJPH.82.12.1621 [PubMed: 1456336]
- Card NA, Stucky BD, Sawalani GM, Little TD. Direct and indirect aggression during childhood and adolescence: A meta-analytic review of gender differences, intercorrelations, and relations among adjustment. *Child Development*. 2008; 79:1185–1229.10.1111/j.1467-8624.2008.01184.x [PubMed: 18826521]
- Cochran, WG. *Sampling Techniques*. 3. New York: Wiley; 1977.
- Courtenay WH, McCreary DR, Merighi JR. Gender and ethnic differences in health beliefs and behaviors. *Journal of Health Psychology*. 2002; 7:219–231.10.1177/1359105302007003216 [PubMed: 22114246]
- Crimmins EM, Saito Y. Trends in health life expectancy in the United States, 1970–1990: Gender, racial, and educational differences. *Social Science and Medicine*. 2001; 52:1629–1641. [PubMed: 11327137]
- Davis-Kean, P.; Jager, J. The use of large-scale data sets for the study of developmental science. In: Laursen, B.; Little, TD.; Card, NA., editors. *Handbook of Developmental Research Methods*. New York: Guilford Press; 2011. p. 148-162.
- Diala C, Muntaner C, Walrath C, Nickerson KJ, LaVeist TA, Leaf PJ. Racial differences in attitudes towards professional mental health care and the use of services. *American Journal of Orthopsychiatry*. 2000; 70:455–464.10.1037/h0087736 [PubMed: 11086524]
- Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods*. 2007; 39:175–191. [PubMed: 17695343]
- Geoffroy MC, Cote SM, Borge AIH, Larouche F, Sequin JR, Rutter M. Association between nonmaternal care in the first year of life and children's receptive language skills prior to school entry: The moderating role of socioeconomic status. *Journal of Child Psychology and Psychiatry*. 2007; 48:490–497.10.1111/j.1469-7610.2006.01704.x [PubMed: 17501730]
- Griffin KW, Botvin GJ, Scheier LM, Diaz T, Miller NL. Parenting practices as predictors of substance use, delinquency, and aggression among urban minority youth: Moderating effects of family structure and gender. *Psychology of Addictive Behaviors*. 2000; 14:174–184.10.1037/0893-164X.14.2.174 [PubMed: 10860116]
- Groves, RM. *Survey Errors and Survey Costs*. New York, NY: Wiley; 1989.
- Hahn RA, Stroup DF. Race and ethnicity in public health surveillance: Criteria for the scientific use of social categories. *Public Health Reports*. 1994; 109:7–15. [PubMed: 8303018]
- Heeringa SG, Wagner J, Torres M, Duan N, Adams T, Berglund P. Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES). *International Journal of Methods in Psychiatric Research*. 2004; 13:221–240.10.1002/mpr.179 [PubMed: 15719530]
- Henry, GT. *Applied Social Research Methods Series*. Vol. 21. Newbury Park, CA: Sage; 1990. Practical Sampling.
- Hohmann AA, Parron DL. How the NIH guidelines on inclusion of women and minorities apply: Efficacy Trials, effectiveness trials, and validity. *Journal of Consulting and Clinical Psychology*. 1996; 64:851–855.10.1037/0022-006X.64.5.851 [PubMed: 8916610]

- Humes, KR.; Jones, NA.; Ramirez, RR. Overview of race and Hispanic origin: 2010. US Department of Commerce, Economics and Statistics Administration, US Census Bureau; 2011.
- Jager J. A developmental shift in Black-White differences in depressive affect across adolescence and early adulthood: The influence of early adult social roles and socioeconomic status. *International Journal of Behavioral Development*. 2011; 35:457–469.10.1177/0165025411417504 [PubMed: 22282639]
- Kessler RC, Crum RM, Warner LA, Nelson CB, Schulenberg J, Anthony JC. Lifetime co-occurrence of DSM-III-R alcohol abuse and dependence with other psychiatric disorders in the National Comorbidity Survey. *Archives of General Psychiatry*. 1997; 54:313–331.10.1001/archpsyc.1997.01830160031005 [PubMed: 9107147]
- LaVeist TA. Disentangling race and socioeconomic status: A key to understanding health inequalities. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*. 2005; 82(Suppl 3):26–34.10.1093/jurban/jti061
- Levy, PS.; Lemeshow, S. *Sampling of Populations: Methods and Applications*. 4. New York: Wiley; 2011.
- Mays VM, Ponce NA, Washington DL, Cochran SD. Classification of race and ethnicity: Implications for Public Health. *Annual Review of Public Health*. 2003; 24:83–110.10.1146/annurev.publhealth.24.100901.140927
- National Center for Health Statistics. *Health, United States, 2010: With Special Feature on Death and Dying*. Hyattsville, MD: National Center for Health Statistics; 2011.
- National Institutes of Health. Office of Extramural Research. NIH Policy and Guidelines on The Inclusion of Women and Minorities as Subjects in Clinical Research – Amended. 2001 Oct. Available online: http://grants.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm
- Onwuegbuzie AJ, Collins KMT. A typology of mixed methods sampling designs in social science research. *The Qualitative Report*. 2007; 12:281–316.
- Phinney JS. When we talk about American ethnic groups, what do we mean? *American Psychologist*. 1996; 51:918–927.10.1037/0003-066X.51.9.918
- Poulin C, Hand D, Boudreau B, Santor D. Gender differences in the association between substance use and elevated depressive symptoms in a general adolescent population. *Addiction*. 2005; 100:525–535.10.1111/j.1360-0443.2005.01033.x [PubMed: 15784067]
- Sachs-Ericsson N, Burns AB, Gordon KH, Eckel LA, Wonderlich SA, Crosby RD, Blazer DG. Body mass index and depressive symptoms in older adults: The moderating roles of race, sex, and socioeconomic status. *American Journal of Geriatric Psychiatry*. 2007; 15:815–825.10.1097/JGP.0b013e3180a725d6 [PubMed: 17804833]
- Snowden LR, Yamada A. Cultural differences in access to care. *Annual Review of Clinical Psychology*. 2005; 1:143–166.10.1146/annurev.clinpsy.1.102803.143846
- Sue S. Science, ethnicity, and bias: Where have we gone wrong? *American Psychologist*. 1999; 54:1070–1077.10.1037/0003-066X.54.12.1070 [PubMed: 15332528]
- Watters JK, Biernacki P. Targeted sampling: Options for the study of hidden populations. *Social Problems*. 1989; 36:416–430.10.1525/sp.1989.36.4.03a00070
- Williams DR. Race, socioeconomic status, and health: The added effects of racism and discrimination. *Annals of the New York Academy of Sciences*. 1999; 896:173–188.10.1111/j.1749-6632.1999.tb08114.x [PubMed: 10681897]
- Zilberman NL, Tavares H, Blume SB, el-Guebaly N. Substance use disorders: Sex differences and psychiatric comorbidities. *The Canadian Journal of Psychiatry*. 2003; 48:5–13.

Sampling is a key feature of every study in developmental science.

We evaluate 4 prominent sampling strategies in developmental science.

We judge these sampling strategies by criteria, such as representativeness and generalizability.

We tally the use of the four sampling strategies in five prominent developmental science journals.

Finally, we make recommendations about best practices for sample selection and reporting.

Table 1

Ethnicity distribution in the United States in 2010

Ethnicity	Percentage
White ^a	63.75%
Hispanic	16.35%
Black ^a	12.21%
Asian ^a	4.69%
American Indian/Alaskan Native ^a	0.73%
Hawaiian/Other Pacific Islander ^a	0.16%

Note. Adapted from Table 1 in Humes, Jones, and Ramirez (2011).

^aNonhispanic.

Table 2

Evaluation of sampling strategies based on five criteria

	Population-based ¹	Convenience	Quota	Homogeneous
Generalizable estimates of target population	Yes	No	No	No
Generalizable estimates of sociodemographic differences	Yes	No	No	No
Ease of recruitment	Low	High	Intermediate	High
Power to detect sociodemographic group differences	Yes	No	Yes	No
Absence of sociodemographic “noise” in results	No ²	No	No ²	Yes

¹ Includes simple random sampling, stratified sampling, and/or clustered sampling.

² Although the effects of the “noise” can be mitigated statistically.

Table 3

Characteristics of U.S. samples that fell into each sampling category from 2007 to 2011 in five high-profile developmental journals

Journal	5-year impact factor	N samples ^a	Median sample size	Range of sample sizes	Percentage of samples drawn using each sampling strategy			
					Population-based ^b	Convenience ^b	Quota ^b	Homogeneous ^b Not reported ^c
JAACAP	5.976	443	126	10–16,128,828	3.8/7.2	45.6/85.6	1.1/2.1	2.7/5.1 46.7
D&P	6.688	226	220	20–390,350	3.1/4.2	65.5/88.1	0.9/1.2	4.9/6.5 25.7
DP	5.123	647	121	8–67,124	3.9/5.5	58.1/82.8	1.4/2.0	6.8/9.7 29.8
CD	5.700	517	101	6–21,255	3.5/5.1	57.8/84.7	1.7/2.5	5.2/7.6 31.1
DS	4.597	289	36	10–10,200	0.3/3.1	8.7/78.1	0.0/0.0	2.1/18.8 88.9
Total	--	2122	105	6–16,128,828	3.2/5.5	49.5/84.5	1.2/2.0	4.7/8.0 41.4

Note. The 5-year Impact Factor was abstracted from the 2011 Thomson Reuters Journal Citation Reports Database. JAACAP = *Journal of the American Academy of Child and Adolescent Psychiatry*; D&P = *Development and Psychopathology*; DP = *Developmental Psychology*; CD = *Child Development*; DS = *Developmental Science*.

^aSamples drawn from the United States only.

^bPercentages left of the slash (/) include “not reported” in the total; percentages right of the slash exclude samples for which ethnicity was not reported.

^c“Not reported” was coded when no ethnic information was reported about a sample, or when only gross categories that did not provide enough information to code (e.g., white vs. non-white) were reported.