Published in final edited form as:
Multivariate Behav Res. 2009 November 1; 44(6): 711–740. doi:10.1080/00273170903333574.

Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration

Sonya K. Sterba
University of North Carolina at Chapel Hill

Abstract

A model-based framework, due originally to R. A. Fisher, and a design-based framework, due originally to J. Neyman, offer alternative mechanisms for inference from samples to populations. We show how these frameworks can utilize different types of samples (nonrandom or random vs. only random) and allow different kinds of inference (descriptive vs. analytic) to different kinds of populations (finite vs. infinite). We describe the extent of each framework's implementation in observational psychology research. After clarifying some important limitations of each framework, we describe how these limitations are overcome by a newer hybrid model/design-based inferential framework. This hybrid framework allows both kinds of inference to both kinds of populations, given a random sample. We illustrate implementation of the hybrid framework using the High School and Beyond data set.

Nonrandom sampling involves selecting units (e.g., persons) with unknown probabilities of selection from a finite population of units. This finite population may be poorly defined (e.g., all persons who saw a flier posted on a community bulletin board) or well defined (e.g., all children attending licensed daycare centers in Dayton, Ohio). Nonrandom, or purposive, sampling is common in observational psychology research. For example, 76% of observational studies in 2006 issues of Journal of Personality and Social Psychology, Developmental Psychology, Journal of Abnormal Psychology, and *Journal of Educational Psychology* used nonrandom samples (Sterba, Prinstein, & Nock, 2008). Psychologists often raise the following two questions about nonrandom samples in observational research (e.g., Jaffe, 2005; Peterson, 2001; Sears, 1986; Serlin, Wampold, & Levin, 2003; Sherman, Buddie, Dragan, End, & Finney, 1999; Siemer & Joorman, 2003; Wintre, North, & Sugar, 2001).

1. Can statistical inferences be made from nonrandom samples; if so, under what conditions and to what population?
2. Do inferences made from nonrandom samples differ from those possible under random sampling?

According to some psychology research methods texts, the answer to the first question is no: “Although these purposive sampling methods are more practical than formal probability sampling, they are not backed by a statistical logic that justifies formal generalizations” (Shadish, Cook, & Campbell, 2002, pp. 24, 356; see also Cook & Campbell, 1979, pp. 72–73).[1] However, according to other psychology research methods texts, formal statistical

Correspondence concerning this article should be addressed to Sonya K. Sterba, Department of Psychology, University of North Carolina at Chapel Hill, Davie Hall, CB# Box 3270, Chapel Hill, NC 27599-3270. ssterba@email.unc.edu.

inferences from nonrandom samples are possible under certain conditions (e.g., Cronbach, 1982, pp. 255, 158–166).

Again consulting psychology research methods texts, the answer to the second question remains unclear. For example, Shadish et al. (2002) note that *randomly* selecting units—that is, sampling units with known probabilities of selection from a well-defined finite population of units—facilitates generalization from those sample units to the finite population by ensuring a “match between sample and population distributions on measured and unmeasured attributes within known limits of sampling error” (p. 343; see also Cook & Campbell, 1979, p. 75).[2] But specifics are not provided as to whether the known probabilities of selection actually feature in statistical inference. We also are not told whether different methods of analysis need or can be used with random samples versus nonrandom samples in order to achieve such inference.

The first goal of this article is to comprehensively address these two questions by describing alternative inferential frameworks for nonrandom and random samples. To answer the first question, we present a *model-based* statistical framework, due originally to Ronald Fisher, for inference from nonrandom or random samples to what we will term *infinite populations*. We make explicit the statistical logic that allows formal generalization under this framework, and we describe the extent of this framework's implementation in psychology. To answer the second question, we then present a *design-based*[3] statistical framework, due originally to Jerzy Neyman, for inference from random samples only to what we will term *finite populations*. We make explicit that different methods of analysis and different kinds of inferences are available exclusively under random sampling, and we describe the extent of this framework's implementation in psychology. However, we then show that each framework has a set of important limitations. The second goal of this article is to explain how these limitations can be overcome using a newly developed hybrid model/design-based framework. The hybrid framework allows inference from random samples to finite or infinite populations and offers some unique strengths. We demonstrate its strengths by showing that it can correct potential limitations of an often-cited High School and Beyond study analysis (Raudenbush & Bryk, 2002;Singer, 1998).

BACKGROUND: BEFORE SAMPLING

To orient ourselves, consider that no statistical framework for inference from a sample to a population was available until the early 20th century. Before that point, social, health, and economic data on a state or country was generally gathered via complete enumeration (Stephan, 1948). However, desire to obtain estimates at lower cost eventually prompted consideration of sampling. Kiaer (1895) suggested nonrandom sampling whereas Bowley (1906) suggested random sampling. Initially, both sampling methods were distrusted for lacking a viable statistical framework for inference from the sample to a population. But instead of one inferential framework, two were proposed. Philosophical differences between

[1]Cook and Campbell (1979, pp. 72–73) indicate that nonrandom sampling precludes statistical inference (which they term strict generalizing) from samples to populations. They further state that generalizing from a sample to a population logically presupposes generalizing across subgroups within a population (e.g., boys vs. girls). Nevertheless, because of the rarity of random samples, they state that they will deemphasize the first step (generalizing to a population) to focus on the second step (generalizing across subpopulations).

[2]For this reason, Shadish et al. (2002, pp. 472–473) state that randomly selecting units facilitates *external validity*. Moreover, Shadish et al. (2002, pp. 55–56) also imply that random selection would facilitate *internal validity* by decreasing risk of *selection bias* (defined later).

[3]Note that the term “design-based inference” is not used here in the familiar research design sense. That is, by design-based, we are not referring to using research designs (e.g., regression discontinuity design) to aid causal inference and minimize validity threats. As discussed later, we are referring to using random selection probabilities (i.e., sampling design) as the sole basis for analysis and inference once the data are collected.

prominent statisticians (Fisher vs. Neyman/Pearson) regarding the definition of a population and the role of models in data analysis resulted in these two alternative frameworks (Lenhard, 2006). Fisher's (1922) inferential framework came to be called the *model-based* framework. Neyman's (1934) inferential framework came to be called the *design-based* framework. In what follows, we deemphasize the rhetoric of Fisher and Neyman's often-heated disagreements (they clashed on many other topics as well, including statistical hypothesis testing and confidence intervals; Dawid, 1991; Fienberg & Tanur, 1987). We focus instead on their frameworks' requirements and logic.[4]

MODEL-BASED INFERENTIAL FRAMEWORK

Fisher's perspective was that empirical random sampling would not always be feasible, particularly for observational studies in sociology and economics (e.g., Fisher, 1958, p. 264). Fisher also held that statistical modeling should play a central role in data analysis; that is, model building and modification should mediate between real-world problems and mathematical testing with the data at hand (Fisher, 1955, pp. 69–71; Lenhard, 2006). Hence, Fisher developed an inferential framework that relied on modeling—particularly, distributional assumptions—to mimic random sampling, even when empirical random sampling was absent.

Fisher's *model-based* framework acknowledged at the outset that nonrandom sampling indeed affords no statistical basis for generalizing from sample statistics to parameters of a particular *finite* population. Here, a *finite population* is defined as all units which had a nonzero probability of selection into that particular sample (see Figure 1). However, although nonrandom sampling does not permit finite population inference, Fisher (1922) showed that a different type of inference was possible using nonrandom sampling—infinite population inference. To implement Fisher's framework for infinite population inference, three prerequisite steps were necessary, as follows (see also Cronbach, 1982, chaps. 5–6). As we discuss later, psychology researchers often implement the first two required steps but partially or fully neglect the third.

Step 1

As a first step, a statistical model needs to be formulated by the researcher (Fisher, 1922, pp. 311–312). A statistical model describes how the dependent variable(s) are thought to have been generated. An example statistical model is a simple linear regression model which posits that the dependent variable y_i is generated as a function of a known, fixed independent variable (x_i) and error:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (1)$$

All possible y-values that could be generated by the model make up a *hypothetical, infinite population* (Fisher, 1922, pp. 311–312). The targets of inference under Fisher's framework are the model parameters (e.g., regression coefficients, β), which characterize this hypothetical, infinite population. The purpose of the statistical model is to provide a link between the observed units in the sample and the unobserved units in the infinite population, enabling causal or analytic inferences to pertain to these unobserved units as well (Royall,

[4]Fisher and Neyman had some agreements as well, including the importance of random assignment to treatment in experiments (Fisher, 1925, 1935; Neyman, 1923). However, Fisher kept his work on experimentation (not discussed here) and inference (discussed here) markedly separate, such that he ironically advocated at times for minimizing scope of modeling in experimental data analysis (Kempthorne, 1976).

1988; see Figure 1). In Cronbach's (1982) terms, "The model is used to reach from *u* [sampled units …] to *U* [population units]" (p. 161; see also p. 163).[5]

Step 2

Naturally, we need to be able to consider our observed *y*-values as realizations of a random variable in order to be able to explain their variability. But this leap is not automatic. Indeed, at this point, we have established no grounds under which to consider observed *y*-values as anything but fixed quantities pertaining to sampled units. One rationale for considering *y*-values as realizations of a random variable would be if each *y*-value was associated with a known, nonzero probability of selection. However, this rationale is not available to us because such selection probabilities are unknown under nonrandom sampling. Instead, as a second step, a parametric distributional assumption needs to be imposed on the model in order to convert the fixed *y*-values obtained for the sampled units into realizations of a random variable *y* (Fisher, 1922, p. 313). An example of a parametric distributional assumption is the requirement $\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$, that is, that the errors in our regression model are independently and identically distributed with mean 0 and variance σ^2. This assumption serves to convert the error term into a *random* variable that, in turn, converts the dependent variable *y* into a random variable (Neter, Kutner, Nachtsheim, & Wasserman, 1996, Theorem A.40). Hence, by imposing a parametric distributional assumption, we render our *y*-values *epistemically random* (Johnstone, 1989) under our model—*regardless* of whether or not empirical random sampling was actually used at the data collection stage. Because random variation in the observed outcome *y* is introduced by model assumption, not by design, data analysis under Fisher's framework lacks a *formal* requirement of empirical random sampling (see Johnstone, 1987, 1989). Instead, Fisher's framework requires only that the distributional assumption(s) imposed by the model be reasonable in light of the sample selection mechanism actually used (see also Cronbach, 1982, pp. 164–165). For example, by invoking the *iid* assumption, we claim that our distribution of *y*-values does not differ meaningfully from the distribution that would have been generated by empirical simple random sampling. This is because an independent and identical distribution of *y*-values is actually the same distribution as would be obtained if empirical simple random sampling were repeatedly performed (Kish, 1996).

Step 3

Fisher recognized that under certain circumstances, a researcher's sample selection mechanism *would* meaningfully depart from that which would have been generated by empirical simple random sampling. Under these circumstances, the sample selection mechanism could not be ignored during data analysis, and Fisher's framework required a third and final step (Fisher, 1956, pp. 33–34, 36). These circumstances have since been made explicit (see Skinner, Holt, & Smith, 1989; Smith, 1983a).[6] The first circumstance occurs when sampling units in the finite population were *stratified* (divided into nonoverlapping categories such as employed/unemployed, inpatient/outpatient, rural/urban) before being independently selected from each stratum. The second circumstance occurs when sampling units in the finite population were *clustered* (aggregated into groups, such as schools, classrooms, households, mouse litters) before multiple units from the same group were selected or before whole intact groups were selected. The third circumstance occurs when the probability of selecting sampling units from the finite population was *disproportionate*—such that probabilities of selection were related to the outcome variable

---

[5]Cronbach (1982) is not only concerned with inferences from sample to population *u*nits (e.g., persons) but also generalizations from sample to population *o*bservations, *s*ettings, and *t*reatments—together called *utos*. We focus on *u* in this article.

[6]Cronbach (1982, chap. 6) discusses this step in the specific context of treatment-outcome designs, which is not our focus here.

even after controlling for independent variables. Next, we illustrate the consequences of ignoring each stratification, clustering, and disproportionate selection via case examples. We describe what Step 3 of the model-based framework would require in each case. For each case example, the Equation (1) model is true in the infinite population with parameters β_0 = 2, β_1 = 3, and σ^2 = 1.

First, consider the case in which units were selected from each of several strata (where strata variables are typically assumed to neither interact with nor correlate with independent variables). Ignoring this fact during data analysis typically results in standard errors that are too large (Kish & Frankel, 1974). For example, standard errors are inflated by 49% for β_0 and 46% for β_1 if there are four strata and the stratification variable correlates with the outcome at r = .50.[7] For this reason, Step 3 of Fisher's framework requires the researcher to condition his or her model on any strata indicators so that, after conditioning, the infinite population is “subjectively homogeneous and without recognizable stratification” (Fisher, 1956, p. 33). This third step is often called Fisher's *conditionality principle* (see Johnstone, 1987; Lehmann, 1993). According to Fisher, after conditioning, nothing should distinguish the observed set of *n y*-values from any other set of *n y*-values that could have been generated by the model for the hypothetical, infinite population (Fisher, 1955, p. 72; 1956, pp. 55–57). In the case of stratification, conditioning would amount to expanding the Equation (1) model to include strata indicators as fixed effects (see Skinner et al., 1989). In Equation (2) dummy variables are included for three of the four strata:[8]

$$y_i=\beta_0+\beta_1 x_i+\beta_2 strat1_i+\beta_3 strat2_i+\beta_4 strat3_i+\varepsilon_i. \tag{2}$$

As long as this parametric model is properly specified, the conditional (model-based) variance of $\widehat{\beta}_1$ across repeated samples that could be generated by the model can then be used to make inferences about the target parameter β_1 in the infinite population.

Next, consider the case in which clusters, rather than individual units, are selected. Ignoring this fact during data analysis typically results in standard errors that are too small (Kish & Frankel, 1974). For example, standard errors are shrunken by 55% for β_0 and 57% for β_1 if the residual correlation among units within cluster is .15 and only β_0 varies across clusters.[9] In the case of clustering, fulfilling Fisher's *conditionality principle* could amount to expanding the Equation (1) model to include cluster indicators as random effects (Raudenbush & Bryk, 2002). That is, we could allow β_0 and β_1 to vary across clusters using the multilevel modeling specification in Equation (3):

$$\begin{aligned} y_{ij} &= \beta_{0j}+\beta_{1j}x_{ij}+\varepsilon_{ij} \\ \beta_{0j} &= \gamma_{00}+u_{0j} \quad \text{where} \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right). \\ \beta_{1j} &= \gamma_{10}+u_{1j} \end{aligned} \tag{3}$$

For example, in Equation (3), β_{1j} varies across clusters with mean γ_{10}, group-specific deviation from this mean u_{1j}, and variance of these group-specific deviations τ_{11}. As long as

[7]The magnitude of *SE* inflation was estimated using Asparouhov's (2004) procedure. The sample selection mechanism was implemented 500 times. The Equation (1) model, which ignores stratification, was fit to each selected sample. Then, the empirical *SD* was divided by the average analytical *SE*.

[8]It is important to note that if stratification variables *did* interact and correlate appreciably with independent variables, additional product terms would need to be added to the model to account for this.

[9]Estimates of *SE* deflation were obtained by replicating this sample selection mechanism 500 times, analyzing each sample ignoring clustering, and then dividing the average analytical *SE* by the empirical *SD*.

this parametric model is properly specified, the conditional (model-based) variance of $\widehat{\gamma}_{10}$ across sets of samples that could be generated by the model can then be used to make inferences about our target parameter γ_{10} in the infinite population.

Finally, consider the case in which units are disproportionately selected, but no stratification or clustering occurred. Ignoring this fact during data analysis would have different consequences for β_0 and β_1 estimates depending on precisely how the selection variables relate to the outcome after conditioning on independent variables (e.g., Berk, 1983; Graubard & Korn, 1996; Skinner et al., 1989; Sugden & Smith, 1984). To see this, consult Figure 2. Line A in Figure 2 is the true population-generating regression line from Equation (1)— before sample selection. The other plotted lines in Figure 2 show the effects of several different selection mechanisms. (Plotted lines are calculated by averaging postselection regression coefficient estimates from 500 repetitions of each selection mechanism).[10] Line B shows that selecting on a design variable z_i—a variable used in recruitment that is not of substantive interest in the investigation—when z_i correlates with independent variable results in intercept and slope bias (here, $r_{x,z}$ = .50). Line C shows that selecting on a design variable z_i that is uncorrelated with independent variable x_i and does not interact with x_i results only in intercept bias. Line D shows that selecting on a design variable z_i that is uncorrelated with independent variable x_i and does interact with x_i results only in slope bias. Line E shows that selecting units directly on the outcome y_i results in both intercept and slope bias. The bias evident in Lines B–D is often termed *omitted variable bias* and the bias evident in Line E is often termed *selection bias*. Line B–E scenarios are often said to threaten *external validity* because statements made about *x*–*y* relations in the whole population on the basis of Lines B–E will be incorrect. Line B–E scenarios are also said to threaten *internal validity* (Berk, 1983) because statements made about *x*–*y* relations within the selected subpopulation only (e.g., persons with *y*-scores ≥ 2 in Scenario E) will also be incorrect. Line F shows that *only* when we select on an independent variable x_i do we end up with no intercept bias and no slope bias. So Line F selection could be considered “conditionally proportionate.” The Line F results suggest how the disproportionate selection scenarios depicted in Lines B–D could be accounted for in data analysis in order to fulfill Fisher's *conditionality principle*. For scenarios B–D, we can simply expand Equation (1) to include the measured selection variable z_i, and possibly its interaction term with independent variables (xz_i), as covariates:[11]

$$y_i=\beta_0+\beta_1 x_i+\beta_2 z_i+\beta_3 xz_i+\varepsilon_i. \tag{4}$$

On the other hand, in the selection scenario depicted in Line E, Fisher's *conditionality principle* would require expanding Equation (1) to account for the fact that the dependent variable is observed only when a selection threshold *t* is exceeded:

$$\begin{aligned} y_i^* &= \beta_0+\beta_1 x_i+\varepsilon_i \\ y_i &= \begin{cases} y_i^* & \text{if} \quad y_i^*>t \\ \text{missing} & \text{if} \quad y_i^* \le t \end{cases} . \end{aligned} \tag{5}$$

This entails a truncated regression model (see, e.g., SAS Proc QLIM for implementation; SAS Institute Inc., 2004).[12]

[10] In Lines B–F, cases were included if their scores on the designated selection variable were ≥ the mean.

[11] Another literature, stemming from Pearson (1903), suggests algebraic adjustments for selection on *z* that do not involve including *z* as a covariate. These usually require restrictive assumptions (e.g., homoscedasticity, linearity) and have been demonstrated for very simple models.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Of course, these scenarios depicted in Figure 2 are simplistic because they depict disproportionate selection when *no* clustering or stratification occurred in the study design. In practice, there often may be disproportionate selection of clusters, not of individuals, and/or there often may be disproportionate selection at more than one stage of recruitment. Such complexities correspond to many additional possible patterns of parameter bias from disproportionate selection beyond those shown in Figure 2 (Grilli & Rampichini, 2003). To illustrate, consider the case in which disproportionate selection occurred at Level 2 (cluster-level) or Level 1 (individual-level), or both, and a researcher knew to account for clustering using a multilevel specification, but the researcher ignored the disproportionate selection. Suppose further that the researcher is substantively interested in the effect of a Level 1 predictor x_{ij} on the outcome y_{ij} (i.e., γ_{10} in Equation (6)) and in the effect of a Level 2 predictor w_j on the outcome y_{ij} (i.e., γ_{01} in Equation (6)). Hence, the researcher specifies the following model:

$$\begin{aligned} y_{ij} &= \beta_{0j}+\beta_{1j}x_{ij}+\varepsilon_{ij} \\ \beta_{0j} &= \gamma_{00}+\gamma_{01}w_j+u_{0j} \qquad \text{where} \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right). \\ \beta_{1j} &= \gamma_{10}+u_{1j} \end{aligned} \tag{6}$$

Table 1 shows that if such a model is estimated ignoring disproportionate selection occurring at Level 2 (cluster-level) or Level 1 (individual-level), or both, there will be fixed effect bias as well as random effect variance bias. Expanding this model to include (a) cluster-level selection variables as Level 2 covariates, (b) individual-level selection variables as Level 1 covariates, and possibly (c) interaction terms between these covariates and other independent variables in the model would serve to account for such multilevel disproportionate selection—thus fulfilling Fisher's *conditionality principle*. Or, employing multilevel truncated regression would serve to account for disproportionate selection on *y* (see, e.g., aML program for implementation; Lillard & Panis, 2000).

In sum, this section has provided an answer to the first question posed earlier: “Can statistical inferences be made from nonrandom samples; if so, under what conditions and to what population?” Statistical inferences *can* be made from nonrandom samples to infinite populations under a model-based framework—if Fisher's requirements described in Steps 1–3 are met.

IMPLEMENTATION OF THE MODEL-BASED FRAMEWORK IN PSYCHOLOGY

Although the models themselves that Fisher proposed (e.g., analysis of variance [ANOVA]) were widely adopted in psychology, the conditionality principle from the third step of his inferential framework was not. To see this, we revisit the observational studies that used nonrandom samples in Sterba et al.'s (2008) review. Twenty-eight percent of these nonrandom samples reported one or more complex sampling features (stratification, clustering,[13] or disproportionate selection) but accounted for all of them in statistical models (thus fulfilling Fisher's conditionality principle). Fifty-eight percent of these nonrandom samples reported one or more complex sampling features but did not account for all of them in statistical modeling (thus violating Fisher's conditionality principle). The remaining 14% of these nonrandom samples reported no complex sampling features—either because none

[12]In truncated regression, independent and dependent variables are unobserved when a unit is not selected. In another variant, censored regression, only dependent variables are unobserved when a unit is not selected.

[13]Instances of clustering solely due to repeated measures within person were not counted toward this total.

were used or because those used were unobserved or unknown. Hence, although it could be argued that sometimes observational researchers are prevented from fulfilling Fisher's conditionality principle because selection variables, strata indicators, or cluster indicators are unobserved/unknown, it is often the case that *known*, observed complex sampling features *recorded in the data set* are simply not ones on which conditioning occurs in statistical models.

Researchers were more likely to account for complex sampling features when they were viewed as relevant to substantive hypotheses, rather than a nuisance induced by the design. For example, when researchers have substantive hypotheses about how the factor structure of a measure differs across a particular demographic variable z_i (say, race or gender) involved in sample selection, measurement invariance testing is often performed. Such measurement invariance testing can be seen as a special case of the conditioning procedures from Equation (5). That is, testing for factor loading invariance across levels of continuous or binary variable z_i for a generic item y_i on a one-factor model, we have

$$y_i=\upsilon+\lambda_1\eta_i+\lambda_2 z_i+\lambda_3\eta_i z_i+\varepsilon_i, \tag{7}$$

where λ's are slopes, υ is an item intercept, and η_i is a latent independent variable. Equation (7) follows the same logic as Equation (4), except the former measured independent variables x_i are now latent independent variables η_i. But measurement invariance testing was not consistently employed for selection variables in all studies, but rather only when it garnered substantive interest.

It may be the case that Fisher's conditionality principle is inconsistently applied in psychology because the analysis of nonrandom samples is typically motivated on pragmatic grounds—for example, budgetary limitations—rather than the aforementioned statistical grounds (Serlin et al., 2003, p. 529; Shadish et al., 2002, pp. 92, 342, 348). Perhaps because motivations for analyzing non-random samples are disconnected from Fisher's statistical framework, published guidelines for analyzing nonrandom samples are as well. For example, Cook (1993, pp. 42, 61) and Lavori, Louis, Bailar, and Polansky (1986, pp. 62–63) note that merely mentioning selection criteria and clinically relevant facts about participants (presumably in a methods or discussion section) can “substitute for random selection when the latter is not possible.” No mention is made of requiring conditioning on complex sampling features. Such recommendations are reinforced by the APA's Task Force on Statistical Inference (1999) that asks members to “describe the sampling procedures and emphasize any inclusion and exclusion criteria. If the sample is stratified (e.g., by site or gender) describe fully the method and rationale” (p. 595). Although the Task Force subsequently notes that “interval estimates for clustered and stratified random samples differ from those for simple random samples” and that “statistical software is now becoming available for these purposes,” (p. 595) it does not note that (a) the same effects of stratification and clustering occur in *nonrandom* samples as well and (b) worse effects result from disproportionate selection. Most disconcerting, the Task Force again only gives the directive to *describe* complex sampling features in prose—not statistically account for them in model specification, per Fisher's framework.

In sum, whereas *in principle* observational psychologists are allied with Fisher's model-based inference approach for nonrandom samples, *in practice* the approach has often become dislodged from Fisher's strict requirements (e.g., the conditionality principle).

DESIGN-BASED INFERENTIAL FRAMEWORK

In contrast to Fisher, Neyman and Pearson (1933) deemed the construction of hypothetical infinite populations, and construction of models, to be fallible and subjective. Neyman (1957) remarked that "a model is a set of invented assumptions regarding invented entities such that if one treats these invented entities as representations of appropriate elements of the phenomena studied, the consequences of the hypotheses constituting the model are expected to agree with observations" (p. 8). Neyman did not want inferences from a sample to the finite population from which it was drawn to depend on appropriate specification of a model and appropriate conditioning on all selection and design variables (Lenhard, 2006). That is, Neyman did not want models to have a mediating role in the validity of inference.

Motivated by his work with Pearson, Neyman developed an alternative *design-based* inferential framework. Its target parameters were not hypothetical/infinite population parameters, as in the model-based framework, but rather were *finite* population parameters. Example finite population parameters are functions of the dependent variable y: the mean of y in the case of a census of the finite population, the total of y in the case of a census, or a ratio of totals. In the design-based framework, the outcome y is converted into a random variable, not through the introduction of *epistemic* randomness via imposition of distributional assumptions, as in the model-based framework, but exclusively through the introduction of *empirical* randomness from the random sampling design (Kish, 1965)—as follows.

Step 1

As a first step, rather than specifying the statistical model *hypothesized* to have generated the outcome y in the hypothetical/infinite population, Neyman's design-based framework required specifying a random sampling frame, design, and scheme that together actually *did* generate y in the finite population (Neyman, 1934, pp. 567–570). The *sampling frame* is a list of primary sampling units in the finite population; the *sampling design* assigns nonzero probabilities of selection to each sample that could be drawn from the frame; the *sampling scheme* is a draw-by-draw mechanism for implementing the sampling design (Cochran, 1977). For example, suppose we are interested in estimating the total number of drinking and driving episodes, t, experienced by high school students in a particular region. Suppose that we wanted to stratify the region on a geographic variable correlated with the outcome (rural vs. urban), creating $H = 2$ strata. Suppose further that we wanted to select $n_h = 5$ clusters (high schools) with unequal probabilities and with replacement separately in each strata. Moreover, we wanted those unequal probabilities (denoted π_{hi}, where h corresponds to stratum and i to cluster) to be proportional to a cluster-level covariate correlated with the outcome (e.g., percentage of students qualifying for free lunch). Our *sampling frame* would be a list of primary sampling units (schools) in the region along with each school's urban/rural location and percentage free lunch qualifiers. Suppose further that, at a second stage of selection, we wanted to sample $m_{hi} = 20$ students (secondary sampling units) from M_{hi} students in cluster i, with equal probabilities. Then this stratified, clustered *sampling design* would assign selection probabilities $\pi_{hi} \times \frac{m_{hi}}{M_{hi}}$ to students in cluster i of stratum h. Various *sampling schemes* exist for implementing this design (Lohr, 1999, chap. 6), which have been automated (see SAS Proc Surveyselect; SAS Institute Inc., 2008).

Step 2

Using only the known, nonzero probabilities of selection, cluster indicators, strata indicators, and observed y-values for sampled units—not a statistical model—a finite population parameter and its variance can be estimated (Cassel, Sarndal, & Wretman, 1977). To do so for our example, we would calculate a *sampling weight* as the inverse of the first stage selection probability times the second stage selection probability $w_{hi}=\frac{1}{\pi_{hi}}\times\frac{M_{hi}}{m_{hi}}$. The weight for a selected student indicates the number of students in the finite population that he or she represents. This weight contains all information needed to construct a point estimate $\hat{t}$ for our finite population parameter:

$$\hat{t}=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hi}y_{hij}. \quad (8)$$

The unconditional (design-based) variance of $\hat{t}$ across all possible samples that could be generated by the design can be approximated, adjusting for stratification and clustering. In our example, the design-based variance of $\hat{t}$ is approximated (using taylor linearization) as the variance of cluster-specific weighted totals within each strata—summed across strata (Cochran, 1977). Note that this approximation can ignore details of the sampling design below the cluster level assuming clusters were selected independently with replacement. Note also that if the ratio of the sample size (here, number of clusters) to finite population size is nontrivial (> 10%), and samples are drawn without replacement, the design-based variance would approach zero. Multiplying it by a finite-population correction (fpc) prevents this. However, in practice, the fpc is unnecessary for most large public-use surveys (Kish, 1965, p. 44).

This example shows that associating positive probabilities of selection with observed *y*-values is *all* that is needed to convert the latter into realizations of a random variable *y* under the design-based framework (Neyman, 1923). Unlike in the model-based approach, no model distributional assumptions were needed to convert *y* into a random variable. Hence, randomness of the sampling design is a mandatory requirement under Neyman's framework because it is the sole basis for the probabilistic treatment of the results during data analysis (Fienberg & Tanur, 1996).

However, a disadvantage of not specifying a model is that *y*-values of sampled units in the finite population and *y*-values of unsampled units in the finite population are not meaningfully related; they are related only to the extent that they both had a chance of being selected. Furthermore, none of these *y*-values in the finite population are meaningfully related to *y*-values outside the finite population. Consequently, only *descriptive inference* is possible with respect to the finite population parameters in the design-based framework (see Figure 1; Godambe, 1966). Descriptive inferences have the property that, if all finite population units were observed without error (a perfect census), there would be no uncertainty in the inference (Smith, 1993). *Analytic or causal inference*, about what will occur or what would have occurred under different circumstances, requires postulating a more meaningful link between sampled and unsampled units. Under Fisher's framework, this link was established by requiring sampled and unsampled units to be jointly distributed according to a parametric model (Royall, 1988).

In sum, this section has answered the second question posed earlier: "Do inferences made from nonrandom samples differ from those possible under random sampling?" Different kinds of inference (descriptive rather than analytic) to different kinds of populations (finite rather than infinite) are possible exclusively under random sampling, and explicit models are not required to achieve these inferences.

IMPLEMENTATION OF THE DESIGN-BASED FRAMEWORK IN PSYCHOLOGY

Neyman's design-based framework was soon taken up by observational survey researchers in epidemiology, sociology, health sciences, and government census and polling agencies—but not in observational psychological research (Smith, 1976). Target parameters for inference in epidemiology, health sciences, and government polling were often descriptive quantities (e.g., frequency of an outpatient medical procedure in a finite population). Additionally, researchers in those fields often had to produce thousands of estimates while knowing little about the population at hand. Hence, they could lack both the time and knowledge to construct plausible hypothetical/infinite population models for their research questions and understandably did not want the validity of their prevalence estimates to be predicated on hastily constructed, fallible models (Kalton, 2002). In contrast, observational psychologists had less interest in enumeration of particular finite populations and more interest in constructing theory-driven models to explain causal mechanisms and predict future behavior. Hence, they gravitated toward the model-based rather than design-based framework (see Deming, 1975).

LIMITATIONS OF THE PURE MODEL-BASED AND PURE DESIGN-BASED FRAMEWORKS

Following the introduction of the model-based inferential framework by Fisher and the introduction of the design-based inferential framework by Neyman, survey sampling statisticians began to identify their respective weaknesses.

With regard to the model-based framework, sampling statisticians found that conditioning on all stratification and selection/recruitment variables, and allowing for their potential interactions with independent variables, complicated model specification (Pfeffermann, 1996). Such conditioning also complicated interpretation of substantively interesting model parameters and swallowed needed degrees of freedom (Pfeffermann, Krieger, & Rinott, 1998). Additionally, such conditioning was found to be error prone; particularly if little was known about the sample selection mechanism, relevant selection/recruitment variables could easily be unknowingly omitted (Firth & Bennett, 1998; Graubard & Korn, 1996; Neyman, 1934, p. 576–577).

With regard to the pure design-based framework, sampling statisticians felt limited by restrictions on the type of parameters that could be estimated (simple statistics such as means, totals, and ratios) and the type of inference that could be obtained (descriptive, finite population inference; Graubard & Korn, 2002; Smith, 1993). Additionally, statisticians increasingly realized that the design-based framework's arguably greatest purported advantage (according to Neyman, 1923, 1934) is not entirely true: it does not provide inference free of all modeling assumptions. True, the design-based framework does not involve explicit attempts to write out a model for the substantive process that generated *y*-values in an infinite population. However, the sampling weight itself entails an implicit (or hidden) model relating probabilities of selection and the outcome (Little, 2004, p. 550). Adjustments to the weight for nonsampling errors such as under-coverage and nonresponse require further implicit modeling assumptions (Little, 2004; Smith, 1983b).[14] Another

[14]For example, multiplying sampling weights by nonresponse weights (inverse of the probability that a unit would respond, if selected) requires (a) dividing the sample into classes according to covariates known for respondents and nonrespondents and related to the outcome and (b) invoking the *implicit* assumption that all units within a class have the same response propensity (Biemer & Christ, 2008).

drawback is that types of nonsampling errors requiring explicit models (e.g., measurement error) cannot be accommodated by the design-based framework at all.

AN INTEGRATION OF THE MODEL-BASED AND DESIGN-BASED FRAMEWORKS

To summarize, sampling statisticians viewed the pure model-based framework as susceptible to bias incurred from incomplete conditioning on the sampling design. Additionally, sampling statisticians viewed the design-based framework as incongruent with analytic statistics (e.g., regression coefficients), causal inferences, and certain nonsampling (e.g., measurement) errors. This raises the question "*How can these limitations be overcome?*" Since the 1970s, work has been under way on a hybrid, integrated framework that overcomes key weaknesses of its predecessors. In the last 5 years, software implementations of this framework have greatly expanded (for review, see online Appendix: http://www.unc.edu/~ssterba/).

The hybrid framework that emerged has several main features: (a) It can produce analytic statistics (e.g., regression coefficients) from complex random samples, adjusting for disproportionate selection, stratification, and clustering—without needing to condition on all of these complex sampling features during model specification. (b) It permits causal or descriptive inference about these analytic statistics to infinite or finite populations. (c) It is flexible enough to take into account measurement error. (d) It can accommodate situations in which researchers desire to condition on some complex sampling features but not others. Although there are variations in the rationale and theoretical details of the hybrid framework (e.g., Chambers & Skinner, 2003; Firth & Bennett, 1998; Kalton, 2002; Sarndal, Swensson, & Wretman, 1992), we trace the emergence of some of its key, crosscutting developments.

(1) Account for the sampling design during model estimation not in model specification

To fix ideas, suppose we hypothesized that the Equation (1) model generated our data in the infinite population, and suppose we desire to make analytic/causal inferences about β_1. But suppose our sampling design involved disproportionate selection, stratification, and clustering. A first major breakthrough for the hybrid framework was Kish and Frankel's (1974) demonstration that, despite the disproportionate selection, we could specify the *exact* model in Equation (1) and make infinite-population inferences about β_0 and β_1—*without* conditioning on selection variables. We would simply adjust for disproportionate selection during *estimation* of the coefficient vector $\hat{\beta}$ rather than conditioning on selection variables in model *specification*. Conventionally, we would think to estimate regression coefficients in Equation (1) using ordinary least squares, that is, $\hat{\beta}=(\mathbf{X'X})^{-1}\mathbf{X'y}$, where **X** is a design matrix for independent variables and **y** is a vector of dependent variables. However, to adjust for unmodeled disproportionate selection, we instead use weighted least squares, $\hat{\beta}_w=(\mathbf{X'WX})^{-1}\mathbf{X'Wy}$. Although in conventional weighted least squares estimation the weight matrix **W** is diagonal with variance weights (i.e., inverses of individuals' error variances) as diagonal elements, here the diagonal elements are sampling weights (inverses of individuals' probabilities of selection).

A second major breakthrough for the hybrid framework was Fuller's (1975) and Binder's (1983) demonstrations that, despite this complex design, we could specify the exact model in Equation (1) and make infinite-population inferences about β_0 and β_1—*without* conditioning on strata or cluster variables. We would simply adjust for stratification and clustering during $Var(\hat{\beta})$ *estimation* rather than conditioning on them in model *specification*. The typical (model-based) weighted least squares variance estimator, that is,

$Var\left(\widehat{\beta}_w\right)=\widehat{\sigma}^2(\mathbf{X'WX})^{-1}$, did not serve this purpose; one problem was that this estimator assumed that weights were proportional to residual variances (unlikely) and another problem was that it assumed no clustering or stratification. Intermediate solutions corrected one problem but not the other (see Kish & Frankel, 1974; Nathan, 1988, for discussion). To remedy both problems, Fuller proposed a design-adjusted variance estimator using taylor linearization $Var\left(\widehat{\beta}_w\right)=(\mathbf{X'WX})^{-1}\widehat{\mathbf{G}}(\mathbf{X'WX})^{-1}$. The $\widehat{\mathbf{G}}$ matrix is a covariance matrix of totals of independent variables multiplied by weighted residuals (i.e., $\widehat{\mathbf{G}}=E\left(\mathbf{X'W}\widehat{\varepsilon}\widehat{\varepsilon}'\mathbf{WX}\right)$), where $\widehat{\varepsilon}$ is a vector of estimated person-specific residuals from Equation (1) and E denotes the expectation operator. Crucially, this $\widehat{\mathbf{G}}$ matrix automatically adjusts for any *arbitrary*, *unmodeled* stratification, clustering, or disproportionate selection scheme for any arbitrary, unmodeled number of stages of selection below the cluster level (see Sarndal et al., 1992, Sec. 5.10) for the following reason. As long as clusters had been selected independently and with replacement within stratum from a large finite population, totals in $\widehat{\mathbf{G}}$ are simply aggregated to the level of the cluster. Then, the covariance of aggregated totals is calculated across clusters for each stratum and summed across strata (Wolter, 2007, Sec. 6.11). Binder (1983) extended this approach from linear regression to a variety of other outcome distributions; his strategy is now widely implemented in software. Beyond taylor linearization, other variance estimation methods from the design-based literature were also applied (e.g., sample-weighted bootstrapping; Sarndal et al., 1992).

(2) Make infinite and/or finite population inference

Another major breakthrough for the model/design-based framework was the articulation of its greater inferential possibilities. Fuller (1975) and Godambe and Thompson (1986) showed that model estimates produced under the hybrid framework serve as estimates of finite population parameters (i.e., a regression coefficient in the case of a census) when the sample and finite population size are large—whether or not the model is correctly specified. Additionally, these model estimates serve as estimates of infinite population parameters when the model is correctly specified (see Figure 1). Hence, descriptive, *finite*-population inferences are mainly independent of a correctly specified model (as in the design-based framework) and analytic *infinite*-population inferences are mainly dependent on a correctly-specified model (as in the model-based framework; Kalton, 2002; Knott, 1991). We said "mainly dependent" because, in contrast to the pure model-based framework, the sample weighting aspect of the hybrid framework does provide some robustness to misspecifications in the fixed effects portion of the model (Binder & Roberts, 2003; Pfeffermann, 1993, 1996). Also, the design-adjusted variance estimation aspect of the hybrid framework avoids altogether needing to properly specify random effects. Furthermore, even if the fixed effects portion of a model is misspecified, the standard errors of parameter estimates will be close to traditional design-based standard errors for large finite population and sample size (Binder & Roberts, 2003).

(3) Account for measurement error

More recent breakthroughs in the hybrid framework have involved the extension of its design-based features (sample weighting and design-adjusted variance estimation) from least squares estimation of regression models to maximum likelihood estimation of structural equation models (e.g., Asparouhov, 2005; du Toit, du Toit, Mels, & Cheng, 2005; Muthén & Satorra, 1995; Stapleton, 2006, 2008). Structural equation models use multiple observed measures of latent variables to account for measurement error—something the design-based framework could not do.

(4) Account for the sampling design partially in model estimation, partially in model specification

The most recent work on the model/design-based framework involves extending it to situations in which the researcher wishes to account for particular complex sampling features during model specification but simply adjust for others during model estimation. For example, suppose the researcher wishes to account for clustering via a multilevel model (the model-based way) and account for disproportionate selection via sample weights (the design-based way) and account for stratification via standard error corrections (the design-based way). To do so, sampling weights are incorporated into the estimation of a multilevel model (e.g., Asparouhov, 2006; Asparouhov & Muthén, 2006; Korn & Graubard, 2003; Kovacevic & Rai, 2003; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002). The twist is that a weight could now be needed at each level. That is, for a two-level model, a Level 2 weight (inverse of the probability that the cluster is selected) and a Level 1 weight (inverse of the probability that the individual is selected given the cluster is selected) could be needed.

IMPLEMENTATION OF THE HYBRID MODEL/DESIGN-BASED FRAMEWORK IN PSYCHOLOGY

We have seen that model/design-based framework is *hybrid* in the sense that it allows both kinds of inference (finite and infinite) and in the sense that it allows models but does not require their full or completely correct specification. However, the model/design-based framework is not hybrid in the sense that it allows both types of samples (random and nonrandom). As can be inferred by the use of sampling weights, the hybrid framework is applicable to *random* samples only. That is, given a nonrandom sample, a researcher's only choice remains the pure model-based framework. Yet, psychologists are increasingly analyzing complex random samples through electronically available public-use data sets, for example, National Longitudinal Study of Adolescent Health (Add-Health), Early Childhood Longitudinal Study (ECLS), National Education Longitudinal Study (NELS), National Longitudinal Survey of Youth (NLSY), High School and Beyond (HSB), and National Survey of Child and Adolescent Well-Being (NSCAW), to which this framework does apply. Moreover, psychometric software programs have recently added the capability for fitting models under the hybrid model/design-based framework. Yet this capability is little discussed in psychology research methods texts. To foster implementation of this framework in psychology, in this article we provide (a) an explanation of the relative merits and interpretation of this framework (see previous section), (b) a review of software for implementing this framework (see online Appendix), and (c) an illustrative example (see next section).

Illustrative Analysis Using the Hybrid Model/Design-Based Framework

Our example uses a theoretical model from Raudenbush and Bryk (2002, chap. 4) and Singer (1998). This model stipulates that math achievement (*MATHACH*) varies across schools according to school average socioeconomic status (*MEANSES*), controlling for school *SECTOR* type (Catholic or public). This model also stipulates that the effect of school mean centered child socioeconomic status (*CSES*) on *MATHACH* varies across schools, but the strength of this relationship differs according to *MEANSES*:

$$\begin{aligned} MATHACH_{ij} &= \beta_{0j} + \beta_{1j} CSES_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} MEANSES_j + \gamma_{02} SECTOR_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} MEANSES_j + \gamma_{12} SECTOR_j + u_{1j} \\ \text{where} & \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right) \end{aligned} \quad (9)$$

Our example uses the High School and Beyond (HSB) data set, whose sampling design includes clustering, stratification, and disproportionate selection. Specifically, HSB's frame of 26,096 clusters (a list of U.S. high schools) was stratified into nine strata, largely according to school type (public, Catholic, private) and school racial composition.[15] Within some strata, schools were selected with probabilities proportional to estimated enrollment, but within other strata, schools were oversampled. In total, 1,122 schools were selected at the primary stage of selection. At the secondary stage of selection, 36 seniors and 36 sophomores were selected with equal probabilities within each selected school. Figure 3 shows resultant variation in the probabilities of selecting clusters and probabilities of selecting individuals within cluster. Diagnostics showed that both sets of probabilities were significantly related to our outcome MATHACH after controlling for independent variables ($\gamma_{\pi ij} = .767$, $SE = .056$, $p < .001$; $\gamma_{\pi i|j} = -1.207$, $SE = .090$, $p < .001$). This means that ignoring disproportionateness of selection risks bias.

The Equation (9) model has previously been fit to HSB data exclusively using the model-based framework (Raudenbush & Bryk, 2002; Singer, 1998). But, the Equation (9) model specification does not account for HSB's disproportionate selection, partially accounts for HSB's stratification (by conditioning on school type but not school racial composition or their product), and fully accounts only for HSB's clustering (by specifying a multilevel model for students nested within schools). Table 2, Column 1, depicts the results of this model-based analysis. We show that this original, model-based analysis likely incurred bias due to incompletely conditioning on the sampling design. We show that a hybrid analysis allows us to fully, and more flexibly, account for the sampling design to avoid this problem.

NIH-PA Author Manuscript

A hybrid analysis affords us the flexibility of choosing whether to account for each of HSB's complex sampling features the design-based way or the model-based way, depending on our substantive goals. In this particular hybrid analysis, we chose to adjust for disproportionate selection in a design-based way (including sampling weights at both levels during estimation; see online Appendix) rather than the model-based way (including selection variables as model covariates). We made this choice because selection variables were a nuisance here, not of substantive interest. We chose to account for stratification in the model-based way (including strata variables as model covariates) rather than the design-based way (standard error adjustments using the HSB-provided strata indicator *SCHSAMP*). That is, we included fixed effects for *SECTOR*, high percentage Black enrollment (*BLACK*), high percentage Hispanic enrollment (*HISPANIC*), and their product terms (*SECTOR* × *BLACK* and *SECTOR* × *HISPANIC*). We made this choice because one variable involved in stratification (*SECTOR*) was of substantive interest in the original model and was thought to interact with independent variables. In contrast, design-based adjustments for stratification typically assume no interaction between strata and independent variables. Finally, we chose to account for clustering the model-based way (inclusion of random effects for cluster)

NIH-PA Author Manuscript

[15]A few strata were further divided (e.g., by urbanization, geographic region), but because accounting for this in model specification did not alter results, it is not discussed further. Within sector, stratification according to school racial composition involved classifying schools according to whether they were high Cuban (≥ 30%), high other-Hispanic (≥ 30%), high Black (≥ 30%), or Regular. For purposes of including stratification variables as model covariates, high Cuban and high other-Hispanic were collapsed into a high Hispanic variable for lack of a school-level percentage Cuban flag in the HSB data set.

rather than the design-based way (standard error adjustments using the HSB-provided cluster indicator *SCHLID*). We made this choice because distinguishing between- from within-effects was of substantive interest. M*plus* 5.2 (Muthén & Muthén, 1998–2007) code for all fitted models is provided in the online Appendix.

Including sampling weights during estimation of the Equation (9) model yielded the results in Table 2, Column 2. Comparison of Columns 2 and 1 indicates that some bias was likely incurred in prior (model-based, unweighted) analyses due to ignoring disproportionate selection. Although the conditional slope of *CSES* on *MATHACH* is still significant in Column 2, and still varies across schools, the slopes for *CSES* no longer significantly differ according to school *MEANSES*. That is, the cross-level interaction of *CSES* by *MEANSES* is now nonsignificant. Further, there is now nonsignificant covariation between intercepts and slopes in Column 2, meaning that the effects of *CSES* on *MATHACH* no longer covary with the average *MATHACH* of the school. Also including omitted strata variables as model covariates completes the hybrid analysis; these results are shown in Table 2, Column 3. Comparing Columns 3 and 2 indicates that, in this case, more fully accounting for stratification does not markedly change conclusions. However, note that not only do standard errors change from Column 2 to Column 3 but in this case several parameter estimates do as well. Recall that we earlier mentioned that stratification should affect *only* standard errors, not parameter estimates when stratification variables neither interact with nor correlate with independent variables. This is clearly not the case here. We do not explore here whether school racial composition interacts with student or school socioeconomic status.

It is important that this hybrid analysis also provides us with choices in drawing inferences. For example, we can make descriptive inferences about γ_{10}, the conditional effect of a unit increment in *CSES* on *MATHACH* in the finite population of U.S. high schools, without assuming an entirely correct model. Or we can make analytic inferences about γ_{10} in the infinite population—assuming a correct model.

CONCLUSIONS

This article began by posing two often-asked but incompletely answered questions about inferences from nonrandom versus random samples in observational psychology research. To address these questions, we began by reviewing two alternative inferential frameworks from samples to populations and discussing the extent of each framework's implementation in psychology. In reviewing the model-based inferential framework, we showed that it does in fact provide a formal logic for making statistical inferences from nonrandom (or random) samples to infinite populations. Second, in reviewing the implementation of the model-based framework in psychology, we showed that its requirements are often not completely fulfilled in psychological research, even when measured indicators of stratification, clustering, and/or disproportionate selection are available in a data set to make this possible. We suggested that psychologists' long tradition of simply reporting, but not fully conditioning on, complex sampling features contributes to the inconsistent fulfillment of these requirements. In reviewing the design-based inferential framework, we showed that different kinds of statistical inferences (descriptive rather than analytic) to different populations (finite rather than infinite) were possible exclusively under random sampling—and their accuracy was not dependent on the proper specification of a hypothetical model. In reviewing the implementation of the design-based framework in psychology, we provided reasons for its scant use in psychology. Finally, having addressed the two original questions often asked by psychologists, we pushed the dialogue a step further, asking, what are the limitations of the model-based and design-based frameworks, and how can these be overcome? We showed that the model-based framework's central limitation lies in the need to tediously condition on all complex sampling features in model specification, and the design-based framework's central limitation lies in the inability to address analytic/causal hypotheses and account for nonsampling errors. We therefore described a newer hybrid model/design-based framework that overcomes these limitations and can be used for analyzing large, complex random samples from public-use data sets—a practice that is becoming more common in psychology. To facilitate greater implementation of the hybrid inferential framework in psychology, we provided an empirical illustration and reviewed applicable software in an online Appendix. We hope that this article spurs readers to attend to the requirements of their chosen inferential framework and provides motivation to take advantage of newer, more flexible inferential frameworks where possible.

Acknowledgments

Support for this work was provided by NIMH F31 MH080484-03. The author would like to thank Daniel Bauer, Robert MacCallum, Kristopher Preacher, and Christopher Wiesen for helpful comments on earlier versions of this article.

REFERENCES

Asparouhov, T. Stratification in multivariate modeling. 2004. Unpublished manuscript. Available from www.statmodel.com

Asparouhov T. Sampling weights in latent variable modeling. Structural Equation Modeling 2005;12:411–434.

Asparouhov, T. General multilevel modeling with sampling weights. 2006. Unpublished manuscript. Available from www.statmodel.com

Asparouhov, T.; Muthén, BO. ASA Section on Survey Research Methods. 2006. Multilevel modeling of complex survey data. Available from http://www.statmodel.com

Berk RA. An introduction to sample selection bias in sociological data. American Sociological Review 1983;48:386–398.

Biemer, P.; Christ, S. Weighting survey data. In: de Leeuw, E.; Hox, J.; Dillman, D., editors. International handbook of survey methodology. Erlbaum; Mahwah, NJ: 2008. p. 317-341.

Binder DA. On the variance of asymptotically normal estimators from complex surveys. International Statistical Review 1983;51:279–292.

Binder, DA.; Roberts, GR. Design-based and model-based methods for estimating model parameters. In: Chambers, RL.; Skinner, CJ., editors. Analysis of survey data. Wiley; Chichester, UK: 2003. p. 29-48.

Bowley AL. Address to the economic and statistics section of the British association for the advancement of science, York, 1906. Journal of the Royal Statistical Society 1906;69:540–558.

Cassel, C.; Sarndal, C.; Wretman, J. Foundations of inference in survey sampling. Wiley; New York: 1977.

Chambers, RL.; Skinner, CJ. Analysis of survey data. Wiley; Chichester, UK: 2003.

Cochran, WG. Sampling techniques. 3rd ed.. Wiley; New York: 1977.

Cook TD. A quasi-sampling theory of the generalization of causal relationships. New Directions for Program Evaluation 1993;57:39–82.

Cook, TD.; Campbell, DT. Quasi-experimentation: Design and analysis issues for field settings. Houghton Mifflin; Boston: 1979.

Cronbach, LJ. Designing evaluations of educational and social programs. Jossey-Bass; London: 1982.

Dawid AP. Fisherian inference in likelihood and prequential frames of reference. Journal of the Royal Statistical Society, Series B 1991:79–109.

Deming WE. On probability as a basis for action. The American Statistician 1975;29:146–152.

du Toit, SHC.; du Toit, M.; Mels, G.; Cheng, Y. Analysis of complex survey data with LISREL: Chapters 1–5. 2005. Unpublished manual. Available from www.ssicentral.com

Fienberg SE, Tanur JM. Experimental and sampling structures: Parallels diverging and meeting. International Statistical Review 1987;55:75–96.

Fienberg SE, Tanur JM. Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. International Statistical Review 1996;64:237–253.

Firth D, Bennett KE. Robust models in probability sampling. Journal of the Royal Statistical Society, Series B 1998;60:3–21.

Fisher RA. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, Series A 1922;222:309–368.

Fisher, RA. Statistical methods for medical workers. Oliver & Boyd; Edinburgh, UK: 1925.

Fisher, RA. The design of experiments. Oliver & Boyd; Edinburgh, UK: 1935.

Fisher RA. Statistical methods and scientific induction. Journal of the Royal Statistical Society, B 1955;17:69–78.

Fisher, RA. Statistical methods and scientific inference. Hafner; New York: 1956.

Fisher RA. The nature of probability. Centennial Review 1958;2:261–274.

Fuller WA. Regression analysis for sample survey. Sankhya: Series C 1975;37:117–132.

Godambe VP. A new approach to sampling from finite populations. Journal of the Royal Statistical Society, Series B 1966;28:310–328.

Godambe VP, Thompson ME. Parameters of superpopulation and survey population: Their relationships and estimation. International Statistical Review 1986;54:127–138.

Graubard BI, Korn EL. Modeling the sampling design in the analysis of health surveys. Statistical Methods in Medical Research 1996;5:263–281. [PubMed: 8931196]

Graubard BI, Korn EL. Inference for superpopulation parameters using sample surveys. Statistical Science 2002;71:73–96.

Grilli, L.; Rampichini, C. Sample selection in random effects models. Universita di Firenze; 2003. Unpublished manuscript

Jaffe E. How random is that? Association of Psychological Science, Observer 2005;18:9.

Johnstone DJ. Tests of significance following R. A. Fisher. British Journal of Philosophy of Science 1987;38:481–499.

Johnstone DJ. On the necessity for random sampling. British Journal of Philosophy of Science 1989;40:443–457.

Kalton G. Models in the practice of survey sampling (revisited). Journal of Official Statistics 2002;18:129–154.

Kempthorne O. Discussion of “On rereading R. A. Fisher” by Leonard J. Savage. Annals of Statistics 1976;4:495–497.

Kiaer AN. Observations et expériences concernant des dénombrements représentatifs. Bulletin of the International Statistical Institute 1895;9:176–183.

Kish, L. Survey sampling. Wiley; London: 1965.

Kish L. Developing samplers for developing countries. International Statistical Review 1996;64:143–162.

Kish L, Frankel MR. Inference from complex samples (with discussion). Journal of the Royal Statistical Society, Series B 1974;36:1–37.

Knott P. A model-based look at linear regression with survey data. American Statistician 1991;45:107–112.

Korn EL, Graubard BI. Estimating variance components by using survey data. Journal of the Royal Statistical Society, Series B 2003;65:175–190.

Kovacevic MS, Rai SN. A pseudo maximum likelihood approach to multilevel modelling of survey data. Communications in Statistics 2003;32:103–121.

Lavori, PW.; Louis, TA.; Bailar, JC.; Polansky, H. Designs for experiments: Parallel comparisons of treatment. In: Bailar, JC.; Mosteller, F., editors. Medical uses of statistics. New England Journal of Medicine; Waltham, MA: 1986. p. 61-83.

Lee VE, Bryk AS. A multilevel model of the social distribution of high school achievement. Sociology of Education 1989;62:172–192.

Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association 1993;88:1242–1249.

Lenhard J. Models and statistical inference: The controversy between Fisher and Neyman-Pearson. British Journal of Philosophy of Science 2006;57:69–91.

Lillard, LA.; Panis, CWA. aML: Multilevel multiprocess statistical software (Release 1) [Computer software]. EconWare; Los Angeles: 2000. Available from http://www.applied-ml.com/index.html

Little RJA. To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association 2004;88:546–556.

Lohr, SL. Sampling: Design and analysis. Brookes/Cole; Pacific Grove, CA: 1999.

Muthén, LK.; Muthén, BO. Mplus users guide. Author; Los Angeles: 1998–2007.

Muthén BO, Satorra A. Complex sample data in structural equation modeling. Sociological Methodology 1995;25:267–316.

Nathan, G. Inference based on data from complex sample designs. In: Krishnaian, PR.; Rao, CR., editors. Handbook of Statistics. Vol. Vol. 6. Elsevier; Amsterdam: 1988. p. 247-266.

Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. Applied linear statistical models. 4th ed.. McGraw-Hill; Chicago: 1996.

Neyman J. On the application of probability theory to agricultural experiments: Essay on principles (translated in 1990). Statistical Science 1923;5:465–480.

Neyman J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 1934;109:558–606.

Neyman J. Inductive behavior as a basic concept of philosophy of science. Revue d'Institute International de Statistique 1957;25:7–22.

Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, A 1933;231:289–337.

Pearson K. Mathematical contributions to the theory of evolution, XI. Philosophical Transactions of the Royal Society of London, Series A 1903;200:1–66.

Peterson RA. On the use of college students in social science research: Insights from a second-order meta-analysis. Journal of Consumer Research 2001;28:250–261.

Pfeffermann D. The role of sampling weights when modeling survey data. International Statistical Review 1993;61:317–337.

Pfeffermann D. The use of sampling weights for survey data analysis. Statistical Methods in Medical Research 1996;5:239–261. [PubMed: 8931195]

Pfeffermann D, Krieger AM, Rinott Y. Parametric distributions of complex survey data under informative probability sampling. Statistica Sinica 1998;8:1087–1114.

Pfeffermann D, Skinner CJ, Holmes DJ, Goldstein H, Rasbash J. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society, Series B 1998;60:23–56.

Rabe-Hesketh S, Skrondal A. Multilevel modeling of complex survey data. Journal of the Royal Statistical Society, Series A 2006;169:805–827.

Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. 2nd ed.. Sage; Thousand Oaks, CA: 2002.

Royall, RM. The prediction approach to sampling theory. In: Krishaniah, PR.; Rao, CR., editors. Handbook of statistics. Vol. Vol. 6. Elsevier Science; Amsterdam: 1988. p. 399-413.

Sarndal, C.; Swensson, B.; Wretman, J. Model assisted survey sampling. Springer; New York: 1992.

SAS Institute Inc.. SAS OnlineDoc 9.1.3. Author; Cary, NC: 2004. SAS procedures: The QLIM procedure.

SAS Institute Inc.. SAS/STAT 9.2 user's guide. Author; Cary, NC: 2008. Chapter 87: The Surveyselect procedure.

Sears DO. College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. Journal of Personality and Social Psychology 1986;51:515–530.

Serlin RC, Wampold BE, Levin JR. Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: A comment on Siemer and Joorman (2003). Psychological Methods 2003;8:524–534. [PubMed: 14664687]

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

Sherman RC, Buddie AM, Dragan KL, End CM, Finney LJ. Twenty years of *PSPB*: Trends in content, design, and analysis. Personality and Social Psychology Bulletin 1999;25:177–187.

Siemer M, Joormann J. Power and measures of effect size in analysis of variance with fixed versus random nested factors. Psychological Methods 2003;8:497–517. [PubMed: 14664685]

Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. Journal of Educational and Behavioral Statistics 1998;24:323–355.

Skinner, CJ.; Holt, D.; Smith, TMF. Analysis of complex surveys. Wiley; New York: 1989.

Smith TMF. The foundations of survey sampling: A review. Journal of the Royal Statistical Society, Series A 1976;139:183–204.

Smith TMF. On the validity of inferences from non-random samples. Journal of the Royal Statistical Society, Series A 1983a;146:394–403.

Smith TMF. An evaluation of model-dependent and probability-sampling inferences in sample surveys: Comment. Journal of the American Statistical Association 1983b;78:801–802.

Smith TMF. Populations and selection: Limitations of statistics. Journal of the Royal Statistical Society, Series A 1993;156:144–166.

Stapleton LM. The incorporation of sample weights into multilevel structural equation models. Structural Equation Modeling 2002;9:475–502.

Stapleton, LM. Using multilevel structural equation modeling techniques with complex sample data. A second course in structural equation modeling. Hancock, GR.; Mueller, R., editors. Information Age Publishing; Greenwich, CT: 2006. p. 345-383.

Stapleton LM. Variance estimation using replication methods in structural equation modeling with complex sample data. Structural Equation Modeling 2008;15:183–210.

Stephan F. History of the uses of modern sampling procedures. Journal of the American Statistical Association 1948;43:12–39.

Sterba, SK.; Prinstein, MJ.; Nock, M. Beyond pretending complex nonrandom samples are simple and random. APA Convention; Boston, MA. Aug. 2008

Sugden RA, Smith TMF. Ignorable and informative designs in survey sampling inference. Biometrika 1984;71:495–506.

Wilkinson L, the Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist 1999;54:594–604.

Wintre M, North C, Sugar L. Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. Canadian Psychology 2001;42:216–225.

Wolter, KM. Introduction to variance estimation. 2nd ed.. Springer; New York: 2007.

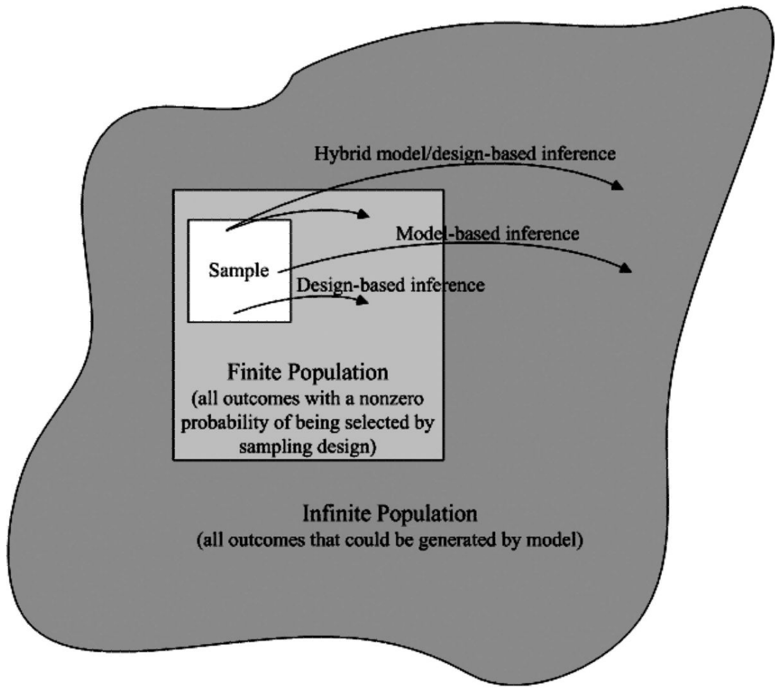

FIGURE 1.
Schematic of alternative populations of inference and mechanisms for inference.

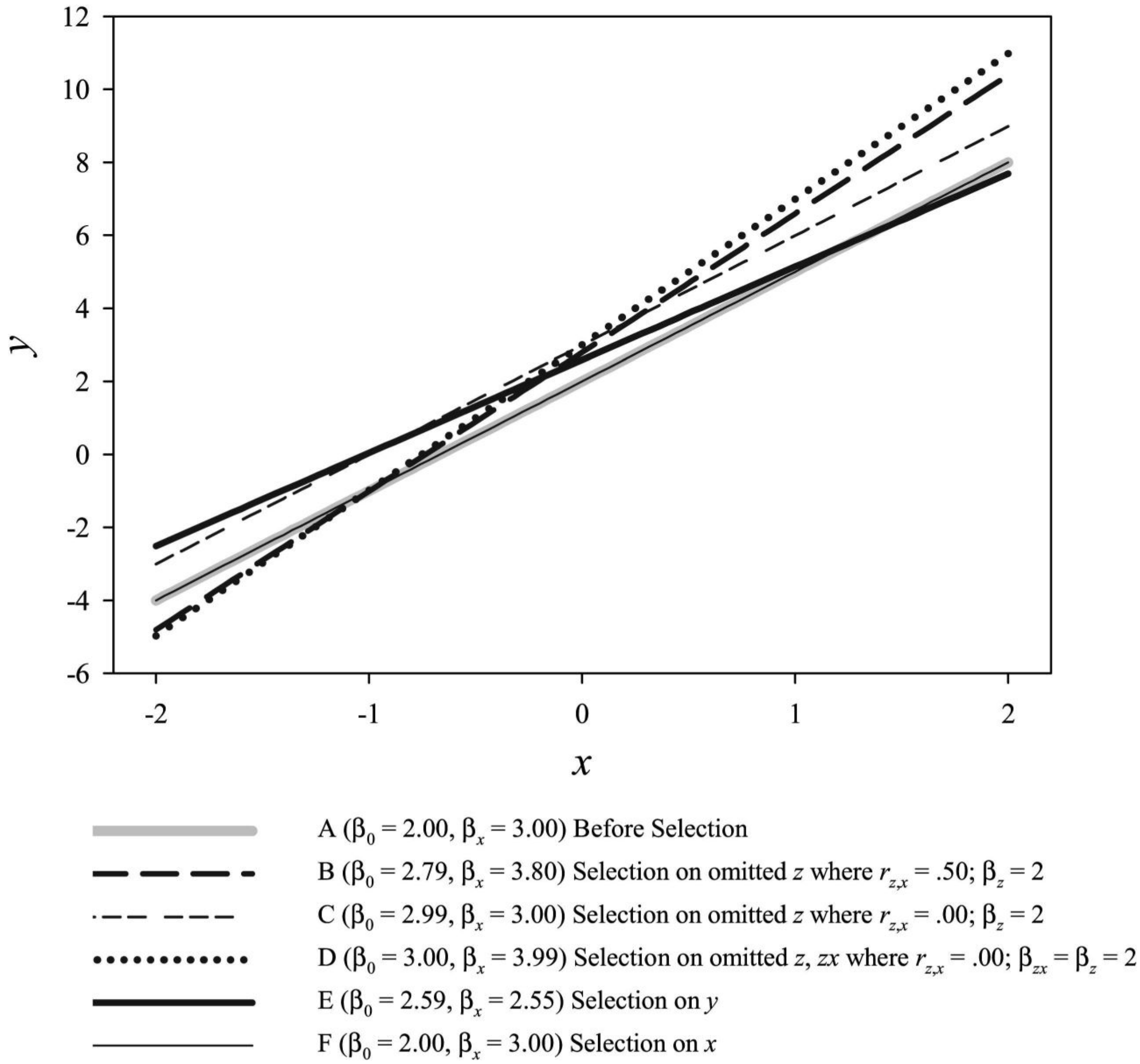

FIGURE 2.

Simulation demonstration: Effects of disproportionate selection in a single-level model.

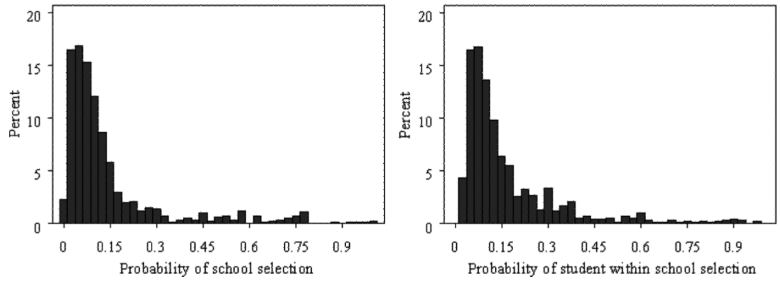

FIGURE 3.

Distributions of students' and schools' probabilities of selection in the High School and Beyond data set.

TABLE 1

Simulation Demonstration: Effects of Disproportionate Selection in a Multilevel Model

	Before Selection	After Disproportionate Selection at Cluster-Level Only				After Disproportionate Selection at Individual-Level Only		After Disproportionate Selection at Both Levels
	Infinite Population Parameters	Select on z_j $r_{w_j,z_j} = .50$	Select on z_j $r_{w_j,z_j} = .00$	Select on z_j ; $x_{ij}\ z_j$ $r_{w_j,z_j} = .00$	Select on z_j ; $w_j\ z_j$ $r_{w_j,z_j} = .00$	Select on w_j	Select on x_{ij}	Select on y_{ij}	At Stage 1, Select on z_j At Stage 2, Select on z_{ij} $r_{w_j,z_j} = r_{z_{ij},x_{ij}} = .50$
Fixed effects									
γ_{00}	2	6.98	7.59	7.60	7.59	2.00	2.00	3.43	12.41
γ_{10}	3	3.00	3.00	8.60	3.00	3.00	3.00	2.59	3.43
γ_{01}	3	3.44	3.00	3.00	8.60	3.00	3.00	2.40	3.43
Variance comp.									
τ_{00}	1	2.29	2.45	2.46	8.27	1.00	1.00	.71	2.29
τ_{01}	0	.00	.00	1.45	.01	.00	.00	−.09	.01
τ_{11}	1	.99	1.00	2.45	1.00	1.00	1.00	1.03	1.01

Note. x_{ij} is a Level 1 independent variable from Equation (6). w_j is a Level 2 independent variable from Equation (6). z_j is an omitted Level 2 selection variable; $\gamma_{z_j} = 2$. $x_{ij}z_j$ is an omitted interaction term between a Level 1 independent variable and Level 2 selection variable; $\gamma_{x_{ij}z_j} = 2$. w_jz_j is an omitted interaction term between a Level 2 independent variable and Level 2 selection variable; $\gamma_{w_jz_j} = 2$. z_{ij} is an omitted Level 1 selection variable; $\gamma_{z_{ij}} = 2$. Note that selection always occurred at the mean of the designated selection variable. Variance comp. = variance components.

TABLE 2

Illustrative Hybrid Design/Model-Based Analysis Using the High School and Beyond (HSB) Data Set

	1. Original, Model-Based Analysis[a]	**2. Hybrid Analysis, Intermediate Step**	**3. Hybrid Analysis, Final Step**
Fixed Effects	**Accounts for Clustering; Partially Accounts for Stratification**	**Model 1 Plus Weights to Account for Disproportionate Selection**	**Model 2 Plus Covariates to Fully Account for Stratification**
INTERCEPT	7.27 (.06)**	7.59 (.11)**	7.76 (.12)**
CSES	2.09 (.07)**	2.16 (.11)**	2.16 (.11)**
MEANSES	4.43 (.14)**	4.37 (.25)**	3.60 (.28)**
SECTOR	−0.06 (.24)	−0.01 (.36)	0.15 (.36)
CSES × *MEANSES*	0.62 (.17)**	0.39 (.28)	0.39 (.28)
CSES × *SECTOR*	−1.50 (.19)**	−1.63 (.27)**	−1.63 (.28)**
HISPANIC			−0.96 (.32)**
BLACK			−1.83 (.29)**
HISPANIC × *SECTOR*			0.25 (.59)
BLACK × *SECTOR*			−1.42 (.74)
Variance Components			
τ_{00}	1.86 (.15)**	2.05 (.27)**	1.73 (.24)**
τ_{01}	0.31 (.13)*	0.27 (.16)	0.18 (.16)
τ_{11}	0.29 (.10)**	0.53 (.26)*	0.55 (.26)*
σ^2	21.89 (.25)**	21.09 (.24)**	21.08 (.34)**

[a]The model-based analysis results in Column 1 differ somewhat from those of Raudenbush and Bryk (2002, chap. 4) and Singer (1998) for two reasons. First, our variables were taken directly from HSB's 1982 public-use datafile for the sophomore cohort (see online Appendix and www.icpsr.umich.edu). Raudenbush and Bryk (2002) constructed and used factor score composites of 1980 and 1982 datafile variables for sophomore and senior cohorts (Lee & Bryk, 1989). Second, we used all public and Catholic schools and they used a random subset. MATHACH = math achievement; MEANSES = school average socioeconomic status; SECTOR = Catholic or public school; CSES = school mean centered child socioeconomic status; BLACK = high % Black enrollment; HISPANIC = high % Hispanic enrollment.

* $p < .05$.

** $p < .01$.