# SECONDARY ANALYSIS OF LARGE SOCIAL SURVEYS

Jeremy Freese

Robert Wood Johnson Scholars in Health Policy Research, Harvard University

Department of Sociology, University of Wisconsin-Madison

*jfreese@ssc.wisc.edu*

Version date: July 5, 2007

Word count: 8,968 plus references

 I am grateful to Brian Powell and Eszter Hargittai for helpful comments.

A prospective graduate student once explained his resolute lack of interest in quantitative social research to me by saying the work seemed to him akin to dying before one was dead. Although quantitative methodology encompasses a broad array of otherwise quite distinct types of research, those who view it as a soulsucking craft seem often to be thinking primarily about the secondary analysis of large social surveys. Somehow, the prospect of a career spending much of one's time analyzing numeric data collected by other people does not capture the romantic imagination in quite the same way as doing fieldwork in some intriguing locale, having long face-to-face interactions with a small number of selected informants, or sitting in a coffeehouse debating finer points of French social theory. Plus, as human lives and the processes by which human fates diverge are vastly more complicated than anything that can be represented by a few hundred numbers, one might be leery of the whole premise that important insights into social life can be gained by gazing into one's monitor at the consequences of operations on these numbers. In brief, a prospective quantitative social researcher might seem to be signing up for a career of boring, solitary labor contributing to a dreary mount of incremental oversimplifications on whatever narrow topic they choose to specialize. Talk about dying before being dead!

I have to confess that I started graduate school with similar preconceptions, and I planned to avoid quantification and questionnaires as much as possible. Now, however, much of my research involves working with large surveys, and, at this writing, I remain physically and otherwise alive. What I came to realize was that large-scale social surveys provide the best means for addressing a broad class of questions about social life that are theoretically interesting and important for broader social understanding or for social policy. The ability to work with large survey data knowledgably allows one to engage such questions when they intersect one's

intellectual agenda. In addition, these surveys offer important advantages when one wishes to do work that seeks to engage contentious topics, as I will discuss. The labor itself turns out to be intricate, challenging, and subtle, requiring both creativity and discipline to be done well, and even then involves the same kinds of setbacks and problem-solving that characterize the research process more generally.

Doing quantitative research of course requires understanding matters covered in statistics or econometrics texts, but those are not the purview of this essay. Instead, my purpose is to introduce some of the practical issues that are part of the craft of analyzing large-scale surveys, and to provide my thoughts on navigating those issues. To keep matters concrete, I proceed using a single project as an extended illustration—a project I did as a graduate student (in essence, my own introduction to large-scale survey analysis) that was later published with my dissertation advisor and another collaborator in the *American Sociological Review* (Freese, Powell, and Steelman 1999). While I have mixed feelings about excavating a project I completed nearly a decade ago, the project works for illustrating the main points I wish to make, the methodology is fairly straightforward and transparent, and the benefit of distance affords the chance to reflect on ways the project could have been better. Even so, technological advances have made the practical work of quantitative analysis different now than it was then, and I make note of key differences along the way. At the end, I also discuss some possibilities for using survey data for stronger and more satisfying conclusions than what we could do in our study.

## BACKGROUND

The story begins in 1996, with me reading a *New Yorker* article about a new book by a scholar named Frank Sulloway titled *Born to Rebel* (hereafter *BTR*). The article intrigued me so

much that I had a friend drive me to a bookstore so I could buy the book; then, I stayed up all night and read the entire thing. By the time I was done, I decided I wanted to try to test some of the book's claims for myself to see if they were true.

Sulloway has an impressive intellectual biography that includes being a member of the Harvard Society of Fellows and a recipient of a MacArthur "genius" grant. His theory in *BTR* is that sibling competition for parental resources causes children to develop ways of thinking and acting that maximize the resources they receive from their parents, and that these differences persist into adulthood. That is, children develop in ways that help them stake out a "family niche." Even though sibling rivalries play out uniquely within each family, firstborns have a consistent edge when competing with their younger siblings because, in the early years, firstborns tend to be larger, stronger, and more intellectually developed. As a result of all these advantages, Sulloway theorizes firstborns tend to occupy the dominant position among their siblings, and they develop the attitudes and personalities that are optimally suited for protecting this position—namely, attitudes that are more conservative and tough-minded and personalities that are more jealous, conscientious, and dominant. Meanwhile, laterborn children are chronic underdogs who have to work to differentiate themselves to increase their share of parental resources. Sulloway argues that laterborns tend to develop attitudes that are more liberal and compassionate and personalities that are more open and rebellious.

Granted, when put so briefly, the theory feels a little hokey to me even as I type this. What makes *BTR* impressive is the voluminous evidence Sulloway presents in support of his theory, which he had spent over twenty years amassing. This included data on over 3000 scientists from the 18th through 20th centuries who participated in 28 different scientific controversies. He found that laterborns were twice as likely as firstborns to adopt scientific

innovations early. For example, laterborn scientists were disproportionately likely to adopt Darwin's theory of natural selection and Einstein's theory of relativity, while firstborns were overrepresented among doubters. He also gathered data on participants in numerous historical events, finding that those who were laterborns tend to drive liberal social movements while firstborns tend to be conservative and uncompassionate reactionaries. As examples, laterborns were overrepresented among the early leaders of the Protestant Reformation and French Revolution, and firstborns were overrepresented among their foes. In addition, Sulloway conducted an extensive review of the existing literature on birth order--notorious for its mixed results--and reported that when methodologically weak studies were excluded, the pattern of results was, in fact, quite consistent and supported the implications of his theory.

All this evidence certainly impressed others. Blurbs on the back of a book's jacket are invariably enthusiastic, and yet *BTR* featured what may be still the most impressive jacket blurbs I have ever seen. Edward O. Wilson, the eminent biologist and winner of two Pulitzer prizes, is quoted as calling *BTR* "one of the most authoritative and important treatises in the history of the social sciences." Robert K. Merton, arguably the most important American sociologist of the twentieth century, adds "A quarter century in the making, this brilliant, searching, provocative and readable treatise promises to remain definitive for twice as long." In case these endorsements were somehow not emphatic enough, a renowned anthropologist predicts *BTR* would have "the same kind of long-term impact as Freud's and Darwin's." Elsewhere, the editor of *Skeptic* magazine—whom you might expect to be a tough sell—called *BTR* "the most rigorously scientific work of history ever written" (Shermer 1996, p. 63) and announced the possibility of the book being "comparable to the Kepler-Galileo-Newton impact in changing astrology into

astronomy" (p. 66). Freud, Darwin, Kepler, Galileo, Newton: rarely are comparisons to these names made for trade books of social science!

All this, and yet the book did not contain any first-hand examination of contemporary population data, only Sulloway's own analyses of historical figures and his summary of studies of contemporary populations conducted by others. However, if Sulloway's theory was true, one should be able to observe implications of it readily in social survey data, provided one did not make any of the methodological errors or alternate decisions that Sulloway argued had obscured real birth order effects in earlier research. As I had bought and read Sulloway's book immediately upon its release, I thought if I worked quickly I could be the first one to address the theory using contemporary data. I approached a faculty member in my Ph.D. program, Brian Powell, and we agreed to work on the project with a longtime collaborator of his, Lala Carr Steelman.

**FINDING DATA**

The plan was always to test the theory using some existing set of data, rather than attempting to collect new data on birth order and personality or social attitudes ourselves. The obvious advantage of using secondary data is one's project benefits from all the hours and expertise involved in putting the original data together. After all, major survey projects convene experts in sampling, questionnaire, and survey fielding operations, and then they spend millions of dollars and employ dozens of interviewers to implement their design. The obvious disadvantage of using secondary data is one has to take the data on its terms, which may be more

or less well-suited to one's animating question.[1] In our case, we believed we could find secondary data that would allow us to contribute to the evaluation of Sulloway's theory far more efficiently and effectively than anything we could do collecting our own data. Internet surveys today do make it more plausible for graduate students to conduct surveys cheaply on geographically distributed samples, but at least for now, this works much better for projects with well-defined samples of individuals expected to be online (i.e, surveys of academics), not when one is looking to do research on the general population.

Originally, we set to work looking at the National Educational Longitudinal Study of 1988 (NELS:88), which is based on a representative sample of American students who were eighth-graders in 1988. In retrospect, I have no idea why we decided this was a good dataset for studying Sulloway's theory, other than that Powell was already using it in other projects and we knew that it had birth order information. The NELS dataset includes various indicators of student performance and achievement, and while Sulloway's book makes some claims about achievement, it was a stretch to think any of the measures there could really make a direct contribution to testing the parts of his theory I thought were most interesting, namely those related to political and personality differences by birth order. Even so, I tried mightily to convince myself that I could make it work, as if I could somehow will new and more suitable questions to appear magically on a survey that had been administered years earlier. I spent so much time considering the data and how I might make it work for the project that I had an

[1] One might presume this changes radically when one gets to be involved in constructing items for a survey, but the economy of large-scale surveys is so tight that one has to choose only a subset of the items one would ideally ask, and my experience has been that when one learns from the items one does ask leads to remorse about the items that were cut.

entirely separate idea for a paper that Powell and I wrote and published in the *American Journal of Sociology* while the project on *BTR* floundered (Freese and Powell 1999).

Then one day I was sitting in the sociology department's computer lab and started idly flipping through a General Social Survey (GSS) codebook that was sitting on a table. GSS has many measures of social attitudes and so would have been a more obvious fit for our study than NELS, but it did not collect information on respondent's birth order. I discovered from the codebook, however, that in 1994 GSS had asked respondents to list all their brothers and sisters, dead and alive, and the years they had been born. (I presume this is obvious, but if you know when a person was born and when their siblings were born, you can infer their birth order.) I told Powell about this and immediately we dropped the idea of using NELS and instead focused on testing Sulloway's theory using GSS.

In this respect, I was lucky, and I suppose could wax enthusiastic about the role of serendipity in social research. The more apt lesson, I think, is that it would have been well worth my time to have made a more systematic and thorough search of available data resources rather than focusing immediately on NELS even when it seemed unpromising. The paper I did get using NELS was developed "inductively," by looking at the data and realizing it would be appropriate for testing a different theory, while my efforts to use it more "deductively," for a question for which it was ill-suited, went nowhere. As one becomes more familiar with a data resource that contains all kinds of information, ideas emerge on how the unique strengths of those data can be used to answer other questions. When one begins with a question one wants to answer, meanwhile, time searching for and confirming that one is using the right and best data is time well-spent. Expanding online resources are making it ever easier to plumb the contents of different datasets quickly, and technological advances and institutional imperatives are also

increasing the availability of data to researchers outside the immediate circle involved in the original data collection.

To draw an even more general lesson, a common tendency in social research is to carry over a belief that is useful and true in many life contexts, which is that the ultimate quality of a finished product with our name on it is basically up to us. That is: if we work hard enough, are clever enough, write lucidly and insightfully enough, or whatever, the paper we are working on can attain a level of quality suitable for any goal we might have, such as publication in the top journal of one's field. Journal editors, however, are not judges in some abstract intellectual virtue pageant, but instead they are looking for convincing work that contributes to the field. The quality and suitability of one's data set a ceiling on how convincing an argument a researcher can ultimately make, just as a chef in charge of salad can only do so much with wilted lettuce and stringy carrots. Investing time in finding the best data, and working to secure access to the best data if such work must be done, is not just usually rewarded with better publication prospects but can also save time, as it reduces the number of iterations a paper goes through as the author tries to figure out how to make the best of a bad situation. (When the limits of data are plain, the better route may be to finish a modest paper intended for a modest venue, and seek either to find a more tractable question for the same data, or better data to answer the same question. The modest paper you finish can provide a call for better data, and, at least later in one's career, one might use this call for better data to attempt to convince a granting agency to give one the resources to collect it.)

As it was, GSS still only allowed me to address part of Sulloway's theory that I found interesting—the relationship between birth order and social attitudes. (By social attitudes, I mean political and moral opinions, such as attitudes about abortion, animal rights, giving benefits

to immigrants, the importance of teaching children obedience, or the trustworthiness of government.) I was as interested in the parts of Sulloway's project that concerned personality as social attitudes, but trying to tendentiously interpret various GSS items to press them into service as personality items would have not been especially convincing. Instead, we narrowed the question to what the GSS could answer well, and went ahead as though social attitudes were our primary interest all along. Ultimately, because one is working with data that have already been collected and thus one is stuck with what one has, I think secondary data analysis often proceeds best when one begins with a broad research question that has room for shaping and narrowing when applied to real data. The ideal is to be able to present a focused and interesting question for which the data allow the possibility of compelling evidence.

**MAKING ANALYTIC DECISIONS**

The analysis strategy for the project was relatively straightforward: take some social attitude items in GSS, and see if they were associated with respondents' birth order. Simple as this sounds in the abstract, however, trying to enact it invoked numerous complicated concrete details. What I quickly learned is that trying to test theories using social surveys involves a large number of small decisions. Obvious for the study of birth order is what to do about respondents who had reported stepsiblings--does having an older stepsibling make one a laterborn? What to do about only children? Twins? What measures of social attitudes should be used? What variables should be used as covariates ("controls") in the models? How should these covariates be measured? What should be done with cases that did not answer one of the items used to construct one of the covariates? Usually minor decisions do not make major differences for one's results, but sometimes they do. If a particular study involves 25 decisions between two

alternative ways of doing the analysis versus another—25 being, at least in my experience, a low-end estimate of the number of such decisions in an actual study—this implies there are over 30 million ($2^{25}$) different configurations of decisions one could make.

Many questions involve statistical points that have been subject to much exposition and debate, and my purpose here is not to get into the specifics of those. Instead, I wish only to reflect upon the general strategy for dealing with the rapidly multiplying decision forks in quantitative studies. My basic counsel is that, with each decision, one should figure out first what one thinks is the right way of proceeding given the purpose of the study. At the same time, however, one should consider the consequence of different ways of doing the analysis, especially for decisions that seem major or especially debatable.

Most emphatically, the point of doing the analysis alternative ways is not that one should reconsider decisions when they turn out to make a difference. It is intellectually dishonest to use one's decision latitude in quantitative research to simply hunt for the results that one wants. Plus, post hoc decisions are all too easy to rationalize after the fact. One gets a result one doesn't like, then runs it another way and something more desirable and then—not necessarily disingenuously—"realizes" there is some reason why the first analysis was a mistake and the second way is right. This is dangerous reasoning, and certainly makes the *p*-values and significance tests reported in one's analysis misleading.[2] Instead, the point of doing the analysis multiple ways is that readers will often wonder whether particular decisions are consequential or not, and one can pre-empt their questions by telling them. Still, the infeasibility of looking at all

[2] *p*-values can be misleading for various other reasons I cannot get into here, but the heuristic value of *p*-values is especially undermined by making post hoc decisions and then reporting *p*-value and significance tests as though those decisions were a priori.

configurations of possible decisions means that the usual strategy is to look at the consequences of different ways of making any one decision when all the other decisions are made according to what one has decided is best.[3]

What is the right decision? When a study is testing a theory, decisions should, to the maximum extent possible, follow from the substantive implications of the theory. In our study, we were testing Sulloway's theory, and we sought to think through the implications of his theory as he articulated it in *BTR*. Because the GSS did not provide information sufficient to draw implications about whether stepsiblings were part of the child's early childhood environment, we decided simply to exclude them from our analysis, although we also looked at the implications of this decision. (If models are correctly specified, excluding cases based on explanatory variables does not bias estimates for the more restricted sample that remains; by contrast, excluding cases based on values of the outcome—a.k.a., "sampling on the dependent variable" (see King, Keohane, and Verba 1994: 129-135)—does bias results.)

In general, my experience is that when people are trying to answer abstract research questions rather than speak to specific theories or at least ideas about processes, decisions about how to do the analysis become much harder. If we had been trying to answer "What is the effect of birth order on attitudes?," questions of, for example, what to do about only children or children with a much older siblings cannot be answered straightforwardly, as different answers follow from different ideas about *why* birth order might affect attitudes. For example, there exist

---

[3] This point does not deny the use of model fit (e.g., R-squared) as a criterion for some matters, perhaps such as how a control variable should be measured. One can decide that the right decision before running the models is to make it on the basis of model fit. Again, what is wrong is to run the models, look at the results, and then choose the results one prefers and rationalize this result on the basis of model fit.

biological theories of birth order effects that attribute effects to differences in the mother's reproductive system with successive births; these imply that older siblings who died in infancy should still be counted and thus make the second child a laterborn. By contrast, an environmental theory like Sulloway that attributes birth order to interactions among siblings and their parents would suggest that children who die in infancy should not count. Indeed, if there was evidence of birth order effects generally, these would be competing theories, and the case of children whose only other sibling died in infancy would be one way of testing between them.

In practice, the vague character of much theory in social science--even for theories that are "formalized"--only goes so far toward providing direction for the many decisions made during a survey analysis project (e.g., Raftery 1995). Having some principled grounds based on a substantive or statistical reasoning is better than just acting arbitrarily. Following conventional practice in an area can serve as grounds, and is especially helpful in contexts in which one frames work as following an established literature except for some key innovation that is the intended contribution of the paper. (The political scientist Gary King [2006] has said this is one of the best strategies for graduate students to get an early publication that makes a genuine contribution.) I worry about research that is slavish to conventional practice in ways that can seem intellectually lazy, but unexplained deviations from conventional practice are easy targets for reviewers. Researchers often refer to including "the usual suspects" covariates (a set that includes at least age, sex, race, educational attainment, and some income measure), many times without any additional explicit rationale for doing so, and omitting one of these variables as a covariate probably can prompt a reviewer to wonder why. Reviewers find enough unpredictable things to object to, and too many reviewers also can allow one or two readily addressed problems color their entire reaction to a paper, so one wants to be cautious to close down as many obvious

lines of objection as one can. More broadly, writing successful research articles requires being able to take the perspective of a reviewer when reading the article yourself. You do not get the opportunity to answer questions for the reviewer, or elaborate points to clear up misunderstandings, and so the text needs to read well to a reader who may not be giving it complete attention, and may not understand the methodological and substantive issues as well as you do. [4]

Our paper estimated the effect of birth order on different social attitudes using several different model specifications. This allowed readers to see how estimates were affected by the addition of different covariates. In our case, as I will discuss more shortly, results did not change as new variables were added. When results are substantively affected by the addition of covariates, one wants to be sure to know what covariates yield the change and consider why (in the abstract, covariates change the coefficient of a key explanatory variable when they are associated with *both* the explanatory variable and the outcome, and the exact change depends on the product of the direction and magnitude of both associations).[5]

Just so we are clear, when we are using regression analysis as a tool for testing theories, the reason why one includes covariates is usually to attempt to adjust for sources of spuriousness (or "confounding") that would otherwise bias our interpretation of the influence of the theorized

---

[4] Helpful for this, of course, is having other people read one's paper. The most useful readers in this respect are those who are as much as possible like those who would be candidate reviewers for the paper. Worth emphasizing again about the feedback one receives from readers is that any elaborations or clarifications you make them afterward is exactly the kind of interaction you will not get to have with a reviewer—one needs to *pre-empt* concerns, not have ready answers to them.

[5] Sometimes researchers will show that the bivariate relationship between two variables is dramatically reduced when several covariates are added to the model, but then not give any indication which covariates were actually responsible for the change. I always find this unsatisfying.

explanatory variable on the theorized outcome (much can be said about the limitations of this approach, see Berk 2004). For example, in the case of birth order, an obvious confound is family size. Sulloway's theory concerns differences between firstborns and laterborns. A family with two children has one firstborn and one laterborn. A family with six children has one firstborn and five laterborns. Consequently, the average laterborn has more siblings than the average firstborn. If family size is itself causally related to an outcome and one's study is based on a sample of unrelated individuals--about which I will say more later--this will make birth order correlated with that outcome even if birth order *per se* has no influence. As a result, we need to adjust for family size, so we are effectively comparing firstborns and laterborns with the same number of siblings. Using regression methods and including family size as a covariate (or "control") is a strategy for adjusting for the potential spuriousness of family size. Note that a very closely related strategy would be to restrict analyses to only respondents with only one sibling, or run separate analysis for respondents with one sibling, two sibings, three siblings, etc.. (We did this as well, and found the same results as in our combined analyses.)

To take a somewhat different example, in the birth order literature, it is commonly argued that one also needs to adjust for the socioeconomic status of the respondent's families. Sulloway's and other attempts to quantitatively summarize ("meta-analyze") the birth order literature has typically excluded studies that do not adjust for both family size and socioeconomic status. The argument goes that lower socioeconomic status is associated with family size, which means that laterborns are more likely to come from lower socioeconomic status families than firstborns. This argument does not actually follow: if one already adjusts for family size, then this solves the problem without also needing to adjust for parents' socioeconomic status. Issues like this arise regularly in quantitative research, where

conventional ways of conducting research are either not optimal or not correctly justified. Practically speaking, one is put in a position of having to choose one's battles. In this case, there is no reason to expect that controlling for socioeconomic status of one's family would bias results, and it turns out that one gets similar results whether or not it is controlled. In our study, we included socioeconomic status in the models we presented, rather than risk failing to persuade reviewers that controlling for socioeconomic status was unnecessary.[6] However, if an alternative way of doing the analysis both seems more justified and does produce different substantive conclusions, then obviously the alternative analysis is what one must present, even if relying on convention would be easier.

Importantly different is the idea of adjusting for respondents' *current* (i.e., adult) socioeconomic status. In social science, one will commonly see researchers "control" for variables determined after the key causal variable has been established. We did exactly this in our study, by including respondents' education and income in what we present as our "final model." If birth order affects respondents' socioeconomic status and that status, in turn, affects social attitudes, then socioeconomic status is a mediating variable—part of one of the *mechanisms* by which birth order affects attitudes—rather than a source of spuriousness. What we say in our paper is "To account for the possibility that observed birth-order effects may be caused by birth-order differences in achievement, our final model adds controls for the respondent's education and occupational prestige" (p. 215). Sulloway's theory is quite clear that

---

[6] For that matter, one can make a reasonable argument why controlling for socioeconomic status is still a good idea, e.g., if one is worried about either measurement or specification error in the family size measure. Also, if it is unnecessary, the expectation is that it will have no effect on the estimated effect of birth order (which is mainly what we observed), not that estimates will be biased.

the birth order effects shape attitudes by shaping psychology, not by shaping social attainment outcomes that later shape attitudes, so this wording is correct. Still, I wish that we had been clearer that, when including these measures, we had moved from estimating the total causal effect of birth order to including covariates that were actually potential mediators of the effect of birth order on attitudes. Sometimes people approach regression analysis as though the appropriate strategy is to include everything at one's disposal as a covariate; variables that are consequences of the key explanatory variable are not "controls" and should only be included in an analysis when the researchers intends to be isolating the effects of mediators of the relationship.[7]

Covariates in regression analysis are commonly used as a strategy for "ruling out" other explanations; for example, that an observed bivariate association between birth order and an outcome is really the result of a spurious relationship with family size. Important to emphasize about this strategy is that it only works to whatever extent the proposed source of spuriousness is accurately and fully measured. There are many social science research questions, for instance, for which researchers believe that socioeconomic status poses a very plausible confound that absolutely must be taken into account. This is commonly done by including a measure of the respondent's educational attainment and family income (or, the respondent's parents' educational attainment and income). To whatever extent what is relevant about "education" is not totally captured by response to survey reports of highest degrees attained, and what is relevant about "financial resources" is not totally captured by response to survey reports of annual family

[7] A separate consequence of this is that regression analyses are much more readily interpreted when considered in terms of attempting to estimate the effect of a single explanatory variable rather than imagining that one is estimating the causal effect of all the explanatory variables in the regression at once.

income, and what is relevant about "socioeconomic status" is not exhausted by "education" and "financial resources", this strategy will not fully control for the spurious influence of socioeconomic status. If the coefficient of one's key causal variable is partly diminished by the inclusion of a faulty control, one would expect it to be further diminished had the control been more completely measured. Ways of correcting for measurement error in control variables exist (although still involve some technical debate when the outcome variable is not continuous). My hope is that social research that relies on the strategy of providing statistical controls for sources of spuriousness will make greater use of these techniques in the future.

**ESTABLISHING FINDINGS**

What quickly became clear when I started looking at the relationship between birth order and social attitudes using GSS was that there did not appear to be any relationship. Powell had expected this; I was enchanted enough by Sulloway's book that I had thought there was a good chance our study would support his theory. In any case, what we had were "null findings," the general term for results in which a theoretically predicted relationship between an explanatory variable and an outcome variable is not observed. Null findings are said to be much harder to publish, especially in prominent venues, than are studies that find positive support for a theory. I have no doubt this is true, and yet our study with null findings was successfully published in the *American Sociological Review*. How?

In order to publish null findings, you need to do two things. First, you need to establish that there are people who believe the theory that your null findings call into dispute. We could argue that Sulloway's book had made the long-studied topic of birth order timely again, and

selections from the extremely positive press the book had received could be used as grounds that the theory was one some people believed. Second, you need to establish that your study really does provide a good test of the theory. Null findings are exactly what one expects if measures are so weak or biased that they are effectively meaningless, or if the sample is small and the effects observed are of a magnitude that would be statistically significant in a reasonably-sized sample. Another way of putting the point is that you need to establish that your findings really are null rather than just equivocal.[8] Here, I do not think null findings are especially different from positive findings: the researcher needs to establish that one's findings are *news to someone* and that they are *credible*. The greater the combined success on these two fronts, the greater the ultimate prospects of the paper.

Toward making our own results as credible as possible, we never considered looking at only one or two measures of social attitudes. If we did, then no matter what we found, readers might wonder why we didn't look more broadly given all the measures of social attitudes available in GSS. They might worry we were only showing them results for the measures that happened to fit our conclusions. Instead, we considered 24 measures that we had decided beforehand most closely reflected the dimensions of attitudes on which Sulloway focused in *BTR*. We selected these 24 measures by first identifying what we regarded as the six broad areas of attitudes that were both given much attention by Sulloway and had coverage in GSS: conservative political identification, opposition to liberal social movements, resistance to racial reforms by Whites, belief in traditional gender roles, support for authorities, and "tough-mindedness." We then chose what seemed the most apt measures available in GSS for each area.

---

[8] In this last respect, null findings are probably easier to publish than findings that are highly ambiguous as to whether they do or do not support a theory.

Thus, for each measure we used, we could provide a justification why it followed from some part of Sulloway's argument.

In our models that included all controls, birth order was significantly associated with only three of these 24 measures (p. 216). Moreover, for all of these three measures, the association between birth order and the attitude was actually *opposite* the direction predicted by Sulloway's theory. For instance, despite Sulloway writing a chapter on how firstborns are more tough-minded—reflected, as one example, in the disproportionate number of firstborns who voted for the execution of King Louis XVI—firstborns in the GSS were actually less likely to support the death penalty than were laterborns.[9]

Even here, we were worried that we could be open to the charge that we simply picked measures of social attitudes that happened to support a particular conclusion. So, in addition to looking at these 24 measures, we looked at all 202 measures in GSS that could be interpreted as tapping into liberal-conservative differences. We found that again that there was no pattern of significant results supporting Sulloway's theory, and in fact the number of significant results and number of results in the predicted direction were roughly in line with what one would expect by chance alone (p. 221).

All that said, I was bothered that our study (as well as most of Sulloway's) was ultimately deficient because it compared firstborns with laterborns from other families. Stronger, certainly, would be to compare firstborns and laterborns from the *same* family, as that strategy, known more generally as a fixed-effects design (see Halaby 2004), would solve problems caused by

---

[9] Not that our conclusion is that firstborns and laterborns really differ in their attitudes about capital punishment: given that one expects roughly one result out of 20 to be significant by chance alone if one is using a significance level of $p < .05$, three out of 24 measures being significant seems like it may also just reflect chance.

differences between respondent families of origin by only comparing individuals with their own siblings. As I learned by following up on an obscure footnote in the GSS codebook, the 1994 GSS included a companion study, the Survey of American Families (SAF), of telephone interviews with a randomly selected sibling of GSS respondents. This allowed us to compare siblings from the same family for those cases where one of the selected respondents was a firstborn and another was a laterborn. This comparison was not without problems of its own: the SAF did not have as good a response rate as the GSS, and the survey included fewer measures, but that analyses of SAF also yielded null results increased my confidence that our findings were correct and strengthened the presentation in our paper. An advantage of drawing on multiple data sources in the same paper is that concerns about the imperfection of each may be strengthened if they yield similar or complementary results.

Worth reflecting here is that many questions that social researchers study are fraught with moral and political implications, and often the people who choose to study a question are some of the people who have the strongest opinions on the topic. I did not have any psychological investment in Sulloway's theory being true or false, but that isn't to say a reader could not wonder if our findings partly reflected some secret determination we had to contradict his work. Here is where the secondary analysis of publicly available data shows one of its greatest strengths. Recall that *BTR* was based on remarkable data that Sulloway himself compiled over many years. I have talked to some people who have cited all his labor as *grounds for skepticism*: since Sulloway was himself responsible for so much of the data collection, perhaps some of the decisions and actions during the decisions resulted in data that were biased in favor of his theory (see Harris 1998). One does not have to suspect *conscious* bias to raise this possibility: to be absolutely clear, I regard Sulloway as a remarkably conscientious and committed scholar.

Sulloway indeed did various checks against undue subjective influence, but some of these were not fully convincing. As one example, a key way that Sulloway argues that his study was insulated against the possible influence of his own biases was that many ratings were obtained by independent experts. Many of these ratings, however, were obtained in interviews conducted by Sulloway himself, and this could have opened the possibility of his participation affecting the resulting ratings in various ways.

Contrast this with our study. We did *far less* work than Sulloway but had an important advantage *precisely because* of that. For we had nothing to do with the collection of the 1994 General Social Survey—all we did was download it. Someone disinclined to believing our results cannot argue that the data on the GSS website have been contaminated by any biases of ours against Sulloway's theory. The data are publicly available if anyone wonders whether we reported our results accurately. Although the steps used to generate results are straightforward and described in the paper, I also have the computer code that I used to generate the results and can share it with anyone who asks. From this, a person can reproduce the numbers in the paper, meaning that one does not have to take our word for that the numbers in my paper really did come from an analysis of the General Social Survey.[10]

People are skittish around words like "objective" for good reason, but how much more "objective" can results get than being able to hand someone exactly the same data that one used—data one had no hand in collecting—and saying: *here, I beseech you to look at these data*

[10] Besides, as already noted, I looked at GSS as comprehensively as I was able, which strengthens my own confidence that there is not some defensible alternative way of doing the analysis that would have yielded different conclusions from those we drew.

*yourself and tell me how one might reasonably arrive at any other conclusion than mine*.[11] In sum, a major strength that helps quantitative research on large surveys to speak effectively to policy questions is that researchers can show precisely and compellingly how their conclusions follow from available data and thereby limit the degree to which others can dismiss their work as just one person's (misleading, wishful, ideologically biased, etc.) interpretation.

Importantly then, when conducting quantitative analysis of secondary data, one should conduct analyses in such a way as to have computer code that reproduces all the steps necessary to go from an original dataset (e.g., the GSS file I downloaded) to all the numbers presented in the paper.[12] Statistical software packages that allow one to conduct analysis using point-and-click menus still typically generate code that allow one to automatically reproduce the commands being executed by the menus. Any software that does not allow one to do this should be avoided. Major economic journals presently require authors to deposit this code (and also the data if they have the rights to do so) at the time their article is accepted for publication, which is then posted on the Internet. My hope is that other social sciences will follow suit with some version of this practice (Freese forthcoming). If a key strength of secondary quantitative analysis is that findings are less open to subjective biases--or, at least, that the influence of subjective biases is easier to establish because of the potential for other researchers to scrutinize the same data--then such analyses are especially convincing when work makes the most of this advantage by making the procedures used as transparent as possible.

---

[11] Or, alternatively, why the data are inappropriate for the conclusion drawn from them.

[12] The code needs to be carefully documented, so one can figure out what one is doing. I have gotten much better at this with time, less because of increasing skill but because of greater appreciation of the importance of replicability and greater humility about the fallibility of my memory.

## CONCLUSION

We live in an age of rapid technological change, and large-scale survey analyses may stand to benefit more than any other common social research methodology in how ongoing changes will expand and strengthen our ability to learn new things. Repeated investments in major survey instruments—both those based on asking similar questions to new cross-sections of the population in different years (like the GSS), and those based on tracking the same individuals over time (longitudinal studies like the National Longitudinal Studies of Youth or the Health and Retirement study) allow these studies to afford stronger and more comprehensive research each time they are fielded. Ubiquitous digital record-keeping has created all kinds of novel possibilities for matching survey and administrative records to these data. For example, the Health and Retirement Study allows researchers to use data from the Social Security Administration on earnings and from Medicare claims data regarding health care utilization.

Advances in Internet surveying will allow for the possibility of more frequent and much cheaper collection of repeated measures, affording possibilities for a much more fine-grained understanding of the unfolding of biography than even those studies that talk to the same respondents every year or two (although the Waleiko and Williams papers in this volume note some challenges to these approaches). Advances in the collection of biomarkers portends possibilities for integrating studies of biological, psychological, and sociological processes (see, e.g., the Adam paper in this volume). Advances in the collection and manipulation of spatial data will allow better work on the effect of environments on individuals' lives. Sure, there will always be buzzkills who regard disciplined research with numbers as anathema to their personal

aesthetic, but exciting possibilities lay ahead in quantitative research involving large-scale survey resources.

This is not to say all is bright. Surveys have suffered from declining response rates in recent years, for a variety of reasons (one of which is that the amount of surveying in society has increased so much that, even if individuals spent the same amount of overall time participating in surveys, the response rate of individual surveys would still be lower). Surveys are improving in their understanding of how to collect data to allow for the best possible inferences in the face of nonresponse (Groves et al. 2002), and estimation techniques for working with nonresponse are becoming increasingly well-incorporated into regular data analysis practice (King et al. 2001). Especially as their work has been more infused with cognitive psychology, survey methodologists have discovered all kinds of ways in which survey responses are sensitive to the wording of questions and responses, their placement in the larger survey, and the mode and context in which the interviews are done (Tourangeau, Rips, and Rasinski 2000; Stone et al. 2000). This work will ultimately make surveys stronger, although part of the short-run consequence is to prompt wondering whether we really know all we thought we did—but this is itself an opportunity for new research. Some work I have collaborated in, for example, suggests that many of the subscales of a leading measure of psychological well-being seem actually, once proper statistical adjustments are made, to be largely indistinguishable from one another, raising the possibility that a whole preceding literature of studies finding distinctions among these subscales has been mostly just chasing chance variation (Springer, Hauser, and Freese 2006).

In addition to technological advances, social science is also making methodological strides that are improving the craft. What may be ultimately of most lasting importance here, perhaps, is not what is now happening on some statistical or computational frontier, but rather

the steady elaboration and diffusion of a clearer understanding of how quantitative research interested in identifying causes or predicting the consequences of policy manipulations should think about research questions. Here, social scientists are becoming increasingly adept with thinking about research questions in *counterfactual* terms: that is, thinking about quantitative research on causes as asking about whether and how our expectations about an outcome would be different if the key explanatory variable had been different (Morgan and Winship forthcoming). So, a study of the effect of attending an elite college on later earnings, for instance, is asking how a given person's earnings would be different if they had gone to a non-elite college rather than an elite one (see Brand and Halaby 2006). This, in turn, allows clearer thinking about the appropriate comparison to substantiate any attempted answer to this question, which helps move well beyond simply taking some large dataset, putting a bunch of variables into a regression routine, and looking at the coefficient of the ELITECOL variable. For example, it leads one to question how much value there is to including in one's analysis people whose background and academic record together give them practically no chance of attending an elite college, as these people are so different from those who actually attend elite colleges that the idea that one is estimating a causal effect by comparing the two seems implausible.[13] Some try to expand the heuristic value of thinking about counterfactuals and "potential outcomes" into a full-fledged philosophy of causality (Holland 1986), which runs aground quickly in various ways

[13] To elaborate, one is on much stronger for estimating the effect of attending an elite college if one compares those who attend elite colleges to those are like them, but do not (ideally, for completely random reasons—that this ideal is implausible is the great challenge for research using non-experimental data). In this case, one is estimating the average effect of attending an elite college for those who did attend an elite college, which may be very different from the average effect of attending an elite college for students as a whole. Only by (often implausible) extrapolation can we claim to estimate the causal effect of an event on a group of people if that event has happened to practically no one in that group.

(e.g., Glymour 1986). Practically, though, counterfactual thinking seems an extremely useful cognitive tonic for clarifying various kinds of muddled thinking that have chronically beset many areas of quantitative social research.

As researchers have become more sophisticated in their thinking about causality, this has increased awareness of just how hard it is to draw strong causal inferences from ordinary single surveys, as we used in our study utilizing just one year of GSS. Moreover, developments in this regard have underscored the limited gains from fancier statistical techniques in comparison to improvements in the quality of data. A recent movement in economics and in policy evaluation research has sought increasingly to isolate instances in which data would afford a relatively clear inference about causes in that situation, even at the expense of broader generalizability.

Two strategies, pervasive in economics but only intermittently considered in the quantitative methods curricula of other social sciences, deserve especial mention. One, which I call *inference from discontinuity*, involves identifying a point in a series of data at which we would expect, if a given cause is operating, a relatively discrete change in the observed data. The most obvious application of this is time: if the cause is an event, then we would predict a discrete change just after compared to just before the event occurs. We can use familiar techniques to model any larger background trend over time, and then specifically estimate the change associated with the event. In the case of estimating effects of widowhood on health, for instance, we can estimate the overall trend of aging and then look at whether there is a discrete difference before and after the spouse's death. Inferences from discontinuity are not limited to time, though, as different policies may lead us to predict discontinuous changes immediately above and below specific levels of income (Berk and Rauma 1983), different school district boundaries (Black 1999), or different test scores (Van der Klauuw 2002). When evidence over

multiple cases reveals a sharp discontinuity at a point for which it is hard to imagine some explanation other than the postulated cause, this can serve as quite compelling causal evidence.

The other, which I call *inference from exogenous variation* (or inference using "instrumental variables"), involves trying to isolate a "natural experiment" within some larger set of data and trying to estimate causal effects for that. Consider trying to estimate the effect of family size on educational attainment. This is a notoriously difficult problem because parents who choose to have greater or fewer children differ from one another in ways that are also likely related to how much education their children receive. The strategy of simply controlling for confounding variables is only satisfactory if one really believes all the pertinent differences are accurately and fully captured by the variables included as controls. An entirely different strategy may be derived from the finding in other research that there are some parents who prefer having a child of each sex enough that they will have a third child if they do not get one of each with their first two births. This creates a natural experiment: if having a son or daughter is basically random, some of these parents will have a son and daughter and then stop, whereas some will have two sons or two daughters and then have a third child. If we could somehow isolate these parents, we could estimate the effect of two-child versus three-child families on educational attainment in a way that would not be confounded by background differences across families (since, at least as posited here, it is random—*exogenous* to any parental characteristics—whether the first two children are opposite or same sex, and so random whether these parents on the margin stop or try to have a third child). If we are willing to defend certain assumptions, however, we do not actually have to isolate these families. Instead, putting things simply, we set aside the information on how many total children there are in the family, and we just compare the first two children in cases in which they are both boys or both girls to cases in which they are

one boy and one girl. If we believe the only reason the educational attainment of the former would be different than the latter is that the former are more likely to have additional younger siblings that causally affect educational attainment, then we can rescale this difference to produce an estimate of the effect being in a three-child versus two-child family (see Angrist and Evans 1998; Conley and Glauber 2006 for examples; see Imbens and Angrist 1994 and Morgan and Winship forthcoming for a discussion of these issues in a counterfactual framework).[14]

Perhaps we can imagine other plausible explanations why the educational attainment of same-sex sibling pairs would be different from opposite-sex pairs. Perhaps there is greater competition for parental resources in same-sex pairs in ways that affect achievement.[15] If so, then this would bias our estimates and so not provide an effective strategy. Other instances are harder to devise alternative explanations. In trying to estimate the effect of military service in Vietnam on earnings, we know there are some men who only served in Vietnam because their birthday was selected high in the draft lottery and other men who did not serve but would have if their birthdays had been selected high instead (Angrist 1990). If we observe that men in those cohorts whose birthdays were on March 15, June 9, August 25, September 24, and December 11 have higher average earnings than those born on March 14/16, June 8/10, August 24/26, September 23/25, and December 10/12, and if the former dates were selected high in the draft lottery and the latter were not, then what other explanation would we have other than that the

---

[14] Of course, what one is really estimating is the effect of being in a three-child versus two-child family for those families where the sex composition of the first two children determine whether or not the family has a third effect. Whether the effect can be generalized to three-child versus two-child families more broadly is a question of external validity that is hard for these designs to address and is often underappreciated in evaluating studies using instrumental variables designs.

[15] As a different possible concern, Currie and Yelowitz (2000) use sibling sex composition as an instrument for public housing because public housing supplements in some areas favor families with two opposite-sex children over two same-sex children (giving the former an extra bedroom).

earnings of the former group are higher because of their higher probability of being drafted? The inference is compelling to whatever extent we imagine that the two groups (same vs. opposite-sex sibling pairs; individuals with one set of birthdays vs. another) otherwise would be the same and so the only reason we would observe a difference because of the effect of the "natural experiment." Other examples include using election years as a source of exogenous variation in the size of city's police force (Levitt 1997), exogenous variation in number of rivers and streams for the number of school districts in an area (Hoxby 2000); exogenous variation in how judges are assigned for the length of prison sentences (Kling 2006) (see Angrist and Krueger 2001 for a list of examples). Inference from exogenous variation is tricky and its virtues can be oversold, but it is attractive because if offers the possibility of compelling inferences in cases that would otherwise seem hopelessly ambiguous.

In closing, I will admit I have ambivalence about how much of my work has relied on large-scale survey data. I find surveys often frustrating to work with for how little we can know about the contours of an individual case, and even extensive longitudinal surveys provide only very indirect insights into processes of individual development or social life. Still, however, they are indispensible tools for characterizing populations, and clear-eyed and conscientious survey research has afforded all kinds of subtle insights into the workings of social life not otherwise available. Plus, survey technology continues to improve, as are techniques for analyzing survey data. Even so, myself and many other social scientists find that a comprehensive research agenda involves seeing surveys as a complement to other kinds of research. For carrying forward such an agenda, the three most invaluable resources seem to be an open and enthusiastic mind, broad training, and collaborators.

**REFERENCES**


Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80:313-336.

Angrist, Joshua D. and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15:69-85.

Angrist, Joshua D. and William N. Evans. 1998. "Children and their parent's labor supply: evidence from exogenous variation in family size." *American Economic Review* 88:450-77.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Berk, Richard and David Rauma. 1983. "Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program." *Journal of the American Statistical Association* 78:21-27.

Black, Sandra E. 1999. "Do 'Better' Schools Matter: Parental Valuation of Elementary Education," Quarterly Journal of Economics 114: 577-599.

Brand, Jennie E. and Charles N. Halaby. 2006. "Regression and Matching Estimates of the Effects of Elite College Attendance on Education and Career Achievement." *Social Science Research* 35:749-770.

Conley, Dalton and Rebecca Glauber. 2006. "Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility." *Journal of Human Resources* 41:722-737.

Currie, Janet and Aaron Yelowitz. 2000. "Are Public Housing Projects Good for Kids?" Journal of Public Economics 75: 89-124.

Freese, Jeremy. Forthcoming. "Replication Standards in Quantitative Social Science: Why Not Sociology?" To appear in *Sociological Methods and Research*.

Freese, Jeremy and Brian Powell. 1999. "Sociobiology, Status, and Parental Investment in Sons and Daughters: Testing the Trivers-Willard Hypothesis." *American Journal of Sociology* 106:1704-43.

Freese, Jeremy, Brian Powell, and Lala Carr Steelman. 1999. "Rebel Without a Cause or Effect: Sociobiology, Birth Order, and Social Attitudes." *American Sociological Review* 64:207-231.

Glymour, Clark. 1986. "Comment: Statistics and Metaphysics." *Journal of the American Statistical Association* 81:964-966.

Groves, Robert M., Don A. Dillman, John L. Eltinge, and Roderick J. A. Little. 2002. "Survey Nonresponse." New York: Wiley.

Halaby, Charles N. 2004. "Panel models in sociological research: Theory into practice." *Annual Review of Sociology* 30:507-544.

Harris, Judith Rich. 1998. *The Nurture Assumption*. New York: The Free Press.

Holland, Paul. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81:945-60.

Hoxby, Caroline M. 2000. "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review* 90:1209-1238.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62:467-75.

King, Gary. 2006. "Publication, Publication." *PS: Political Science and Politics* 39:119-125.
King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95:49-69.
Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96:863-76.
Levitt, Steven D. 1997. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime." *American Economic Review* 87:270–90.
Morgan, Stephen L. and Christopher Winship. Forthcoming. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge University Press.
Murray, Michael P. 2006. "The Bad, the Weak, and the Ugly: Avoiding the Pitfalls of Instrumental Variables Estimation." Available at SSRN: http://ssrn.com/abstract=843185.
Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-163.
Shermer, Michael. 1996. "History at the Crossroads: Can History Be a Science? Can it Afford Not to Be?" *Skeptic* 4:56-67.
Springer, Kristen W., Robert M. Hauser, and Jeremy Freese. 2006. "Bad news indeed for Ryff's six-factor model of well-being." *Social Science Research* 35:1119–30.
Stone, Arthur A., Jaylan S. Turkan, Christine A. Bachrach, Jared B. Jobe, Howard S. Kurtzman, and Virginia S. Cain. 2000. *The Science of Self-Report: Implications for Research and Practice*. Mahwah, NJ: Lawrence Erlbaum and Associates.
Sulloway, Frank J. 1996. *Born to Rebel: Birth Order, Family Dynamics, and Creative Lives*. New York: Pantheon.
Tourangeau, Roger, Lance J. Rips, and Kenneth A. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.
Van der Klauuw, Wilbert. 2002. Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach. International Economic Review 43(4): 1249-87.