11

Data Processing Multiverse Analysis of Regnerus and Critics

Now we put together many dimensions of data processing into a comprehensive multiverse analysis. We reanalyze a now infamous study by Mark Regnerus, which was subjected to many criticisms that have featured centrally in Chapter 10: data processing and the potential for dark matter in the research design to be driving the empirical findings.

Is it harmful for children to grow up with gay or lesbian parents? Publishing in the journal *Social Science Research*, Regnerus (2012a) found that the children of lesbian, gay, bisexual, and transgender (LGBT) parents, compared to those raised in "intact biological families" (IBFs), were worse off in many sociodevelopmental ways: They were more likely to be unemployed, be on public assistance, have drug abuse problems, have contemplated suicide, and more. Others have reanalyzed the data and argued that the results depend on analytic choices – particularly on data processing decisions (Cheng and Powell 2015; Rosenfeld 2015).

This article is one of the most hotly contested studies in twenty-first-century sociology. It has been invoked in Supreme Court debates and rulings on same-sex parenthood and adoption rights. An external review at *Social Science Research* considered retracting the article (Sherkat 2012). Many scholars have questioned nearly every detail of the study's data, assumptions, and analysis. We draw on this postpublication discourse to develop a multiverse analysis of the original study.

A widely debated study like Regnerus (2012a) is ideal for this analysis, because a compelling data processing multiverse requires a great deal of specific input from expert analysts – who both identify contested aspects of processing and propose alternative specifications that may be equally or more credible. We construct two distinct multiverse analyses: a control variable multiverse and a data processing multiverse. This allows us to address a central question for this chapter: How do "researcher degrees of freedom" from data processing compare to those stemming from choice of controls? How much variation in results do each of these sources of uncertainty generate?

Our baseline model specification is replicated from Regnerus (2012a), and we draw on two published articles for criticism and alternative specifications (Cheng and Powell 2015; Rosenfeld 2015). We add to this additional processing alternatives we identified through close replication of all three articles. Our full multiverse includes over 2.6 million unique model specifications.[1]

THE STUDY

Regnerus's study collected a new dataset called the New Family Structures Study (NFSS), using an online survey administered by the data company Knowledge Networks. About 15,000 people took a screening survey, answering the question "Did either of your parents ever have a romantic relationship with someone of the same sex?" Response choices were "Yes, my mother had a romantic relationship with another woman," "Yes, my father had a romantic relationship with another man," or "No." The full survey was taken by 2,988 people, including all of the 236 people who indicated in the screener question that one of their parents had been in a gay/lesbian relationship.

Regnerus's original regression was approximately

$$\begin{aligned} y_{ij} = \alpha + b_{1j}\text{gay\_father}_i + b_{2j}\text{lesbian\_mother}_i \\ + [\text{other family types}] + [\text{controls}] + \varepsilon_{ij} \end{aligned} \qquad (11.1)$$

for i individuals and $j = 1, \dots, 40$ unique outcome variables. The reference category is the IBF – people who grew up with both biological parents and whose parents were still alive and still married at the time of the interview. Other family types include single parents, stepparents, divorced parents, and other departures from the two biological parents reference group. The forty outcome variables include educational attainment, economic wellbeing, health outcomes, mental health outcomes, civic engagement, substance abuse, family arrangements, and criminal activity. Some are "positive" outcomes (closeness to biological mother/

[1] The replication package for all of the analyses presented in this chapter is available online at https://osf.io/45ft2/.

father) while others are "negative" (unemployment). Others are not obviously positive or negative, such as being in a homosexual relationship or frequency of TV watching. For each outcome, effects were estimated for lesbian mothers and gay fathers separately. This original formulation produces an unwieldy amount of regression output. With forty outcomes and two categories of LGBT parents there are eighty coefficients of interest, before conducting any robustness tests. We take two steps to streamline this output down to one summary coefficient.

First, following Rosenfeld (2015), we combine the many outcome variables into a single index. Rosenfeld used nineteen outcomes that were unambiguously positive or negative and did not have large numbers of missing values. To create the index, he reverse-signed the positive variables to make an index of negative outcomes, then standardized the variables so they were on the same scale (in which positive values correspond to more negative outcomes, counterintuitively). We rebuilt this index so that (1) the index is correctly signed (positive values indicate better outcomes), (2) we include more outcome variables (twenty-nine versus nineteen), and (3) we more flexibly account for missing data.[2] Compared to Rosenfeld's index construction, this increases the number of outcome variables by 10 and increases the sample size by nearly 500 (including 45 treatment cases – which is 19 percent of the total – and 454 control cases). We believe this outcome index significantly improves on Rosenfeld's construct, but we include both as multiverse options so readers can see what difference it makes to the results.

Second, we pool together lesbian mothers and gay fathers into one group of those with an LGBT parent. Pooling together gay fathers and lesbian mothers in one model provides a weighted average of the separate effects while increasing statistical efficiency (i.e., producing smaller standard errors), as in Mazrekaj, De Witte, and Cabus (2020) who simply report results for those in same-sex parent families.[3] Through these

[2] Rosenfeld (2015) drops all respondents who are missing on *any* of the nineteen outcome variables he used (17.5 percent). In contrast, we use all available responses on the twenty-nine outcome variables we include and afterward check for missingness. We require respondents to have responses for at least ten outcome variables. With this rule, we drop only 0.77 percent of respondents. Following Rosenfeld, we continue to exclude outcomes that are normatively neutral (such as being in a same-sex relationship).

[3] We note that in the baseline model, both these effects are negative and the effect of lesbian mothers is larger in magnitude than that of gay fathers. However, this heterogeneity is not analytically important for our reanalysis, and indeed none of the studies we reviewed in this debate actively discussed or interpreted the different effects of having a lesbian mother versus having a gay father.

two steps, we summarize the eighty coefficients into a single model that provides one summary estimate of the effect of having an LGBT parent. Hence, our reference model is

$$\begin{aligned}\text{index}_i &= \alpha + b_1\text{LGBT\_parent}_i \\ &\quad + [\text{other family types}] + [\text{controls}] + \varepsilon_i \qquad (11.2)\end{aligned}$$

CONTROL VARIABLES

Regnerus (2012a) had selected six control variables : age, mother's education, family-of-origin income, and dummy variables for female, white, and having been bullied while growing up.[4] Cheng and Powell (2015) identify five additional controls: mother's and father's age at birth, region, metropolitan status, and a dummy for the family having received welfare while growing up. This constitutes the eleven-variable "controls only" multiverse with $2^{11} = 2{,}048$ models.

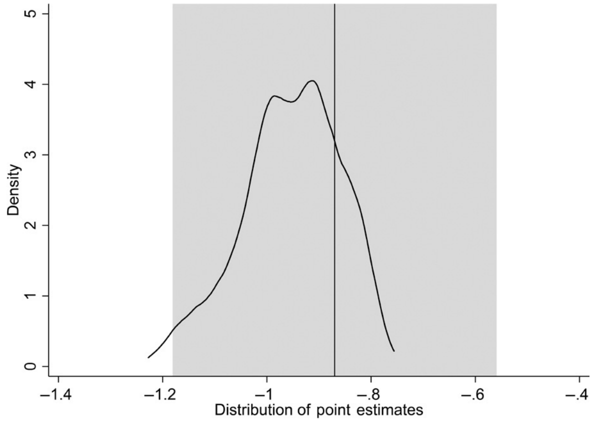

FIGURE 11.1 Distribution of gay/lesbian parenting effects in control variable multiverse

Note: Data from the NFSS, $N \approx 2{,}466$, depending on the specification. Shown are the coefficients from 2,048 models defined by eleven possible control variables. All models estimated using ordinary least squares (OLS) using sampling weights. The vertical line shows the estimate from a close approximation of the preferred model in Regnerus (2012a) and the shaded area shows the conventional 95 percent confidence interval for that estimate.

[4] The original study also included a state-level gay-friendly score, which we cannot include since the public version of the NFSS data does not include state.

Note that Rosenfeld rejects one of these controls, bullied as a child, as endogenous and does not include it. This does not affect our analysis as the default multiverse assumption is that any control could be endogenous and thus potentially excludable.

Figure 11.1 shows the control variable multiverse. The original results are highly robust to the choice among these control variables, with all of the estimates positive and statistically significant. Indeed, the "Regnerus model" (i.e., using just the six controls from that study) yields an estimate that is on the low end of this distribution. None of the control variables have notable influence. The most influential control suggested by Cheng and Powell (2015) is family receipt of welfare while growing up, and including it reduces the multiverse mean by only 11 percent (from –1.0 to –0.9). The modeling standard error is 0.09, which is smaller than the sampling standard error in the Regnerus model (0.16). Thus, the conventional 95 percent confidence interval has wider bounds than the control variable multiverse. Uncertainty about the choice of these controls has little impact on the conclusions: Using all, none, or any combination thereof produces similar results.

DATA PREPROCESSING DECISIONS

From the beginning, criticisms of this study focused not on control variables but rather on many aspects of the data and the processing strategies. Thus, our focus is on two core types of data processing: (1) treatment of anomalous observations and (2) construction of the explanatory and dependent variables.

ANOMALOUS OBSERVATIONS IN NFSS DATA

Many critics have commented on the anomalous observations that can be easily gleaned from the NFSS codebook. "Some number of respondents were likely just having fun filling out the survey"(Perrin, Cohen, and Caren 2013: 333). For example, "20 male respondents have had sex with more than 100 women, while 16 female respondents have had sex with more than 100 men … Ten respondents have been pregnant a dozen or more times" (Sherkat 2012: 1348). One man who said he was raised by a gay father also reports being seven feet and eight inches tall, weighing eighty-eight pounds, and having been married eight times (Cheng and Powell 2015).

A number of scholars have suggested that the structure of the screening survey – which asked 15,000 people if their parents had a gay/lesbian relationship – attracted pranksters to the survey. The screening survey was an economical way of identifying a sample of respondents who are rare in the general population. But a drawback is that it also informed respondents upfront that the survey would focus on a topic that is both highly personal and politically heated (at the time of the survey in 2011 marriage equality was intensely debated in national politics).

Regnerus (2012a) did not exclude any anomalous observations, leading to criticism that "data cleaning was apparently not something in the research agenda" (Sherkat 2012: 1348). Yet this echoes the debate over Jasso (1985) and how to deal with observations that seem implausible or respondents who seem unserious. Should analysts delete or down-weight these observations, or is doing so a way of filtering the data through one's own beliefs and inducing bias through sample truncation? We have argued earlier that the raw data should be included in the baseline analysis, alongside the postcleaning results, so we can understand what impact, if any, data cleaning has on the estimates.

For data cleaning decisions, we follow the close inspection by Cheng and Powell (2015), who argue that roughly 44 percent of the respondents coded as raised by an LGBT parent are misclassified or borderline cases – a substantial error rate if their assessments are correct. Specifically, they identify twenty-nine people with an LGBT parent who gave unreliable or inconsistent responses across different parts of the survey (considered misclassified). They flagged six other cases where information was plausible but suspicious (borderline cases). They also argue that in order for LGBT parents to causally influence a child's upbringing one expects them to have lived with the children. Many of those having LGBT parents co-resided with them only for short time periods. They consider fifty-three people who lived with their LGBT parent for a year or less (in twenty-four cases, never lived with them at all) as misclassified. They also flag fifteen people who lived with an LGBT parent for only two to four years as borderline cases. From this data review, we see three processing options: (1) use all data, (2) drop the cases Cheng and Powell code as misclassified, and (3) drop the borderline cases. In a response to his critics, Regnerus presented models that dropped cases where the respondent had never lived with the parent's same-sex partner, without making any adjustments for people with unreliable and inconsistent responses. We add this as a fourth processing option.

Finally, we consider a method for addressing outlier observations in the NFSS data. The outcome index follows a nonnormal distribution largely due to a long left tail: some people reporting extremely negative outcomes. We include as a multiverse option winsorizing the outcomes index and running OLS on that dataset.

VARIABLE CONSTRUCTION

A key source of disagreement between the author and critics is how to operationalize the comparison of the "treatment group" to the reference group: the family types that people grew up in. The study is a comparison of children growing up in IBF versus those raised by an LGBT parent. Those growing up in other family types fill out the classification: single parents, divorced parents, stepparents, adoptive parents, and "other." Coding up these other family types is needed to restrict the comparison to be between LGBT families and IBFs. Yet there are a host of problems with this construction.

Outside of the reference category of IBF, the family types are not mutually exclusive. People coded as growing up with an LGBT parent "may quite plausibly have been in any one of the other categories as well, and indeed most of them probably were" (Perrin et al. 2013: 331). However, the family type variable sums to one for every individual; anyone with an LGBT parent is coded as LGBT only, even if they also fit into the other categories. Treating these categories as mutually exclusive is confusing and potentially biasing if those with gay/lesbian parents also went through divorces, spent time in single parent families, or had step- or adoptive parents.

Cheng and Powell (2015) relax the constraint that IBFs still be married at the time of the survey. In effect, they argue that people raised by gay or lesbian parents ought to be compared to all people raised by their biological parents, regardless of whether the parents divorced after age eighteen. This expands the size of the reference category (IBF) from 919 to 1,191 people, and Group 3 (divorced after the respondent came of age) is absorbed into the reference group.

Rosenfeld (2015) proposes a fundamental change to how family types are modeled. He dispenses with Regnerus's categories entirely and focuses on the concept of *family transitions*. A family transition is any change in the adult members of the respondent's household while growing up. If one parent moves out of the household, it counts as one transition. If a grandparent moves into the household and then later moves back out, it counts as two transitions. And so on. Rosenfeld replaces the categorical family structure coding with a count of the number of family transitions during childhood. Thus, while the Regnerus coding strategy sets up a comparison between LGBT families and IBFs, the Rosenfeld strategy compares LGBT families with those with the same recorded number of family transitions. Table 11.1 shows the regression structure of these two approaches. Rosenfeld's coding substantially reduces the effect of an LGBT parent in this model, though the effect is still statistically significant in Model 2.

However, measurement error problems frustrate this coding structure as well. People who grew up in IBF were not asked their detailed family history. Hence, for the IBF group, we do not know if grandparents, aunts, or uncles moved into or out of the household at any point in their lives. For non-IBF respondents, these extended family members count for family transitions. If we assume that IBFs had unmeasured transitions of adult family members, then the analysis is biased. What would be the direction of bias? IBFs appear as perfectly stable households, when some of them likely were not. So, IBFs would look *even better* if we could measure and remove the (negative, as shown in Table 11.1) effect of these family transitions.

Model 3 in Table 11.1 shows the effect of splitting LGBT parents into two groups. Both lesbian mothers and gay fathers have similar effect estimates (–0.27 and –0.40) as in the pooled Model 2 (–0.31); however, estimated separately, they have larger standard errors and do not achieve statistical significance. This is a classic case when subgroups should be pooled together for statistical efficiency: The within-group differences are not of interest, and pooled together the results are more informative. Keeping the gay father/lesbian mother estimate separate does not meaningfully affect the coefficient; it only serves to reduce statistical significance. Model 3 does not indicate a null effect; rather, it shows that the analysis runs out of statistical power when the LGBT group is split.

The differences between the Regnerus model structure and the Rosenfeld structure leads to an important debate (and as the multiverse will show, one's stance on this debate is critical in shaping one's conclusions about the research question). Regnerus's main comparison is between LGBT families and IBFs, and many have pointed out the flaws in calling this comparison an "LGBT effect." It is surely an unfair standard to compare LGBT parents to perfect, stable, intact families. Indeed, LGBT are not the only family type that compares unfavorably to IBFs; every other family type compares unfavorably to IBFs, regardless of whether a same-sex couple is involved. To truly estimate an LGBT effect, one would need to disentangle it from the "not-IBF"

TABLE 11.1 *Regression models for effect of family structure on positive outcomes index, Regnerus versus Rosenfeld definitions*

				Model 4
	Model 1	**Model 2**		**Regnerus +**
	Regnerus	**Rosenfeld**	**Model 3**	**Rosenfeld**
Family type				
IBF	(ref.)			(ref.)
Gay/lesbian parents	−0.87***	−0.31*		−0.50**
	(0.16)	(0.16)		(0.17)
Lesbian mother			−0.27	
			(0.15)	
Gay father			−0.40	
			(0.35)	
Divorced later	−0.43**			−0.43**
	(0.15)			(0.08)
Adoptive	0.23			0.23
	(0.17)			(0.17)
Stepfamily	−0.50***			−0.23*
	(0.10)			(0.10)
Single parent	−0.28***			−0.11
	(0.07)			(0.07)
Other family type	−0.40***			−0.31**
	(0.09)			(0.10)
Other controls included?	Y	Y	Y	Y
Family transitions (broad)		−0.07***	−0.07***	−0.06***
		(0.01)	(0.01)	(0.01)
Constant	−0.47**	−0.56**	−0.56***	−0.43*
R^2	0.26	0.26	0.26	0.28

Notes: $N = 2{,}466$. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Weighted regressions predicting positive outcomes index. Other controls are age, biological mother's highest level of education, origin-family income, and dummy variables for female, white, and having been bullied as a child.

effect. Rosenfeld's coding strategy achieves this disentangling, by comparing children of LGBT families to children of non-LGBT families with the same level of instability.

On the other hand, Regnerus pushed back against this critique by arguing that LGBT parents tend to have less stable relationships, and this is one of the reasons why LGBT parenting is bad for children. In other words, he argues that family transitions are on the causal pathway from LGBT parent to child outcomes and thus serve as a bad control

(Regnerus 2012b: 1370). There is no clear winner in this debate; both Rosenfeld and Regnerus make reasonable points.

Rosenfeld argues against the endogeneity interpretation – which is central to his preferred analysis. He notes that while LGBT families indeed have more family transitions than other family types (say, stepfamilies) in the NFSS data, this is frequently because of lesbian parents losing legal custody of their children – something that rarely happens to straight mothers who are divorced or otherwise single (Rosenfeld 2015: 495). Rosenfeld thus argues that the higher family transitions of LGBT parents is caused by a legal system biased against them, rather than being endogenous to LGBT status itself, and thus constitutes an appropriate control. This is a compelling argument as far as it goes, though it is not clear that the legal system fully accounts for potential endogeneity in family transitions.

Another detail in Rosenfeld's coding is the types of family transitions that he includes. He presents two different measures of transitions. His preferred measure includes any adult transition, so some transitions are extended family members (such as grandparents) or unrelated adults entering or exiting the child's household. His second measure includes only transitions involving the loss (or return) of a parent, since parental transitions are likely more important in the lives of young people. Rosenfeld favors the broad "any adult" coding on the basis of an R^2-type test, but in the absence of a theoretical motivation, the R^2 is not necessarily a strong reason to favor a model specification. We include both measures in the multiverse to see how the model choice influences the results.

Rosenfeld presents his coding structure as a complete alternative to Regnerus, but these approaches could be combined: retaining the full family types and simply adding family transitions as an additional element. Rosenfeld (2015, fn. 7) suggests this would not be possible due to a collinearity problem, but that is incorrect. Within family types (other than IBF) there is plenty of variation in the number of transitions, and models with both family types and family transitions readily estimate. This is shown in the final column of Table 11.1, which is a hybrid of the Regnerus and Rosenfeld coding. There is shared variance between family types and family transitions, and retaining family types, at least in this simple model, reduces the effect of transitions and increases the effect of LGBT parent. This suggests that family types are needed for more accurate estimation of the effect of transitions.

Together, there are three broadly different ways of constructing and coding family type – the central explanatory variable in this study. In

Table 11.1, we label these the Regnerus approach (full categorical coding), the Rosenfeld approach (dummy coding with transitions), and the Regnerus + Rosenfeld approach (full categorical with transitions). None of the approaches are perfect given the underlying data problems, and each have different strengths and weaknesses.

There are other details in the data processing multiverse. But in the interest of getting to the results, we will report the remaining data processing alternatives in Table 11.3 of the influence analysis that shows their impact.

FULL MULTIVERSE RESULTS

Figure 11.2 shows three distinct details: (1) the Regnerus estimate, indicated by a vertical line, (2) the controls-only multiverse, previously reported, and (3) the full controls-plus-data processing multiverse. The full multiverse contains all 2,048 models from Multiverse 1, and it also incorporates all combinations of those controls with all the other data processing and modeling choices: (1) alternative coding strategies for four control variables and the outcomes index, (2)

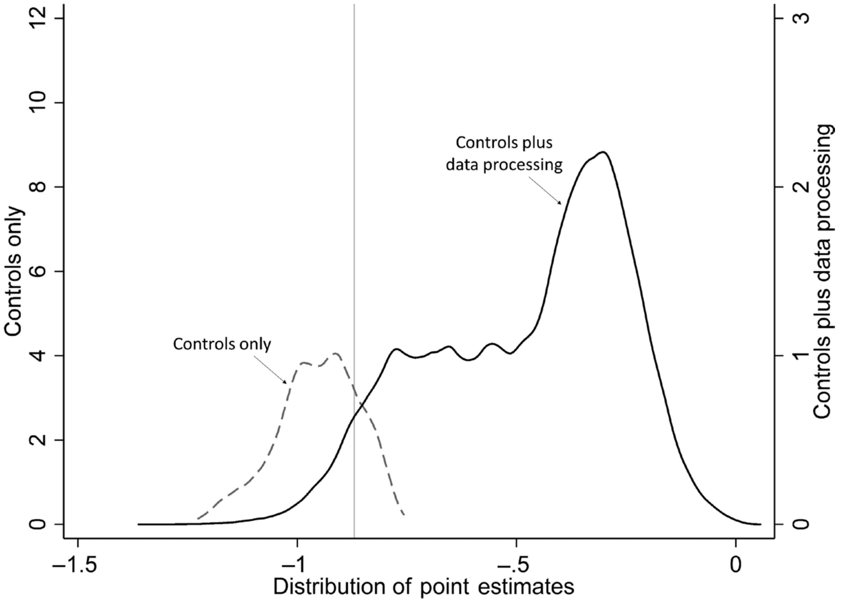

FIGURE 11.2 Distribution of gay/lesbian parenting effects in full multiverse
Note: Data from the NFSS, $N \approx 2{,}466$, depending on the specification. Results from 2.65 million models. The vertical line shows the Regnerus estimate.

alternative ways of dealing with outliers, (3) corrections for misclassification of LGBT parent, (4) alternative definitions of the comparison group, and (5) alternative weighting strategies. This multiverse expands to roughly 2.7 million models.

The full multiverse distribution is much wider and is substantially shifted toward zero compared to the control variable multiverse. The modeling standard error in the full multiverse is more than twice as large as that for the controls only. A wider range of estimates is possible in the full multiverse, and 95 percent of estimates are smaller than the Regnerus estimate (which falls at the fifth percentile of the modeling distribution). Data processing decisions have much more influence on these results than the choice of controls. This is both because there are many more alternative specifications for data processing and because some of those processing decisions are much more influential than the control variables. Our central conclusion is to affirm that data processing decisions deserve a central place in multiverse analysis. While the choice of control variables is much more tractable in a combinatorial multiverse analysis, controls may not be the most important decisions in the overall analysis, nor do they adequately account for the potential variability of findings across different analysts using the same data.

In terms of substantive conclusions, the full multiverse analysis shows that critics were only partly successful in challenging the results. Clearly, very few of the 2.65 million estimates are zero or opposite-signed. Table 11.2 shows the summary statistics. Only 992 models report a positive coefficient for LGBT parent, which is 0.04 percent of the models, and none of those positive estimates are statistically significant. In practical terms, the estimated effects are strictly negative. Some 76 percent of the models are both negative and statistically significant. In the full multiverse, the Regnerus estimate appears more extreme than most alternative models, but both the sign and significance are very stable. The robustness ratio in the full multiverse is 3.14 – easily robust by conventional rules of thumb.

The Regnerus estimate is a borderline tail estimate that depends on multiple contested assumptions. But a conclusion of "no effect" is not supported by these data, either. The most reasonable conclusion is that there is some negative effect of gay/lesbian parenting on children's outcomes in these data, but it is probably smaller than suggested by Regnerus's original study.

To be clear, we do not view the mean of any multiverse as the best estimate. The purpose of multiverse analysis is to show what estimates are

TABLE 11.2 *Multiverse results, gay/lesbian parenting effect*

	Controls only	**Controls plus data processing**
Number of models	2,048	2,654,208
Mean estimate	−0.95	−0.48
Minimum coefficient	−1.23	−1.36
Maximum coefficient	−0.76	0.06
Positive	0	992
Positive and significant	0	0
Negative	2,048	2,653,216
Negative and significant	2,048	2,022,745
Percent negative and significant	100%	76%
Modeling standard error	0.09	0.22
Sampling standard error	0.16	0.16
Total standard error	0.19	0.28
Robustness ratio	4.66	3.14

plausible given the data and a rich set of modeling assumptions. But other things being equal, estimates closer to the center of the distribution are technically easier to justify as they are supported under more bundles of assumptions, while tail estimates are harder to justify because they require a larger set of exact assumptions to be true.

In Table 11.3, we document every data preprocessing decision included in the multiverse and show how each decision affects the empirical conclusion. Column 1 shows the average coefficient for all models invoking that assumption. The "all models" average is −0.48, meaning an effect size of about one half of a standard deviation in the outcome index. Column 2 shows the average standard error for all models invoking that assumption. Column 3 shows the percent change in the average coefficient, relative to the "all models" average. This column indicates the relative influence of the model assumption specified in each row. So, the "Regnerus model" has an estimate of −0.87, which is 82 percent larger than the overall mean. In this last column, a positive percent change means the specification choice leads to a larger absolute estimate, while a negative percent change means it pushes the estimate toward zero. As the table documents, our multiverse considers many different dimensions of an empirical analysis. As the table shows, many factors contribute in small ways to the diversity of possible results. The central specification issues are how to structure the comparison group for LGBT families (options C.1 to C.6) and how to deal with potentially misclassified or borderline cases (options M.1 to M.4).

TABLE 11.3 *Influence effects for gay/lesbian parenting effect*

		(1) Average coefficient	(2) Average standard error	(3) Percent change from "all model" average
All models		**–0.48**	**0.15**	
Regnerus model		**–0.87**	**0.16**	**82%**
Comparison group	C.1: IBFs	–0.76	0.15	59%
	C.2: Both parents until 18, may have divorced later	–0.71	0.14	49%
	C.3: All families with same number of "any adult" transitions	–0.26	0.14	–44%
	C.4: All families with same number of parental transitions	–0.29	0.15	–40%
	C.5: IBF, with "any adult" transition and other family types	–0.38	0.15	–21%
	C.6: IBF, with parental transitions and other family types	–0.47	0.15	–2%
Misclassification of people raised by gay father/lesbian mother	M.1: Raw data	–0.55	0.14	16%
	M.2: Exclude misclassified cases	–0.47	0.13	–1%
	M.3: Exclude misclassified/ borderline cases	–0.38	0.14	–20%
	M.4: Exclude cases where respondent never lived with gay father's/lesbian mother's same-sex romantic partner	–0.50	0.18	5%
Positive outcomes index	P.1: 19 component variables; exclude cases with missing values on any component variable	–0.51	0.16	7%

	P.2: 29 component variables; exclude cases with missing values on more than 19 component variables	–0.44	0.14	–7%
Weights	W.1: Sampling weights	–0.46	0.19	–3%
	W.2: No sampling weights	–0.49	0.11	3%
Outliers	U.1 No adjustment for outliers	–0.47	0.15	–1%
	U.2 Winsorize outcome index	–0.48	0.15	1%
Coding of income	I.1: Categorical, missing kept as category	–0.49	0.14	3%
	I.2: Categorical, missing treated as missing	–0.45	0.15	–6%
	I.3: Continuous, log income	–0.44	0.15	–7%
	I.4: Drop income	–0.53	0.15	10%
Coding of age	A.1: Age	–0.47	0.15	–1%
	A.2: Age and age squared	–0.47	0.15	–1%
	A.3: Drop age	–0.48	0.15	1%
Coding of race	R.1: Dummy for white vs. nonwhite	–0.48	0.15	0%
	R.2: Five categories	–0.48	0.15	0%
	R.3: Drop race	–0.48	0.15	0%
Coding of mother's education	E.1: Missing kept as category	–0.49	0.14	2%
	E.2: Missing treated as missing	–0.47	0.15	–1%
	E.3: Drop mother's education	–0.47	0.15	–1%

Notes: All models run using OLS. In addition to those shown, the multiverse analysis also considers the following possible controls: female, bullied as youth, origin-family welfare receipt, mother's age at birth, father's age at birth, region, and metropolitan status. Replication package at https://osf.io/45ft2/.

An interesting result involves the use of sampling weights (options W.1 and W.2). All the researchers contributing to this analysis used sample weights to render the NFSS data representative of the general population in terms of observable demographics like age, gender, and race (all analysts agreed that the raw data were highly unrepresentative of the US population). However, a priori it is not clear that weights are needed for regression and none of these authors based the decision to use weights on a Hausman test (Bollen et al. 2016). Hence, our data processing multiverse also includes unweighted models. At first glance, the use of weights doesn't seem to matter much: the unweighted results (–0.49) are slightly larger in magnitude than the weighted estimates (–0.46). However, the use of sample weights in this analysis leads to standard errors that are nearly twice as large – reflecting the concern that weighted least squares can be very inefficient (Solon et al. 2015). The unweighted models have an average standard error of 0.11; the weighted models have an average of 0.19. This loss of efficiency has a big impact on the significance level of many models: Among unweighted models, 95 percent have a significant LGBT effect. Among the weighted models, with their slightly smaller effect sizes and much larger standard errors, only 58 percent have a significant effect. Even though the weights change the modeling distribution very little, they have a tremendous impact on the vote count.

A smaller but noticeable processing choice is how to construct the outcomes index: Rosenfeld's approach (P.1) included only nineteen of the outcomes and dropped observations that were missing on any one of those outcomes (i.e., listwise delete before constructing the index). Our alternative (P.2) included twenty-nine outcomes and only dropped respondents that were missing on a large portion of the outcomes. We developed this alternative because we believed the Rosenfeld index resulted in an unnecessary loss of information. Our method results in smaller estimates (–0.44) than the Rosenfeld index (–0.51) – moderate influence in the context of this study.

Many processing choices had only small effects, if any, on their own. Winsorizing the outcomes index had little effect on the results. The coding of race, or even to include race in the analysis, had no influence in these data (options R.1–R.3). Critics had objected to Regnerus's specification of race as a white/nonwhite dummy, but estimates are generally the same if race is coded as a dummy, coded as five categories, or not included at all. The literature also has some disagreement over the coding of income, either (I.1) as income categories, with one of the categories being "income not reported"; (I.2) as income categories, with missing income coded as missing rather than a separate category; (I.3) convert the categories to continuous log income; or (I.4) drop income entirely (as per a control variable multiverse). Among these options, dropping income leads to the largest LGBT effect (–0.53), while log income gives the smallest effect (–0.44). The important point here would be that there is little difference in results between continuous log income and categorical income.

One last data processing choice not included in Table 11.3 is the general treatment of missing data using listwise delete (adopted here, with some noted exceptions) or using multiple imputation. Our main conclusion is that multiple imputation is not computationally feasible within a large multiverse analysis. Because multiple imputation is itself computationally intensive, implementing it as an alternative data processing choice of each one of over 2.5 million models had an unacceptable runtime using a conventional computer – potentially several months. In smaller scale testing, however, we found that multiple imputation had little influence relative to listwise delete in these data.

TWO MULTIVERSES: REGNERUS VERSUS ROSENFELD

A central debate comes down to whether "family" should be operationalized as the types laid out in Regnerus or as the history of transitions advocated by Rosenfeld. In practice, there are many permutations of the two approaches, including combinations of both that are hard to rule out. Nonetheless, Figure 11.3 shows what is at stake in the two core approaches.

The key distinction between these two sets of models is that all models in the "Rosenfeld multiverse" include a control for family transitions (all of options C.3 through C.6), and the models in the Regnerus multiverse do not (options C.1 and C.2). The two approaches yield strikingly different results: The Regnerus models have an average coefficient of –0.74, and 99.9 percent of models have a negative, significant LGBT parent effect. The Rosenfeld models have an average of –0.35, and only 64 percent of models have a significant effect. That's a difference of about four-tenths of a standard deviation on the outcomes index; it is not trivial. If one favors the Regnerus coding, then the LGBT parent effect is negative, large in magnitude, and fully robust. If one supports the Rosenfeld assumptions, the effect is still negative but much smaller and only modestly robust.

The division between the "Regnerus models" and the "Rosenfeld models" reflects fundamentally different conceptions of what constitutes

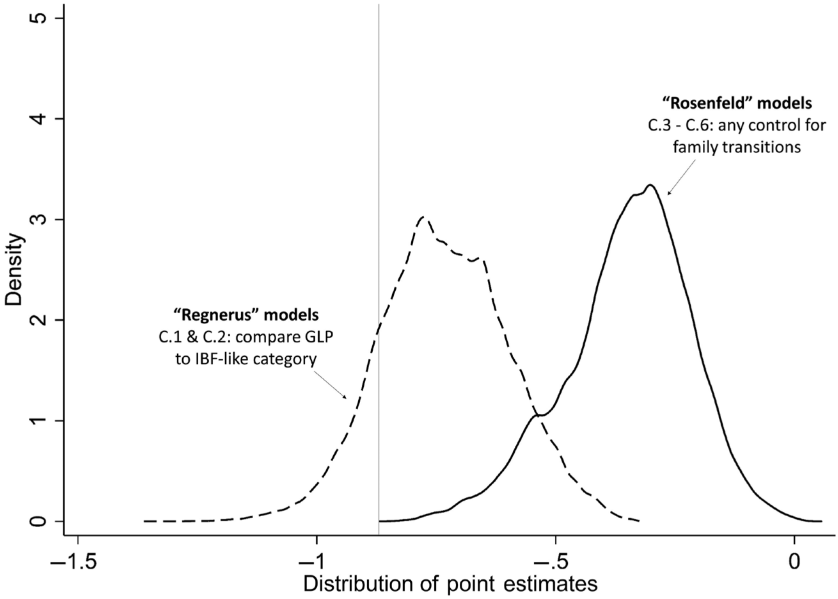

FIGURE 11.3 Distribution of gay/lesbian parenting effects in Regnerus multiverse versus Rosenfeld multiverse
Note: Data from the NFSS, $N \approx 2{,}466$, depending on the specification. Results from 2.65 million models. The vertical line shows the Regnerus estimate.

an LGBT parent effect, and both conceptions have weaknesses. The Regnerus models compare people with LGBT parents only to people from the most stable families, which probably conflates a true LGBT effect with a more general "not-IBF" effect. On the other hand, the Rosenfeld models may be biased downward if family transitions are one of the pathways through which LGBT parenting affects adult outcomes (as Regnerus argues is the case). If one prefers Rosenfeld's conception of an LGBT parent effect, one probably thinks the true effect is closer to the Rosenfeld curve; if they prefer Regnerus's conception, they look at the Regnerus curve.

For those looking to discredit Regnerus's findings, we point to one more critical result: In order to reject those findings, it becomes necessary not only to prefer the Rosenfeld model structure but also to defend the use of sample weights in this analysis. Among the Regnerus models, over 99 percent have a significant effect, regardless of whether weights are used. But among the Rosenfeld models, the weights make a critical difference: 92 percent of the Rosenfeld/unweighted models have a significant result, compared to only 37 percent of the Rosenfeld/weighted models. If this difference occurred because the weights led to smaller point estimates, then the weighted models would surely be more defensible (Solon et al. 2015; Bollen et al. 2016). But that is not the case: Using weights only moves the average coefficient (among Rosenfeld models) from –0.36 to –0.34. The difference in significance levels is driven almost entirely by the widening of the standard errors in the weighted models (from 0.11 to 0.19). The weighted models have about the same effect estimate but are much less precise.

CONCLUSIONS

In our experience of replicating this original study and the two replications of it, we were continually surprised by how much more we learned about the research through replication than through reading alone. We thought we understood each article quite well by reading its methodological descriptions and published results, but in practice, we almost never did and were often surprised by things we learned through replication. This project took much more time and effort than we expected, but we highly recommend the experience to others. Only through replication can one fully embrace the spirit of scholarly skepticism, in full knowledge of how the estimates were produced. To replicate is to appreciate the assumption-laden nature of empirical claims.

We opened the discussion of data processing with the Durante et al. (2013) multiverse, which starkly rejected the claims of that original study: The authors' preferred estimate was the single most extreme finding available in the data; only 4 percent of multiverse estimates were statistically significant, and half of all estimates were opposite-signed – amounting to an overwhelming null result. In contrast, we were surprised by the robustness of the Regnerus finding. Prior to examining the data directly, we accepted the conclusions written by the critics and expected that a comprehensive multiverse analysis would drive their point home in a powerfully conclusive way. Rosenfeld (2015: 478) had written that in reanalysis "same-sex couple parents … are weakly or not at all associated with negative adult outcomes." Cheng and Powell (2015: 615) concluded that the Regnerus results are "so fragile that they appear largely a function of … possible misclassifications and other methodological choices." It is certainly true that in multiverse analysis – recognizing the data processing questions raised by the critics – the estimates become much smaller. Our surprise was discovering that in these data a negative effect is nonetheless still robust and that there are essentially no opposite-signed results.

This multiverse of estimates all derive from one underlying dataset. Regardless of the analysis, there are undeniable problems in the data themselves that have no postcollection analytical solution. Whether one prefers to specify the comparison group for same-sex parents using family types or family transitions (or both, as seems most sensible), both these variables are substantially mismeasured. Transitions are mismeasured since IBFs were never asked about any extended family transitions, and types are mismeasured because a family is falsely restricted to being only one type when in fact the categories are overlapping (Perrin et al. 2013). We do not know how these errors in measurement affect the results because the needed information was not collected. Nevertheless, both critics offer at least partial praise for the data: saying that the collection effort was "certainly impressive" (Cheng and Powell 2015: 617) and had "advantages over other data sources" (Rosenfeld 2015: 479). We defer to this expertise on evaluating the relative quality of the data at the time but remain disappointed that the key measures central to this research have obvious and unresolvable errors in measurement.

We have other reservations about the data. We see it as bad practice to prescreen a survey with the question "are either of your parents gay?" and then include everyone who responds "yes." It has been easy to document that some of the respondents "were having fun with the survey," but the underlying selection mechanism into this kind of politically charged (at the time) prescreened survey is very difficult to understand or adjust for. The problem of nonserious survey respondents is greatly compounded when studying small subsets of the population. For small populations, even small rates of misclassification error can create large amounts of noise and bias in results (Cheng and Powell 2015). In the screening survey of the NFSS data, only 1.7 percent of the respondents indicated that a parent had been in an LGBT relationship. If pranksters are especially drawn to survey questions about sexuality – which seems to be the case – then many more of the respondents might be misclassified.

Social theory seeks to predict and explain *phenomena*, not *data* per se. Theory need not apply to data when they are collected through poor proxy measurements of the phenomena of interest. Scholars hope that data are closely coupled with the underlying phenomena that matter, but measurement systems have many flaws. For this reason, we strongly favor big administrative data for the analysis of small, hard to reach populations. Using such data in the Netherlands, Mazrekaj et al. (2020) find that the children of same-sex parents do better than average at school, implying a supportive and successful home environment. Acknowledging that no dataset is flawless, we hope the literature going forward will aspire for bigger and better quality data.

Finally, we emphasize that critics and replicators can and should be subject to criticism themselves. Publication and replication are a back-and-forth debate in which, it is hoped, the best arguments ultimately win. Replicators should be held to a high standard, as they have the benefit of the original authors' thinking, and should be expected to only improve the quality of evidence. In this view, there are some flaws in the critics' work that deserve mention.

A key mistake that both critics make is to focus exclusively on significance testing, particularly as they drop substantial portions of the data. Neither Rosenfeld nor Cheng and Powell ever report a substantive estimate or regression coefficient for their LGBT effects but instead report only the significance tests. Regnerus's regression tables are themselves unconventional, but readers can at least look at a table and see what is the difference in outcome between any family type on any outcome variable. Readers cannot retrieve this information from either of the critics, who just report whether a regression yielded a "star." A focus on significance testing alone, to the exclusion of effect sizes, is bad statistical practice in any context (Gelman and Stern 2006). It is especially flawed here, as the critics provide new specifications that cut the sample size and greatly reduce the treatment group. Whenever one imposes a sample exclusion – especially from an already small treatment group – this must be evaluated by the change in the parameter estimates, not by a drop in statistical significance. It is not a fair assessment of the data to report that significance levels fall after dropping as much as 44 percent of the treatment group: Of course statistical significance will be lower when the sample is smaller. Our multiverse modeling distribution (Figure 11.2) and influence analysis (Table 11.3) correctly put the focus on effect sizes across different specifications. This is not a criticism of the decision to drop misclassified cases, but it is a reminder of the importance of looking at the change in effect sizes rather than significance levels whenever imposing a sample restriction.

A second problem in this literature is a weakly theorized subdivision of the treatment group: splitting LGBT parents into separate specifications for gay fathers and lesbian mothers. Regnerus originally coded LGBT parents separately in this way, and critics have followed suit in their own statistical models. The separate coding is not wrong but is unnecessary, was not given substantive justification, and dilutes statistical power. Since the effect sizes for gay father and lesbian mother are similar, there is no reason to split up LGBT groups by sex. Indeed, none of the authors ever interpreted or discussed as meaningful the difference between gay fathers and lesbian mothers, leaving it unclear why the separate coding was used. Our models pool LGBT parents rather than splitting by sex, which preserves statistical power and increases the statistical significance of the results.

Critics of the Regnerus study have challenged the data and analysis in many ways that, when implemented, reduce the effect size. But the remaining debate is over the magnitude of the LGBT parent effect or over the quality of the data but not over the existence of an LGBT parent effect *in this dataset*. To the extent that modeling decisions matter, data processing features are more important than the control variables in this analysis.[5]

The fundamental principle of multiverse analysis is open and transparent social science. In the debate around the Regnerus study, we provide three distinct contributions: (1) the distribution of results across over 2.6 million model specifications, with model ingredients from both the original study and from its critics; (2) an influence analysis documenting how different data processing decisions affect the results; and (3) a new replication package (https://osf.io/45ft2/) that combines all this work including all analyses to date so that others can question, learn from, and expand on this multiverse.

[5] Admittedly, debates about how to structure the variable of interest could partly be framed as a debate about control variables, just as questions about, say, selection bias could be framed as an omitted variable problem.