No Basis: What the Studies Don't *Tell Us About Same-Sex Parenting*

By
Robert Lerner, Ph.D., and Althea K. Nagai, Ph.D.

Marriage Law Project, Washington, D.C.
January 2001

Table of Contents

Executive Summary

It is routinely asserted in courts, journals and the media that it makes "no difference" whether a child has a mother and a father, two fathers, or two mothers. Reference is often made to social-scientific studies that are claimed to have "demonstrated" this.

An objective analysis, however, demonstrates that there is no basis for this assertion. The studies on which such claims are based are all gravely deficient.

Robert Lerner, Ph.D., and Althea Nagai, Ph.D., professionals in the field of quantitative analysis, evaluated 49 empirical studies on same-sex (or homosexual) parenting.

The evaluation looks at how each study carries out six key research tasks: (1) formulating a hypothesis and research design; (2) controlling for unrelated effects; (3) measuring concepts (bias, reliability and validity); (4) sampling; (5) statistical testing; and (6) addressing the problem of false negatives (statistical power).

Each chapter of the evaluation describes and evaluates how the studies utilized one of these research steps. Along the way, Lerner and Nagai offer pointers for how future studies can be more competently done.

Some major problems uncovered in the studies include the following:

Unclear hypotheses and research designs
Missing or inadequate comparison groups
Self-constructed, unreliable and invalid measurements
Non-random samples, including participants who recruit other participants
Samples too small to yield meaningful results
Missing or inadequate statistical analysis

Lerner and Nagai found at least one fatal research flaw in all forty-nine studies. As a result, they conclude that no generalizations can reliably be made based on any of these studies. For these reasons the studies are no basis for good science or good public policy.

Four Appendices follow. Appendix 1 is a bibliography of the studies and related publications. Appendix 2 is a table that summarizes the evaluation of each of the studies with regard to each research step. Appendix 3 (by William C. Duncan) is an overview of how these studies have been used in the law. Appendix 4 (by Kristina Mirus) describes how the media has covered these studies.

Foreword

By David Orgon Coolidge
Director, Marriage Law Project

What do existing studies tell us about the impact of same-sex parenting on children?

Nothing.

That's right, nothing.

You would never know that, however, if you were to read most court decisions, law review articles, commission reports or newspaper articles. You would hear the opposite.

The point of the study which follows is *not* to try to answer the question, "Why is this?" Instead, Robert Lerner and Althea Nagai have simply evaluated the studies themselves. They have asked: What are their hypotheses? How do they set about to prove them? What do they conclude? In formulating, executing and analyzing their research, do these studies get it right?

The results are not pretty. Lerner and Nagai identified 49 empirical studies on the subject of same-sex parenting.* After going through them all, inch-by-inch, they found…nothing.

I first saw the need for such an evaluation back in 1996, in Honolulu, Hawaii. I sat through two weeks of testimony in the same-sex "marriage" case, *Baehr v. Miike*. Almost all of the testimony was

* The terms "homosexual" (on the one hand) and "gay and lesbian" (on the other) are both loaded. The studies evaluated here examine parenting by same-sex couples in sexual relationships. To avoid distraction I have used the term "same-sex."

by social scientists. It raised questions I could not shake.

Many of those questions are larger ones, such as how science and morality relate. But other questions were more straightforward: Are these studies well-done by normal standards? Should journals publish them? Should policymakers rely on them?

The fact of the matter is that many people, including policymakers, *are* relying upon these studies in litigation, legislation, scholarly writing, and in the larger public debate. (To confirm this, see Appendices Three and Four by Bill Duncan and Kristina Mirus.)

The least that should be done is to take a serious look at the methodology of the studies. That is what Robert Lerner and Althea Nagai have done. At the risk of damaging their professional and academic reputations, they have done this full-scale evaluation. Here you have the results. You will learn more than you ever wanted to know about how studies should be designed, implemented and evaluated. And you will learn how even the best studies of same-sex parenting fall far short of these standards.

Lerner and Nagai have not only taken apart existing studies, however. By setting their evaluation in the context of a broader discussion of social-scientific research, they have pointed the way toward better studies. They are clearing ground so others can go forward.

In the meantime, the rest of us have decisions to make. How shall we proceed? Lerner and Nagai make no attempt to answer this question. They have only one point to make: Whatever you do, don't do it based on these studies.

Take the time to see what Lerner and Nagai discovered about the same-sex parenting studies. These authors know a better or worse study when they see it, and they tell it like it is. Whether we like it or not, we are all in their debt.

No Basis: What the Studies Don*'t Tell Us About Same-Sex Parenting*

By Robert Lerner, Ph.D., and Althea K. Nagai, Ph.D.

Introduction

"[C]hildren with two parents of the same gender are as well adjusted as children with one of each kind."[1]

This view, revolutionary in its implications, and unheard of five years ago, is now commonly asserted by social scientists, lawyers, policymakers and the media. Numerous studies are routinely offered to show that the sexual orientation of a couple makes "no difference" to the well-being of children. The obvious implication of this view is that two gay "dads" or two lesbian "moms" can raise a child as well as can two married biological parents. Simply being surrounded by two caring adults is thought to be enough to raise most children to be healthy, well-adjusted adults.[2] Is this claim true? Does the research supporting it stand up to scientific scrutiny? These are the questions discussed in this study. Our approach to this question concentrates on an analysis of the methodologies used to carry out existing same-sex parenting studies. We conclude that the methods used in these studies are so flawed that these studies prove nothing. Therefore, they should not be used in legal cases to make any argument about "homosexual vs. heterosexual" parenting.[3] Their claims have no basis.[4]

What Social Science Requires

Social science research is a complex process, but it follows a series of well-defined steps. Each of these steps must be carried out properly to obtain valid conclusions. Like a chain is only as strong as its

Notes for this section begin on Page 9

weakest link, the conclusions derived from any research study are only as reliable as its weakest part.[5]

The typical sequence of social-scientific research involves:

Formulating concepts and research hypotheses
Creating the research design
Establishing measurements for important concepts
Defining the sample and its selection procedures
Collecting the data
Performing statistical tests on the data analysis, and
Based on the above, hopefully reaching valid conclusions.

The studies discussed here will be analyzed by following the typical sequence of social science research methods textbooks. Under each heading we will analyze all the studies to see how well they meet accepted social science standards. Any failures in the process-failure to properly design the study, failure to properly measure the relevant variables, failure to properly control for extraneous variables, and failure to use the proper statistical tests-make a study scientifically invalid. Most importantly of all, if a study claims to find no difference i.e. "non-significant results," and that study failed to carry out one or more of these research links in the proper manner, its conclusions are purely and simply invalid. Why? Because failing to carry out correctly one or more of these essential elements, in and of itself increases the chances of finding non-significant results. In other words, if you look for wrong findings using wrong methods, it is even more likely you'll get wrong results.

Social Science and Public Policy

With one exception, the authors of these studies wish to influence public policy to support same-sex marriage and the adoption of children by homosexual couples. While the authors of these studies have every right to advocate this point of view, as do those who disagree with them, their wish means that the stakes in obtaining valid answers to these research questions are very high. It is not enough for a study to be interesting, or raise important questions about a subject, or to be provocative. While these criteria may be enough to

get a study published, they are not strong enough to justify dramatic alterations in long-established public policies. To justify changes in public policy, studies should be strong enough that policy makers have faith in the study's reliability, and confidence that more research is unlikely to overturn its findings.

This is not an unreasonable requirement. The public policy consequences of relying on inadequate or insufficient studies can be devastating.

In 1973, a literature review undertaken by social scientists Elizabeth Hertzog and Cecelia Sudia purported to find that the effects of growing up in fatherless homes are at most minimal and likely to be due to other factors. The authors did not stop here. They stated it might be a good idea to increase community support for single parents,[6] rather than developing policies that forestall the absence of fathers, or that oppose easy divorce. This study was part of a larger current of expert opinion proclaiming that growing up in a one-parent family had no negative consequences for the children living in these arrangements.

With more rigorous research, these interpretations were challenged and eventually overthrown. Research has demonstrated that divorce is not the costless exercise for children that many had proclaimed it to be. The newer research demonstrated that children growing up in fatherless families do not do as well financially, in school, and emotionally both as children and as adults, as those in families with their married biological parents.[7] Therefore, the standards used here, to investigate studies on the impact of same-sex parenting on children, are necessarily demanding. We owe ourselves nothing less.

How These Studies Were Selected

All of the articles used in this review deal with same-sex couples and/or their children. We excluded dissertations, review articles, and articles in the nonscientific press.[8] We have only analyzed reports of original research studies (i.e., real social science). We have tried to be as exhaustive as possible, although research is exploding in this field.

Working from a variety of angles,[9] we arrived at a final list of 49 studies for analysis that have been either published in professional journals or as chapters in a book.[10] All present the results of original research on homosexual parents and/or their children.

Do these 49 studies offer conclusive proof that there is "no difference" between heterosexual and homosexual households? We believe that these studies offer no basis for that conclusion—because they are so deeply flawed pieces of research. The reader is invited to make his or her own judgment.

Notes to Introduction

1. Harris, 1998, p. 51. Harris cites this body of studies in her controversial book on child development.

2. For example, sociologist Judith Stacey, writing in a recent issue of *Contemporary Sociology,* a book review journal, that focuses on sociology and public policy, writes that "thus far the research on the effects of lesbian parenting on child development is remarkably positive and therefore challenging [the status quo] . . ." Stacey, 1999, p. 21.

3. This is not the same as concluding that traditional family arrangements are better. It simply states that the evidence presented above does not justify the opposite conclusion.

4. Since vocabulary related to homosexuality is extremely contentious, we should explain our terminology. We have tried to generally use the term "same-sex," since the terms "gay and lesbian" (on the one hand) and "homosexual and heterosexual" (on the other hand) are so ideologically polarized. However, the studies themselves use one set of terms or the other, so the reader should expect a variety of terms.

5. For example, one can have a perfectly selected sample, but concepts that are so badly defined and poorly measured that one is unable to conclude anything from the results of the study.

6. Cited and discussed in Popenoe, 1998, pp. 59-61; McLanahan and Sandefur, 1994, pp. 13-14. A well-known study in the same vein was sociologist Jessie Bernard's The Future of Marriage (1972), which became famous or infamous for its comment that to be happy in a traditional marriage a woman must be mentally ill (quoted in Whitehead, 1998, p. 51).

7. For detailed discussion of the extensive research literature see the following

works: Waite, 1995; McLanahan and Sandefur, 1994; Popenoe, 1998; Amato and Booth, 1997; and Whitehead, 1996.

8. Dissertations are original studies, not review articles; but if they go unpublished, the most one can say is that they met the minimum standards for receiving a degree from the university that granted them, and nothing more. Review articles were excluded because they present no original data for assessment. Articles found in the nonscientific press were excluded because their criteria for publication (e.g., popular interest, immediate policy relevance) are not the same as those for assessing the scientific credibility of a study.

9. Graciela Ortiz, M.S.W., conducted initial bibliographic research in the summer of 1998. Additional studies were identified by examining law review articles published by Wardle (1997) and Ball and Pea (1998), briefs filed in *Baker v. State*, the Vermont same-sex "marriage" lawsuit, and *Lesbian and Gay Parenting: A Resource for Psychologists*, Washington, D.C.: American Psychological Association, 1995.

10. There is also one book, Tasker and Golombok (1997), which is part of the study.

Chapter 1
The Good, the Bad, or the Ugly: Formulating the Hypothesis

We've all heard the slogans: "If you don't know where you're going, you can count on getting there," or "If you aim for nothing, you're sure to hit it." The same is true for formulating the hypothesis of a research study: If your goal is to prove no differences, you're bound to reach it. But you won't have proved "no difference," only no basis.

All good studies begin with careful definitions of key concepts and careful delineation of the relationship between these concepts. Formulating the hypothesis is the crux of any scientific design, and its development requires special care. The hypothesis determines the main focus of the study, and frames all subsequent research endeavors.[1] Hypotheses can be Good (affirmative), Bad (fuzzy), or Ugly (null). Of the 49 studies, two are Good, 29 are Bad, and 18 are Ugly. Understanding why requires Social Science Research Methods 101, which we will sprinkle throughout this and other chapters.

What is a Good Hypothesis?

All good social science studies have at their core a positive hypothesis statement. This takes the form of an explicit conceptual relationship between two variables whereby something (an independent variable) "causes" something else (a dependent variable).[2] The researcher posits a direct relationship between the independent and the dependent variables.[3] The hypothesis can and

Notes for this section begin on Page 22

should be stated as a proposition that takes the following form: "the greater the a, the greater the b," where "a" is the independent variable and "b" is the dependent variable.[4]

Hypothesis statements may be either quantitative or qualitative. Consider the following example. A study group of children is enrolled in a social program such as *Head Start*, while a control group of children is not enrolled there. The independent variable here therefore, is "enrolled versus not enrolled." The dependent variable might be something like "readiness for school." Assuming that "readiness for school" is a quantitative variable (i.e., it can be scored on a three or more point scale), then the research hypothesis would compare mean levels of school readiness of those in *Head Start* with those who are not. Assuming "readiness for school" is a qualitative (i.e., "yes" or "no") variable, the research hypothesis would compare the proportion of *Head Start* children who are ready for school with the proportion who are not ready.[5] There are many different possible hypotheses a reseacher might have, depending upon the nature of the problem studied and the level of measurement assumed in the independent and dependent variables.[6] Applying this view to studying a parent's sexual identity and its possible relationship to child outcomes, the investigator should define conceptually the independent variable ("homosexual versus heterosexual" identity),[7] the dependent variable (such as a child's sexual identity, child's psychological adjustment, or the child's sexual behavior), and the posited relationship between the independent variable and the dependent variable(s).

An example of such a properly stated research hypothesis is: "the children of homosexual parents are more likely to grow up to be homosexual than are the children of heterosexual parents."

Good: The Affirmative Research Hypothesis

Only two studies among the 49 studies we examined actually contain an explicit positive hypothesis statement of this sort (**Pagelow, 1980** and **Miller, 1979**).[8] **Pagelow (1980)** hypothesizes that lesbian mothers are more oppressed than heterosexual mothers. The researcher then seeks to measure this by the concept of perceived oppression in the areas of freedom of association, employment, housing, and child custody.[9]

Miller (1979) comes closest to presenting a hypothesis in the proper format: **Miller** asks, A.) "Do gay fathers have children to cover their homosexuality?" B.) "Do they molest their children?" C.) "Do their children turn out to be disproportionately homosexual?" D.) "Do they expose their children to homophobic harassment?"[10] While **Miller** does not put his hypotheses in precisely the y=(f)x format, the hypothesis statements are both specific and decisional (i.e., they can be answered as either "yes" or "no" regarding the homosexuality of the father).

Thus **Miller**'s statements can be easily rephrased into the following testable hypotheses: A.) The reason for gay men having children is to cover their homosexuality (as opposed to other choices provided by the investigator, such as he loved the woman, he was confused, he just wanted children, don't know); B.) Gay fathers are more likely to molest children than are straight fathers; C.) Children of gay fathers are more likely to be homosexual than are children of straight fathers; and D.) Children of gay fathers are more likely to be exposed to homophobic harassment than are children of straight fathers.[11] Stated in this form, the hypotheses can then be verified or refuted by empirical research.

Bad: The Fuzzy Hypothesis

A majority of the studies we examined (29 of them or 59 percent) failed to produce a testable hypothesis. Of these, 12 studies rendered their statement of the research problem in the form of "Are there differences between homosexual and heterosexual parents?"[12] For example, **Bigner and Jacobsen, 1989a** state their research problem as, "an examination of factors that may motivate gay men to become parents, and to explore whether gay fathers may differ from heterosexual fathers regarding the value of children in their life as an adult."[13] **Brewaeys et al. (1997)** poses the problem as an examination of "family relationships and emotional/behavioral and gender role development of 4-8 year old children"[14] in lesbian donor-inseminated families, compared to heterosexual families who conceived their child also by donor insemination and heterosexual families who conceived their child naturally. Hypotheses that are stated in the form of looking for possible differences do not suffice as statements of research hypotheses. Formulation of a hypothesis in terms of pos-

sible differences fails to address any of the causal questions that guide hypothesis formation. Such a formulation is purely descriptive in nature and is not "an explicit conceptual relationship between two variables whereby something (an independent variable) "causes" something else (a dependent variable)."[15]

This kind of formulation, which may seem commendable in its caution, fails the "so what?" test. A proper research hypothesis requires the hypothesizing of some kind of causal mechanism operating in the real world so that some kind of tentative causal conclusion can be drawn from the research results if they are valid and the hypothesis test is successful.[16]

Seventeen studies present the research problem in the form of, "what are the characteristics?"[17] For example, **Gartrell** states, "The aim [of this study] was to learn about the homes, families, and communities into which the children were to be born."[18] **McCandish** writes, "The family dynamics and developmental changes within these families and the implications for the psychotherapeutic treatment of lesbian mother families are the subject of this chapter."[19] **Pennington** declares, "The purpose of this chapter is to discuss the major issues confronted by children living in lesbian mother households."[20]

Hypotheses that take the form of descriptions of characteristics face a different problem from that faced by statements of possible differences. A focus on what is "characteristic" of a population (e.g., the mean, median, or mode) can obscure causal relations that are not "characteristic" of the populations studied, but are nonetheless causal in nature.[21] For example, sociologists Sara McLanahan and Gary Sandefur report that 29 percent of young adults from one-parent families dropped out of high school while only 13 percent of those from two-parent families dropped out. Dropping out is not "characteristic" of children from either type of family structure, yet there is little doubt that a causal relationship between the variables of type of family structure and the propensity to drop out of school exists (1994, Figure 1, p. 41). This problem can be put in more general terms. Focusing on characteristics of populations obscures the necessity for a proper research hypothesis to focus on the relationship between two variables and not the properties of each of

them considered separately. In this respect, focusing on the characteristics of an attribute is misleading and hinders the scientific research enterprise. All of these studies are not *a priori* invalid as instances of exploratory research. Compared to studies that state and test the research hypothesis properly, however, they are much inferior in their level of research sophistication and precision. It tells us to look for and expect other problems in later research steps. Authors with such weaknesses in their formulation of hypotheses are unlikely to produce any conclusions sufficiently robust so to inform public policy debates with any degree of dependability.

The Ugly: Affirming The Null Hypothesis

The remaining 18 studies explicitly seek to find no differences between heterosexual and homosexual parents in child outcomes and to make this formulation a kind of hypothesis statement. While this procedure is superior in some way to those used in the other studies, it is also highly problematic because of the difficulties associated with testing hypotheses purporting to affirm the null hypothesis.

The authors of the null hypothesis-affirming studies seek to show either that children raised by homosexual couples are not more likely to grow up to be homosexual themselves than are those raised by heterosexual couples, and/or that they are not more likely to grow up with psychological problems than are children raised by heterosexual couples, or both.

Eighteen studies explicitly seek to find no differences between heterosexuals and homosexuals.[22] For example, **Flaks et al**, in their study of 15 lesbian couples, 15 heterosexual couples and their children, state, "On the basis of prior research, we expected [to find] no differences between the children of lesbian and heterosexual parents in any of the areas evaluated."[23]

In another case, **Huggins** studied adolescent children of lesbians, expecting that a parent's homosexuality would not result in confusion of the child's sexual identity, inappropriate gender role behavior, sexual orientation, and overall psychopathology.[24]

Likewise, **Patterson**'s studies of donor-inseminated lesbian families all start with the expectation of finding no differences between

the children of lesbian and heterosexual parents.[25] The same is also the case with **Tasker and Golombok (1995, 1997)**.

The "no difference" hypothesis used in the 18 studies discussed above inverts the usual social science quantitative research procedure, which would use a positive research hypothesis in the form described above. This creates two major methodological problems that are unrecognized by the all the authors of these studies save one (**Chan et al, 1998**).

1) Failing to reject the null hypothesis necessarily leads to an indeterminate result because one cannot validly "confirm" the null hypothesis, and
2) Inverting the normal hypothesis testing situation makes it too easy to fail to reject the null hypothesis, which is the outcome favored by these researchers.

This results in an undue partiality in interpreting their research findings and in carrying out the research itself. To see all of this clearly, it is necessary to review the usual statistical testing procedure in quantitative social science. This procedure requires statistical testing of a positive research hypothesis.

A simplified example may help to visualize what is involved. For example, suppose a researcher hypothesizes that political liberalism leads to greater support for abortion rights than does political conservatism. One way to test this research hypothesis might be to use a national sample survey of the American public (e.g., data from the General Social Survey produced by the National Opinion Research Center). With this body of data, which consists of individual responses to questions on a questionnaire, one computes the mean "support for abortion" score of liberals and the mean "support for abortion" score for conservatives.[26] One can assume that if this procedure is carried out, liberals will have a higher average score than do conservatives (and in fact, they do). Since it is extremely unlikely that such a comparison will yield exactly the same average score for both liberals and conservatives, one must question whether this finding is a real difference or whether it could be due to chance factors. The difference in averages alone does not provide sufficient information to determining the likelihood. To answer the question, statisticians

have developed ways of distinguishing between statistically significant and statistically insignificant differences. Insignificant differences might be due to sampling error, measurement error, or just random fluctuations. In fact, these are competing claims that ought to be considered in as possible explanations for any research finding. Another way to state the null claim is, that if it is true, any difference found in the data is due solely to random variation. Chance occurrences of this kind do happen; individuals do win the lottery and draw royal flushes in honest poker games.

To ascertain whether in a given instance random variation explains the findings in the data, the researcher carries out a statistical hypothesis test. This requires a research hypothesis[27] and a null hypothesis.

The research hypothesis is of the kind already discussed. The null hypothesis is a hypothesis of no difference or no effect. In the abortion example, the research hypothesis is that liberals are more pro-abortion than are conservatives.[28] The null hypothesis is that there is no difference between liberals and conservatives in their support for abortion rights. Both of these cannot be true. In carrying out statistical hypothesis testing, the null hypothesis is a statistical device that allows for calculation of the value of a test statistic. The test statistic is calculated to determine the probability that the null hypothesis is true given the data at hand. After the calculations are carried out, the test statistic yields some number. If the number calculated from the test statistic is greater than a certain preset value, which is called the critical value (e.g., $t>=2.00$), the null hypothesis is rejected at the associated level of statistical significance (e.g., $p<.05$)[29] and the research hypothesis is accepted.

For example, suppose we find that the mean abortion rights score for liberals is greater than it is for conservatives. In our example, the relevant test statistic is calculated and the results are checked to see whether they are statistically significant or not at the preset level of statistical significance (in fact, there is a substantial statistically significant difference between liberals and conservatives when this test is carried out). Then we reject the null hypothesis and accept the alternative hypothesis that "liberalism" is correlated with "support for abortion rights." Of course, carrying out this one hypothesis test does not end the researcher's task. In fact, the formal hypothesis test

is just the initial step in analyzing the data. The social scientist then has to show that this difference is not due to other factors (for example, due to differences in education among sample members or to selection biases in the sample). However, at least he has a statistical relationship to work with, to try to either explain or explain away in terms of broader substantive considerations.

There are two subsidiary points that need to be made here. First, the reason the statistical test situation is conceived in the above manner is to yield a determinate outcome. If the null hypothesis is rejected, then the alternative hypothesis is accepted. In our example above, we reject the null hypothesis of no difference between liberals and conservative on their support for abortion rights and accept the alternative hypothesis that liberals have a higher average support for abortion rights scores than do conservatives which is statistically significant (and they really do). Also, statistical tests of this kind place the burden of proof on the investigator to show support for his research hypothesis. That is why the criterion for rejecting the null hypothesis is difficult. Thus, social science researchers conventionally use the $p<.05$ level of statistical significance. It does not have to be this stringent (1 out of 20), but the practice has evolved so that it has become the standard social science research convention in the standard positive hypothesis social science research situation.[30]

What happens if the null hypothesis cannot be rejected? In this situation there are always two competing explanations for this result. The first possible explanation for the failure to reject the null hypothesis is that whatever differences are found really are due to chance factors, so that no statistical, let alone causal, relationship between the two variables really exists. A statistician might say that if the test were repeated an infinite number of times, a zero correlation or a zero difference between the two groups studied would result. The examples of the honest poker game and the honest lottery are relevant here. The second possibility is that the research hypothesis is true but its truth cannot be ascertained by the research results because there is some flaw in the study design or in the statistical test itself, which causes the test statistic to yield a statistically insignificant result or p value.[31] Therefore, failing to reject the null hypothesis by itself does not lead to a determinate result. Since every failure to reject the null hypothesis has two possible explanations, one can-

not simply "accept" the null hypothesis in the same way that one can "reject" the null hypothesis and "accept" the alternative hypothesis. Further investigation or conducting a new study is always in order.

This is the problem with the 18 studies that explicitly sought to confirm the null hypothesis as their research hypothesis. These studies sought to prove the null hypothesis, which, as we have shown, is not the same thing as failing to reject the null hypothesis. In substantive terms, their authors seek to show that homosexual parents produce the same child outcomes as do heterosexual parents. This means that they desire to be able to "accept" the null hypothesis as showing that homosexual parenting has no effect on child outcomes simply on the basis of failing to reject the null hypothesis. This violates the standard statistical hypothesis testing procedure. It is wrong because, as we show above, failing to reject the null hypothesis does not necessarily mean that the null is true.[32]

This is not merely a technical flaw in these studies. These investigators report their failure to reject the null hypothesis and falsely conclude that there is no difference between homosexual and heterosexual parents in child outcomes.[33] This false conclusion invalidates the "findings" of no difference between heterosexual and homosexual parents as reported in the research literature that we have surveyed. Only the authors of one study (**Chan et. al, 1998**) showed any awareness of the problem, but they did nothing to correct for it or to alter their interpretations of their results because of it. If the null hypothesis itself becomes the research hypothesis, and some kind of research hypothesis is to become the new null hypothesis, then the standard testing situation must be radically altered to accommodate this situation and non-standard statistical tools are needed in order to reach defensible results.[34] The studies we surveyed all failed to do this or even to indicate that they saw the need for doing it. This indicates that their authors' understanding of the logic of quantitative social scientific research is suspect. When the hypothesis statement is properly conceptualized, the null hypothesis is used in conducting statistical tests as the comparison hypothesis to the one under investigation. It is no substitute for a properly formulated affirmative hypothesis. It is the objective of properly stated hypotheses, proper design, and proper execution of an empirical research

study to decrease the probability that the relationship uncovered by the investigator is due to chance. [35]

The goal of genuine social-scientific research, in short, is to make the null hypothesis less, not more, likely.[36] Properly speaking, then, one can never prove the validity of the "null hypothesis." When you hear the statement that a study found "no significant difference," what this actually means is that, having done some tests, the investigator can only say, "I looked for differences, and haven't found anything significant yet. But who knows?" In social-scientific terms, the study "failed to reject the null hypothesis." It proved nothing.[37]

In summary, in conducting a statistical test of a hypothesis there are two possible outcomes. The first is to be able to reject the null hypothesis and accept the research hypothesis that a difference between the groups does exist that is not likely to be due to chance factors. The researcher then proceeds to see if his or her hypothesis can stand up to other tests of its validity, by introducing controls for extraneous and confounding factors and the like. These are all the subsequent research steps we will be discussing below.

The second possible outcome is to fail to be able to reject the null hypothesis. This is *NOT* the same as showing that no effect exists. There are many possible reasons why one may fail to reject the null hypothesis yet be in error in doing so. For example, the sample used in the study may be too small to reach the appropriate level of statistical significance for a given effect, the significance level used in the significance test itself may be set too high, or the research instruments used to measure the independent and dependent variables may so highly unreliable that no stable results are possible. Even if none of these factors can explain the absence of positive results, this still does not show that no effect exists. The researcher then proceeds to see if his or her non-finding can stand up to additional tests of its validity, by introducing controls for extraneous and confounding factors that might cause a spurious non-correlation (see more below).

Precisely because the usual and correct research procedure is to try to reject the null hypothesis, projects that aim to demonstrate no significant differences between homosexual and heterosexual parents and/or their children face serious problems. If the investigator starts

with the goal of finding no differences, the investigator may be too quick to assert that there is no relationship between the independent and dependent variables, without taking into account all these other aspects of a study that may go wrong. Serious social scientific research is complex enough that the subsequent elements of a study introduce ample opportunities for poor execution. Ironically, each poorly executed research step, such as setting up comparison groups, sampling, measurement, and statistical analysis, increases the likelihood of finding no difference.

What Went Wrong and What Can De Done About It?

Let us restate briefly the lessons from Step 1: Formulating the Hypothesis.

1) A proper hypothesis defines:
The posited cause (or independent variable),
The posited effect (or dependent variable), and
The posited causal relationship between the two.
Only two of the 49 studies do this.

2) A research project that looks for differences, but fails to state its hypothesis in any clearly causal form, is off to a very bad start. This describes 12 studies.

3) A research project that focuses on characteristics, but fails to state its research hypothesis in a standard form, is also off to a very bad start. This describes 17 studies.

4) A research project designed to "prove a negative"–the null hypothesis–is doomed to great difficulty from the start. Yet 18 of the studies take exactly this approach.

If your goal is to prove no differences, you're bound to reach it, and the poorer research you do, the more successful you will be. But there is nothing inevitable about this. There is no reason that proper research hypotheses cannot be formulated in this area of study, and then implemented through various methods that we will now describe.

Notes to Chapter 1

1. A study that is properly thought-out at this conceptual level would not make the kind of mistakes that show up subsequently in much of the research analyzed in this essay, such an improper sampling, neglect of extraneous variables, and attempts to "prove" the null hypothesis.

2. See Rosenberg (1968) Hirschi and Selvin (1973) Cook and Campbell (1979), Davis (1985), Rossi and Freeman (1995), and Nachmias and Nachmais (1997) for discussions of the conceptual logic of social science research.

3. Mathematically, this is stated, y=f(x), where x is the independent variable and y is the dependent variable.

4. Most causal statements in the social sciences exhibit neither necessary nor sufficient conditions. Hypothesis statements are probabilistic in nature because most social phenomena are due to the operation of multiple causes (see the sources cited above for more discussion of this point).

5. This discussion of levels of measurement is relatively crude, but sufficient for this purpose. Generally investigators distinguish between nominal, ordinal, interval, and ratio scales. See Nachmias and Nachmais, 1997, pp. 158-163.

6. Any hypothesis can be translated into one of three statistical forms suitable for direct assessment: 1) a correlation between two quantitative variables, 2) a difference between two qualitative attributes in their mean scores on a quantitative variable, or 3) a difference in (or a ratio of) proportions between two qualitative attributes.

7. Notice that this example begs all sorts of definitional questions concerning homosexual and heterosexual "orientation." Even a superficial examination of the literature reveals that sexual orientation is far more complex than a simple binary attribute like gender. Not only are there different types of homosexuals, but the self-identification question does not distinguish between identities, feelings, and behaviors. Nor does it describe these properties in the life histories of the individuals studied. For a thorough discussion of these questions, see Laumann et al., 1995, pp. 283-286 and passim.

8. Both of these studies are seriously flawed. Pagelow does not focus on children and Miller has no comparison group of heterosexual parents.

9. Pagelow, 1980, p. 191. Unfortunately, Pagelow subsequently used self-constructed and unreported measurements, and did no statistical testing of her self-selected sample. This fatally flaws her study.

10. Miller, 1979, p. 545.

11. Unfortunately, Miller failed to compare his sample to a group of heterosexual fathers and their children. So even though his hypothesis statement is good, his study is fatally flawed. All the steps are important.

12. These studies are: Golombok et al, 1983; Green, 1978, 1982; Green et al, 1986; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lyons, 1983; Mucklow and Phelan, 1979; Bigner and Jacobsen, 1989a, 1989b, 1992; Brewaeys et al, 1997; Golombok and Tasker, 1996; Koepke et al, 1992; Miller et al, 1982; and Bailey et al, 1995.

13. P. 166.

14. Brewaeys et al (1997, p. 1349).

15. See above.

16. Morris Rosenberg usefully distinguishes a number of causal relationships commonly studied in social scientific research including the stimulus-response and property-response relationships. Studies of the consequences of sexual orientation can be classified as (potential) property-response relationships. Rosenberg, 1968, pp. 13-17.

17. The studies that characterize the problem in the form of "what are the characteristics?" are: Barret and Robinson, 1990; Bozett, 1980; Cameron, and Cameron, 1996; Crosbie-Burnett and Helmbrechty, 1993; Gartrell et al, 1996; Hare, 1994; Hoeffer, 1981; Javaid, 1992; Kirkpatrick et al, 1981; Lewis, 1980; Lott-Whitehead, and Tully, 1992; McCandlish, 1987; O'Connell, 1993; Pennington, 1987; Rand et al, 1982; Riddle and Arguelles, 1989; Ross, 1988; Weeks et al, 1975; West and Turner, 1995; Wyers, 1987.

18. Gartrell et al, p. 274.

19. McCandlish, p. 23.

20. Pennington, p. 59.

21. Hirschi and Selvin (1973, pp. 123-125) develop this point at some length when they show that focusing on the characteristics of attributes is incompatible with the fact that social phenomena nearly always exhibit multiple causes.

22. These studies are: Chan et al, 1998; Crosbie-Burnett and Helmbrechty, 1993; Flaks et al, 1995; Green, 1982; Green et al, 1986; Harris and Turner, 1985; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Miller, 1979; McNeill et al, 1998; Miller et al, 1982; Patterson, 1994a, 1994b, 1997; Tasker and Golombok, 1995, 1997; Turner et al, 1990.

23. 1995, p. 106.

24. 1989 p. 124.

25. E.g., Patterson, 1994a, pp. 157-158.

26. This can be set up in many ways, using many different types of data, and many different test statistics. The account given here is illustrative only.

27. Statistics textbooks use the terms null and alternative hypothesis. We prefer the term research hypothesis because the alternative hypothesis is what the researcher is actually interested in.

28. One might object that the research hypothesis stated here is actually one-sided, while the statistical tests described below are two-sided. While strictly speaking this is correct, there is generally little difference in outcomes between the two situations.

29. A two-sided t-test where the null hypothesis is to be rejected if p<= .05 requires a value of 2.00 or greater with a sample size of 50 or more.

30. This is the two-sided 5 percent level of statistical significance. One-sided tests are sometimes used, but they are vulnerable to a certain degree of manipulation in order to achieve the desired result.

31. There are many possible flaws that could produce this result even when the null hypothesis should be rejected and the alternative hypothesis accepted. These include: too stringent a statistical test, too weak a test statistic, insufficient sample size, measurement unreliability, restriction of variable range, suppressor third variables, and inadequate comparison groups. Some of these are discussed below.

32. In fact, there were only five studies with any results worth considering at all. This means that the authors used a heterosexual comparison group and they included a sustained degree of multivariate testing. These are Brewaeys et al,. (1997), Chan and Patterson et al., (1998), Flaks et al., (1995), and Green et al., (1986). Tasker and Golombok (1995; 1997) is noteworthy as being the only longitudinal study of the issue, but the authors failed to analyze their data properly and their samples were vanishingly small.

33. Strictly speaking it is a logically invalid inference.

34. Statisticians have begun to develop practical means for doing this. For example, Donohue describes how an alternative hypothesis of a particular magnitude can be used as a kind of "null" hypothesis. He then shows how a p-like value can be calculated that may permit rejecting of this "null" hypothesis (1999).

35. There is always the probability, albeit sometimes extremely small (for ex-

ample, one out of 10,000 instances), that any relationship uncovered by the investigator is due to chance. Needless to say, the smaller the probability that the relationship is due to chance, the greater confidence we have in its reality, all other things being equal.

36. This is because the social sciences, like all other sciences, are interested in discovering laws governing the phenomena in question. There are presumably an infinite numbers of ways of getting this wrong.

37. The late Jacob Cohen, a leading psychometrician, pointed out that any attempt to prove the null hypothesis "is always strictly invalid." Cohen, 1988, p. 16. One can only calculate the proper test statistic and then either reject it, or fail to reject it.

Chapter 2
Compared to What?
Methods to Control for Unrelated Effects

Now that you have a hypothesis, you put it to work. You need operational definitions.[1] These are the specific recipes that translate the independent and dependent variables that are part of the researcher's hypothesis into real world operations or actions. The process of operationalization, as it is sometimes referred to, involves translating these variables into concrete measurements that can be recorded and analyzed.[2] For example: if your hypothesis involves "gender," you need an operational definition of this concept. If a researcher plans to carry out a reputable survey, there must be box in the survey where a respondent may mark his or her gender.[3]

Without understanding the operational definitions used by a study, it is impossible to know how the hypothesis was tested.[4] Additionally, another researcher cannot replicate the study. To make matters worse, if operational definitions are imprecise or erroneous they will have one predictable effect: To increase the probability of "finding" that there is "no difference."

Five Kinds of Controls and Why They Matter

If the goal of a study is to test a hypothesis, and one has operationalized the concepts used in the hypothesis, then the stage is set for the next step. The researcher must impose various controls on the research design to eliminate false answers. While the actual process can be complex, the basic idea is simple: If you want to show that A "causes" B, you need to get other causes out of the way.

Notes for this section begin on Page 53

The five key methods for doing this are (1) use a comparison group, (2) control for extraneous variables, (3) control for suppressor variables, (4) use pair or group matching and (5) use multivariate statistical tests. These methods build upon one another and most valid studies use at least three of these methods.

In this chapter, we will define each method, and look at how it was used, not used, or misused in these studies. In brief, our examination of the 49 studies disclosed:

Eighteen used no control method of any kind
Seven used only one control method
Fifteen used only three control methods
Eight used four control methods
One used all five control methods

Method One: Use Control Groups

As an absolute minimum, a study of whether parent sexual identity affects child outcomes needs a study group and a comparison group.[5] If the independent variable is the sexual orientation of the parent, there must be at least two groups of parents, homosexual and heterosexual.[6] Otherwise it is logically impossible to draw any conclusions about the possible effects of parental sexual orientation. Attributes that do not vary are not variables and can explain nothing. Ideally, the study and comparison groups should differ solely on the single variable of the parent's sexual orientation. The groups should be otherwise identical regarding the parent's level of education, the ages of their children, the ages of the parent, etc. ... These other features are extraneous variables[7] whose influence the researcher strives to eliminate as far as possible. The control group should increase the likelihood that any results uncovered by the investigator are actually based on differences in parent sexual orientation.

It is disappointing to discover, therefore, that 21 studies (43 percent) had no heterosexual comparison group at all. This makes them scientifically invalid from the outset.[8]

For example, **Green**'s study of 37 clinical cases of children raised by transsexual and homosexual parents (1978) lacks even one hetero-

sexual comparison group. **Green** notes this as a "limitation," but goes on to conclude (albeit tentatively), that children raised by transsexual or homosexual parents are no different in their sexual identity than those raised by heterosexuals.[9] But the lack of a control group is more than a "limitation." It makes it impossible for **Green** to offer any scientific generalizations.[10]

Similarly, **Miller (1979)**, after formulating a good hypothesis, fails to include any heterosexual control group in the study on gay fathers and their children. He recognizes the need for carefully controlled studies,[11] but his is not one of them. This does not keep him, however, from making at least two significant claims based on his study: (1) "[T] here does not appear to be a disproportionate amount of homosexuality among the children of gay fathers,"[12] and (2) while the children he studied had "problems of sexual acting-out,"[13] the acting-out is more likely to be a function of divorce, not of the father's homosexuality.[14] This is a reasonable hypothesis. If he used proper methods, we might be able to see if it is true. He does not, however, so his claims have no scientific basis.

Lacking comparison groups, these studies tell us nothing about the effects of differing parental sexual identities on child outcomes. They should be simply disregarded.

The studies by **Charlotte Patterson** use comparisons but have similar flaws. The **Patterson** study uses a method that does not allow controls for extraneous variables.[15] In her Bay Area Lesbian Family Study, **Patterson** collected data on 37 lesbian families. The study then compared scores from the study group with national averages for various psychological tests. The study also statistically compared the lesbians' children's scores on Eder's "Children's Self-View Questionnaire" with 60 children from Eder's original study.[16]

Because there is no control group, **Patterson**'s method makes it impossible to know whether the study's sample of lesbian mothers differs significantly from the study's other samples because of "sexual orientation" or because of something else altogether. Other differences might account for Patterson's findings. Comparing samples is not a substitute for a full-fledged comparison group studied at the same time and in the same way. **Patterson**'s colleagues acknowledge this serious methodological flaw.[17]

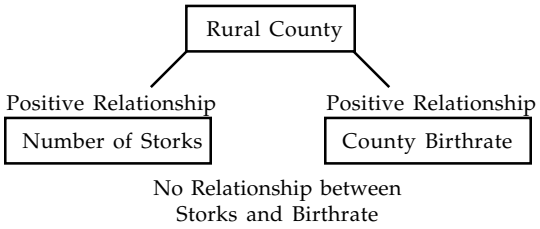
Overall, only 28 studies (if we include **Patterson**'s three studies) actually compared those in a homosexual study group to at least one heterosexual comparison group.[18] Twenty-four studies recruited a separate heterosexual comparison group, and one used a random sample of a general population that included both homosexual and heterosexual parents.[19] Both are common methods of allowing for homosexual/heterosexual comparisons.[20]

Method Two: Control for Extraneous Variables

The study group and comparison group should be identical except for the independent variable. Unless a controlled experiment is possible, however, this is nearly impossible to achieve.[21] Assuming that a controlled experiment is not feasible, either practically or ethically, the investigator should then use some form of control for extraneous variables. This will increase the probability that any changes found in the dependent variable are more likely due to changes in the independent variable, rather than to other variables. [22]

Storks in Sweden. One common example used in statistics classes is the relationship between the number of storks and the birthrate in Swedish counties. Counties with more storks (the independent variable) also had higher birthrates (the dependent variable). If the researcher proceeded mechanically, he or she would erroneously conclude that there is casual connection between storks and babies. Yet there is a correlation. Where does it come from? A third variable: rural-urban differences. Rural areas have both a greater number of storks, and a higher birth rate (see Diagram 2).

Diagram 2.



The only way we know that there is no relationship between the number of storks and the birthrate is because we controlled for urbanization. Controlling for extraneous variables, therefore, is absolutely critical in establishing any kind of causal inference in a non-experimental setting. If extraneous variables are not controlled for, or are improperly controlled for, the investigator cannot conclude that his or her findings have anything (or nothing) to do with homosexual and heterosexual parental identities. Claiming a causal relation (or lack thereof) may be the same as claiming that more storks cause more babies.

Only 23 of the studies (46 percent) control for extraneous variables at all.[23]

The basic rule of thumb is that any time the investigator finds a substantial or significant difference on some third variable between the study and the control group, this third variable must be entered into subsequent statistical analyses as a control variable. Otherwise, the effect will be missed, and the results of the study will then be invalidated.[24] For example, **Golombok et al. (1983)** found a statistically significant difference between their homosexual and heterosexual groups regarding education, psychiatric treatment and contact with fathers. The lesbian mother group had significantly greater levels of education, had more psychiatric treatment, and their children had more contact with their fathers. All these variables should have been taken into account when performing any subsequent analysis. If they are not, all subsequent findings could be a function of the effects of any or all of these variables that produce statistically significant differences. Since these studies did not do this, their results are scientifically invalid.[25]

Method Three: Control for Suppressor Variables

There is also a reverse problem, which is of central importance to the research we are critiquing, that a study needs to address called spurious non-correlation. This happens when a third variable, neither independent nor dependent, causes the false impression that the independent and dependent variables are unrelated. This is called a suppressor variable because it statistically suppresses the truth about the real cause.[26]

The possibility that a suppressor variable is at work increases when (1) there is no relationship (or a weak one) between the independent and dependent variables, and (2) the third variable is positively associated with either the independent or dependent variable and negatively associated with the other variable. Because of its distinctive relationship with the independent and dependent variables, this suppressor variable masks the true relationship between the independent and dependent variables. The statistical effect is to create a false impression that there are "no differences" at work. This false impression, however, can be easily removed from a study if the suppressor variable is controlled for.

Here's an example. Imagine someone putting forth the hypothesis that race (the independent variable) affects the likelihood of voting (the dependent variable). If the study does not use good controls, it appears that African-Americans are less likely to vote than are Caucasians. However, if education is controlled for, the relationship reverses itself—African-Americans are more likely to vote than are Caucasians. This is because: 1.) Education is positively related to the probability of voting (those with more education are more likely to vote than those with less), and 2.) Currently, African-Americans are less educated on average than are Caucasians. Put slightly differently, it appears that race is the key factor in likelihood of voting, but it turns out that the real factor is education. One cause looks real, but it is not. Until controls are used, the other cause is invisible. Then it turns out to be the real cause.

So what? Here's the payoff: If one hopes to show "no difference"–in social science language, that "no effect exists between the independent variable and the dependent variable"–then it is crucial that the study identify and control for misleading suppressor variables. Otherwise real causes, if present, will be missed.[27]

Unfortunately, only one of the 49 studies, **Green et al (1986)**, came close to explicitly addressing the problem of spurious non-correlation, or to controlling for suppressor variables. The Green study incorporates a method for automatically controlling for suppressor effects, although it does so rather mechanically, and does not discuss the problem of suppressor effects in studies that seek to show that "no differences" exist. The **Green** study carries out an initial regres-

sion analysis, using seven variables to predict the child's responses to interviews and tests.[28] Statistically significant variables were included in the final regression model,[29] which automatically includes potential suppressor effects.[30] Unfortunately, having done all this, the Green researchers did not publish their actual regression results, so we do not know which variables were included or dropped. The authors only mention three findings that were statistically significant based on their r^2 values.[31] They do not, however, report any regression coefficients or standard errors, which would allow the reader to perform his or her own significance test. This exclusion is extremely odd. The common statistical practice is to provide a table of regression coefficients (and either their standard errors or their t-statistics) when undertaking a multiple regression analysis.[32] The table displays all variables entered into the equations, and discloses all subsequent results, significant and non-significant, in a table so other investigators can examine the results and reach their own conclusions.[33] As a result of the table's exclusion, there is no way that anyone can treat **Green**'s conclusions as scientifically valid. None of the other studies on homosexual parents control for suppressor variables. But does this really matter? Let us examine what difference it made in several studies that failed to address this issue. We will consider two potential suppressors: prior psychiatric treatment and parental education.[34]

Prior Psychiatric Treatment as a Potential Suppressor. For example, the studies of lesbian families in **Golombok (1983)** found prior psychiatric treatment to be statistically significantly higher for the lesbian group than for the heterosexual group.[35]

The studies, however, did not include prior psychiatric treatment as an extraneous variable to be analyzed simultaneously with the independent variable. Why not is puzzling, but their logic is quite simple. If lesbianism is associated with seeing a therapist, and if parental therapy is positively associated with child outcomes (which could be the case if it is effective), then it follows that the "non-relationship" between homosexual orientation and child outcomes **Golombok** claims might be spurious. The truth might be different.[36]

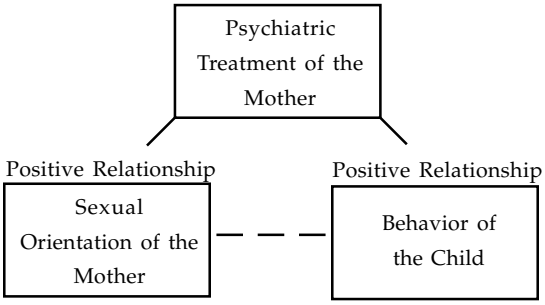
Diagram 3

Simple Bivariate Relationship
No Differences Found

If suppressor variable effects are at work, however, things might actually be as follows:

Diagram 4

Potential Third-Variable Suppressor



Suppressed Negative Relationship Between Independent and Dependent Variables?

It could be, in other words, that the real reason there is "no difference" is that the lesbian mothers were more likely to have had prior psychiatric treatment. By not controlling for that suppressor variable, any possibly negative relationship between a mother's lesbianism and her child's behavior would be effectively shielded from view. All statistical analysis performed by **Golombok et al (1983)** and **Tasker and Golombok (1995, 1997)** should automatically have included, at minimum, any statistically significant extraneous

variables. This would reveal any potential negative relationships between the "sexual orientation" of the mother and child outcomes. Without this analysis, one cannot conclude that there are "no differences," only that there is no information.

Education as a Potential Suppressor Variable. Even the most statistically sophisticated homosexual parenting study we examined overlooks the importance of suppressor effects in explaining their findings. In their study of families created via donor insemination, **Chan et al (1998)** find a statistically significant relationship between educational level and a parent's "sexual orientation." Homosexual parents have significantly more education than heterosexual parents. This is true when comparing the lesbian biological mothers with the heterosexual mothers, and when comparing the lesbian social mothers with the heterosexual fathers. **Chan et al** also find no significant differences in subsequent analyses of the relationship between parent's "sexual orientation" and the child's behavior, when controlling for relationships within the family and the parent's mental health. Parental education, however, is a potential suppressor variable because many child outcomes are positively related to parental education. If lesbian and heterosexual mothers produce similar child outcomes to their heterosexual counterparts even where the lesbians are better educated, it may well be the case that when educational levels are properly equated that heterosexual mothers produce more favorable outcomes than do their homosexual counterparts. As far as we can tell, however, education is not entered into subsequent equations, and therefore remains an un-addressed potential suppressor.[37]

In this instance, what is needed is not simply more research on homosexual parents and their children. Instead, we need much better analysis of existing data. When the researcher is hoping to show validly that no effect is present, this analysis should reflect at least a minimal grasp of the role suppressors might play as extraneous variables.

Method Four: Use Matching

In the non-experimental context, such as the studies examined here, there are a number of ways of controlling potential effects of

extraneous variables. Today, the most widely used method is multivariate statistical analysis.[38] Typically, the investigator draws a random sample of respondents, and then statistically controls for the effects of extraneous variables.[39] But there is an important alternative: matching by various methods.

Matching emerges out of the experimental tradition. Classic experimental design consists of an experimental group and one or more control groups. Subjects are randomly assigned to groups. The experimental group receives the treatment, while the control groups does not. The groups vary only in terms of receiving or not receiving treatment and any differences found between the groups is likely to be due to the experimental treatment.[40] For obvious ethical and legal reasons, these experiments cannot be used to investigate certain research topics. A researcher cannot randomly assign babies to different family structures, nor can he or she intentionally give cancer-causing substances in conducting an experiment. To get around these and other limitations,[41] while still properly comparing study groups and control groups, social scientists have developed quasi-experimental designs for selecting respondents (e.g., Campbell and Stanley, 1966 and Cook and Campbell, 1979).

Matching is often used when the subjects to be investigated are a rare or hard-to-reach population. If researchers were looking for adult children raised by gay fathers, common forms of random sampling would yield huge numbers of people that would not fit the criteria. To avoid such a waste of time and money, researchers often resort to one of two methods of matching: pair matching or group matching. Of the 49 studies examined here, 23 used some form of matching.[42]

For reasons that we explore below, group matching used alone is both unwise and inhibits the complex analysis necessary to fully explore and understand the data.[43] Where a random sample is not possible, pair matching should be the alternative method used.

Pair matching involves matching pairs of respondents based on variables believed to affect the outcome variable. This is easy to visualize in a two-group situation. One person in a matched pair is assigned to the study group, the other to a control group.

Consider this example, taken from sociologists Peter Rossi and Howard E. Freeman (1994), of how one might study the effects of school vouchers. The investigator first chooses particular children as voucher recipients. Then the investigator draws another child from a pool of students not receiving the voucher, to be paired with the voucher recipient. The paired students would be as identical to each other as possible, in terms of age, sex, number of siblings and father's occupation.[44] This is pair matching.

These techniques are used in a number of studies we examined. **Bigner and Jacobsen (1989a)** pair-matched their samples to obtain precise controls, but when they did this, they depleted much of their comparison group. They had 33 homosexual males in their study group, and created the comparison group from a sample of 1,700 fathers, presumed to be heterosexual, drawn from a larger research project on the social competency of children.[45] From this sample they computer-matched the heterosexual respondents with members of the study group according to age, marital status, income, ethnicity and education. This left a final sample of only 33 comparable respondents.

A study by **Green et al (1986)** appears to have pair matched their subjects.[46] They drew their control group from an initial pool of 900 single heterosexual mothers and matched them to the lesbian mothers' age (+ 5 years), race, children's sex and age (+ 12 months), length of time separated from child's father, mother's current marital status, current family income, mother's education, and length of absence of an adult male in the household.[47] This left 50 lesbian mothers and 40 single heterosexual mothers.

Pair matching is superior to group matching, as we will discuss below. As the above example suggest, pair matching can result in a dramatic loss of cases when comparable individuals cannot be found, since cases that are not matched are often dropped from the analysis.

Group matching is sometimes referred to as "block" or "aggregate" matching.[48] Group matching creates study and comparison groups with the same proportion of men and women, or rural and urban inhabitants, depending on the study.

How to Do it Well: Group Matching by Bell, Weinberg, and Hammersmith.

The classic study of homosexuals by Bell, Weinberg, and Hammersmith illustrates the difficulties of obtaining samples with sufficient controls if one relies on group matching.[49] Considerable effort was first spent expanding the pool of potential homosexual respondents. "Individuals who offered to be interviewed were placed in a 'recruitment pool' and assigned to various categories, or cells, on the basis of their recruitment source, age, gender, educational level and race."[50] The team started with a recruitment pool of 3,438 Caucasian homosexual males, 675 Caucasian homosexual females, 316 African-American homosexual males, and 110 African-American homosexual females. They ended up with final homosexual samples of 575 Caucasian homosexual males, 229 Caucasian homosexual females, 111 African-American homosexual males, and 64 African-American homosexual females.

To obtain heterosexual comparison groups, the investigators used probability sampling with quotas, a procedure that was developed by the National Opinion Research Center at the University of Chicago. First the team used random sampling to select census tracts. Then they randomly selected city blocks within the chosen tracts. Quota sampling was then used for each block. This meant that the interviewer determined if a person satisfied the requirements to be included in the control groups, by seeing if the heterosexual person fell into one of 48 needed categories based on age, gender, education level and race.[51]

The interviewers continued to find respondents until they reached the pre-set number of cases for a particular category. For example, since 25 percent of the Caucasian homosexual male sample received a high school diploma or less education, the interviewers needed to obtain a final sample of Caucasian heterosexual males where 25 percent also obtained the same level of education (which they obtained). The care with which they obtained their block-matched samples is seen in Table 1. The percentages of study and comparison groups are matched in terms of the levels of education obtained.

Table 1.
Study and Comparison Groups in Bell, Weinberg, and Hammersmith's Study of Homosexuals.[52]

Education	White-Gay Male	Black-Gay Male	White-Les. Female	Black-Les. Female	White-Les. Male	Black-Het. Male	White-Het. Female	Black-Het. Female
≤High School	25%	27%	23%	27%	25%	32%	25%	28%
Some College	33%	52%	31%	55%	33%	40%	33%	46%
≥Col. Degree	42%	21%	45%	19%	42%	28%	42%	26%

Bell, Weinberg, and Hammersmith were able to obtain almost the same distribution between Caucasian homosexual and heterosexual groups. For example, 25 percent of Caucasian homosexual males have high school degrees or less. They were able to obtain a sample of Caucasian heterosexual males where 25 percent also high school degrees or less. In their final sample of Caucasian homosexual men, 42 percent had a college degree or better. They managed to obtain a quota sample of Caucasian homosexual males where 42 percent of Caucasian heterosexual males have college degrees or more. Likewise, 23 percent of Caucasian homosexual females have a high school degree or less. They obtained a sample of Caucasian heterosexual females where 25 percent have a high school degree of less, and so on.[53]

Precision, however, is sacrificed as the subgroups become more difficult to obtain. Even with their systematic quota sampling, the investigators were unable to obtain the same percentages of African-American heterosexual males and females for the three levels of education compared to African-American homosexual males and females.

The quality of aggregate matching in controlling for extraneous effects is as good as the precision of the match. Less than precise matching introduces biases into the design. This increases the prob-

ability that the results will be non-significant. A process by which the investigator seeks to obtain roughly equal averages between groups (mean age of study group and mean age of control groups) is not sufficient. The distribution of cases across categories must be as identical as possible.

How The Parenting Studies Did It: Not So Well.

Twenty-two studies relied on group matching as one means of controlling the potential effects of extraneous variables.[54] Of these, **Koepke et al, (1992)** compares lesbian couples with and without children, while **Riddle and Arguelles (1989) and Turner et al (1990)** compare gay and lesbian parents without any heterosexual control group.

We will examine a few studies that used group matching to highlight problems associated with the technique. Studies that solely rely on group matching without subsequent statistical controls are highly inferior to those that use both techniques.

Brewaeys et al appears to be a study with aggregate matching on extraneous variables. Their study group consists of 30 lesbian mother families, each with a child between 4 and 8 years old conceived by donor insemination (DI). **Brewaeys et al** had two control groups. One comparison group was made up of all heterosexual DI families with a child born between 1986 and 1990. The investigators made no attempts to match for extraneous variables for the heterosexual DI group, thereby seriously compromising the design of their study. The other comparison group was of heterosexual naturally conceived families, "matched as closely as possible with respect to the age of the biological mother, age of the child, family size and birth seniority" (oldest child was in the study). The exact mechanism of matching is unclear. We assume they mean that if the age of the mother and of the child and the size of the natural family fell within the range of the lesbian family, the family was included in the study. We do not know if the distributions are the same.[55]

Flaks et al (1994) also group-matched respondents on several extraneous variables. **Flaks et al** studied 15 lesbian families and 15 heterosexual families. Respondents were included if they had one child

between the ages of 3 and 10. In addition, families were matched "on the variables of sex, age and birth order of the children as well as on race, educational level and income of the parents."[56] Although they do not explicitly indicate their method, we assume **Flaks et al** mean "group matched" because **Flaks et al** use the wrong statistical procedures if they pair-matched.[57]

Flaks et al present some data regarding these extraneous variables.[58] The "matching" is extremely imprecise, since one is supposed to have roughly the same proportion between study and control groups in each category for the extraneous variable. For example, **Flaks et al** report the mean ages of parents. The lesbian biological mother and the lesbian "social" mother are on average 2.2 years and 3.6 years older than the heterosexual father and the heterosexual mother respectively. **Flaks et al** do not report the age distribution for the four subgroups (lesbian biological mother, lesbian social mother, heterosexual father, heterosexual mother).

The imprecision is even more evident if the distributions of these variables are examined, and if data are presented as percentages, not as raw numbers. In terms of education, **Flaks et al** report raw numbers, which gives the appearance of being roughly the same, because the sub-samples are so small. We have converted them to percentages to show the imprecise nature of their group matching process.

Table 2.
Education level, **Flaks et al's** sample (15 adults in each subgroup, converted by us from the reported raw frequencies[59])

Education	Bio. Mom	Social Mom	Het. Mom	Het. Father
High school Diploma	0%	0%	7%	7%
Some College	20%	13%	0%	7%
College Degree	13%	27%	7%	20%
Some Grad School	0%	7%	27%	7%
Grad Degree	67%	53%	60%	60%

The distribution of educational attainment among the four types of parents does not closely match. This level of "matching" means that substantial bias is introduced into the study, increasing the probability of finding non-significant results.

Even less precision is found when we examine employment between the homosexual and heterosexual parents in Table 3 (15 adults per subgroup). The distribution of employment regarding heterosexual and lesbian biological mothers is not a close match either. None of the lesbian biological mothers are stay-at-home mothers, compared to 27 percent of the heterosexual mothers.

Table 3.
Type of Employment in **Flaks et al (1995),** converted into percentages.

Type of Employment	Bio. Lesbian Mother	Social Lesbian Mother	Heterosexual Mother	Heterosexual Father
Homemaker	0%	7%	27%	0%
Outside Home	93%	93%	73%	100%
Unemployed	7%	0%	0%	0%

"Matching" for individual income also illustrates other problems with group matching (see Table 4). The precision of matching and the distribution of respondents across categories is also a function of how many categories the study has. **Flaks et al** introduce another source of bias in their study that, in turn, increases the chance of non-significant results. They divide income into only two groups (above and below $55,000). This probably also masks major disparities in the distribution of individual income. Since 27 percent of heterosexual mothers do not work, can we assume that a similar proportion of heterosexual mothers have $0 individual income, compared to none of the biological lesbian mothers and only a small percentage of the social lesbian mothers. The distribution of individual income for heterosexual fathers, nevertheless, appears much higher than that of either lesbian parent.

Table 4.
Individual Income in **Flaks et al (1995)** converted into percentages.

Individual Income	Bio. Lesbian Mother	Social Lesbian Mother	Heterosexual Mother	Heterosexual Father
$0-$54,999	53%	73%	93%	47%
$55,000+	47%	27%	7%	53%

In short, the matching of respondents in **Flaks** is noticeably less precise than that used by Bell et al. Less precision regarding matching of extraneous variables naturally and strongly increases the likelihood of finding non-significant results.

The degree of control used in the parenting studies reviewed here is primitive. Sociologists and noted evaluation researchers Rossi and Freeman report "[m]atching has been supplanted to a considerable extent by the use of statistical controls."[60] Biostatistician Joseph Fleiss, in a leading textbook on epidemiological statistics and research design, recommends that "[m]atching should . . . be on a small number of characteristics (rarely more than four and preferably no more than two), with each defined by a small number of categories. . . If the investigator insists on controlling for biasing factors simultaneously, multivariate [statistical] methods . . . have to be used."[61]

Method Five: Supplement Matching with Statistical Tests

Because matching subjects for rare populations is so difficult and so limiting, investigators should include additional statistical tests. At a minimum, differences between groups on various extraneous variables should be controlled not only by matching but also by multivariate statistical analysis.

Eight of the studies that use matching fail entirely in this regard. They relied solely on group matching, without using supplementary statistical analysis to check for extraneous variable effects. That is, at

a minimum, no attempt was made to see if there was a statistically significant difference between the homosexual and heterosexual groups on the variables.[62]

The other 15 studies that used matching fare somewhat better in terms of at least checking for differences. They used some form of statistical check on their group matching to see if the extraneous variables were significantly related to parent's sexual orientation.[63]

Green et al (1986)'s study is the best in terms of choice of method for controlling for extraneous effects. As discussed earlier, **Green et al (1986)** used pair matching supplemented by statistical analysis of extraneous variables. They chose to control for extraneous effects via multiple regression techniques—the optimal method for doing so.

The other 13 studies rely on t-tests or chi-square statistics to find variables that are statistically significant, a method that is not as good as directly entering the variables in a multiple regression equation, since it makes it impossible to pick up interaction effects between the independent and extraneous variable.[64]

Despite using statistical tests to check for significant differences regarding extraneous variables, a number of the important studies failed to take the next step, which is absolutely critical to avoid invalid results from the data analysis. If the investigator should find statistically significant differences on these extraneous variables, the proper procedure should be to then enter these extraneous variables into subsequent prediction equations. Finding statistically significant differences but not entering them in subsequent statistical analysis invalidates the later analyses.

Despite finding significant differences on several extraneous variables (e.g., educational levels between lesbian and heterosexual parents, annual household income between couples versus singles), **Chan et al (1998), Miller et al (1982), and Golombok** and her various colleagues **(Golombok and Tasker, 1996; Golombok et al, 1983; Tasker and Golombok, 1995; 1997)** failed to enter these variables into subsequent statistical analyses. This makes their later analyses and conclusions invalid. **Hoeffer (1981)** finds no significant differences using a t-test on marital status, educational level,

and occupation, but she finds a significant difference in support for feminism (lesbians being more supportive of feminism). **Hoeffer**, however, fails to enter feminism as the extraneous variable in the subsequent analysis, thus invalidating the overall project.

In general, the overwhelming bulk of the studies either failed to control for these most basic demographic variables or controlled for them improperly. We will now discuss some more specific examples, this time looking at how the studies treat specific variables.

Putting It All Together: Variables, Matches, and Statistical Tests

Rossi and Freeman (1995) provide a useful list of *a priori* characteristics for which investigators frequently match cases and groups.[65] If the investigator is looking at individuals, common extraneous variables include: age, sex, education, socio-economic status (income, wealth), occupation (prestige), ethnicity (race, language groups, religion), intellectual functioning (cognitive ability, knowledge), and labor force participation. When controlling for household characteristics, the investigator should look at life-cycle stage, number of household members, number of children, housing arrangements, socio-economic status of members, and ethnicity of members.[66]

We examined the following extraneous variables to see how many studies either group matched, pair matched, and/or statistically controlled for them. These potential extraneous variables are: gender of the child; educational level of the parent, the occupation; income, socio-economic status or social class of the parent; the partnership status of parent (living alone, living with a partner); and the age of child.

Child's Gender. The child's gender is the extraneous variable most frequently controlled. Twenty-one studies report data for girls and boys separately, or directly control for the child's gender.[67]

The child's gender is an important extraneous variable, but controlling for the child's gender highlights the single biggest problem

with all these studies: The samples are too small for the statistical tests used. A credible study of homosexual parenting must have at minimum, a heterosexual control group. When the investigator controls for the gender of the child, the sample is further divided, to get four sub-samples. Unless the investigators obtain sufficiently large sub-samples, the investigators will probably obtain non-significant results, even if the null hypothesis is in reality false.[68]

For example, **Flaks et al (1995)** compare boys and girls conceived through donor insemination and raised by lesbian versus heterosexual couples. The investigators have eight girls and seven boys in each group. Given Flaks et al's two independent variables (sexual preference of mother and child's gender), the probability of finding statistically non-significant differences is extremely high.

Since all these studies have small samples (we will discuss this issue later), the dilemma of introducing extraneous variables, as illustrated by using child's gender, is this: As one introduces more extraneous controls, one increases the probability of finding non-significant results because the samples are small (and because the statistical procedures use up available degrees of freedom). The investigator may arrive at non-significant results as an artifact of small samples; the non-significant results may falsely mask the real relationship. There is no way around this problem except to increase sample size.

If the investigator chooses not to control for third variables, so as to maximize the sample size that is statistically tested, the results will be unconvincing, precisely because these alternative variables were not introduced. There is no way the investigator can know whether the relationship between the independent and dependent variables is in fact due to a third variable that accounts for both the differences in the independent variable and differences in the dependent variable.

Education. As discussed in the sections on suppressor effects, education is a critical variable that is conventionally incorporated in controlling for extraneous effects. We used a fairly generous definition of "controlling for education." It included 1) whether the investigators used education as a potentially confounding variable

statistically, and/or 2) whether the investigator screened respondents for levels of education and then noted whether there was a difference or not between the study and comparison groups.[69]

Fourteen studies, while testing for differences in levels of education, fail to enter education into subsequent statistical analysis despite the fact that education is found to be statistically significant.[70] Note that all these studies generally find no differences between the homosexual study group and the heterosexual control group regarding the dependent variable, but find a significant difference in education between the study and control groups. This should have led the investigators to search for suppressor effects. That is, this situation raises the possibility that the positive relationship between the homosexual study group and education is masking the negative relationship between sexual preference of the parent and the dependent variable.

For example, **Golombok** and colleagues studied 27 lesbian mothers and their 39 children, plus 27 single heterosexual mothers and their 39 children, first in 1976-1977 and then again in 1992-1993.[71] Their analysis of the 1976 data found education level to be significantly higher for lesbian parents compared to heterosexual parents. Despite this significant difference, education level is not entered into subsequent statistical analysis, thus making reported results invalid.[72]

Subsequently, the adult children in the 1992-1993 study was compared with respect to age, gender, ethnicity and education level,[73] with no statistically significant differences found among the adults (although the adult children from the lesbian group have generally a higher level of education). The subsequent studies do not test for differences among the mothers of these adult children, despite the fact that the initial 1976 sample showed a statistically significant difference in education (and other variables) between the lesbian and heterosexual groups. This is a major error. It increases the likelihood that non-significant results are due to the presence of mother's education level as a suppressor.[74]

The same practice of finding significant differences in education but not subsequently entering it into a multivariate statistical analy-

sis is found in **Miller et al's** study of 34 lesbian and 47 heterosexual mothers. **Miller et al (1982)** also found lesbian mothers having higher levels of education compared to heterosexual mothers. Despite the differences being statistically significant, the investigators failed to enter education (and other statistically significant variables) into their subsequent statistical analysis. This is another case of ignoring possible suppressor effects. Failure to enter education into their statistical analysis increases the likelihood that the claims of "no significant differences" between lesbian and heterosexual mothers regarding their views of the caregiving role are not valid.

Brewaeys et al, (1997) found significant differences in educational level among their groups, and did enter education level into their subsequently statistical analysis as a proper control variable. In their study of 30 children of donor inseminated (DI) lesbian couples, 38 children of DI heterosexual couples, and 30 children of naturally conceiving (NC) heterosexual couples, Brewaeys et al found the lesbian couples to be significantly better educated than the DI and NC heterosexual parents. They found no significant differences in behavioral adjustment, the child-parent relationship, and in gender role development. The problem with the study, however, is also one of small samples. Smaller samples increase the likelihood of non-significant results, even if a real relationship exists. Brewaeys et al further add to the probability of finding non-significant results by recruiting as comparison groups an unmatched DI heterosexual sample but a matched NC sample.

Occupation, Socio-Economic Status, or Social Class. Occupation is often an extraneous variable, in many ways similar to educational attainment. Occupational prestige, like educational attainment, is a source of human capital, enabling individuals to better function in modern society. The studies failed to explain how they defined their occupational categories. For example, **Hoeffer** finds that 65 percent worked in a "white collar" occupation. What does "white collar mean?" What occupations make up "white-collar" versus "blue collar?" These classificatory schemes should rely on conventional measures for occupational stratification, but none of the studies referring to occupation use them. This is poor technique.[75]

Eighteen studies controlled for occupation, income, socio-economic status, or social class as a variable.[76] The studies by **Golombok** and colleagues relied on "social class" as the comparable British category. In these cases involving British respondents, there is a standard measure of social class, which they used. In American cases, investigators poorly control for this variable by claiming they found no "class" differences between their lesbian and heterosexual groups regarding class background or state that their subjects are "middle class" or "upper middle class," without a proper operational definition as to what this means.

The most detailed occupational classificatory scheme is found in **Patterson**'s Bay Area Family study.[77] She extensively classifies the occupations of the mothers—professional occupations, technical and mechanical, business and sales, and others (e.g., artist), although the classificatory scheme and the logic behind it is not described.[78] Unfortunately, **Patterson** fails to have a proper heterosexual control group with which to compare lesbian mothers, so the demographic descriptions of the lesbian mother study group are scientifically useless.

In **Patterson**'s study, occupation is a very serious potential suppressor variable. While 62 percent of the sample is classified as having professional occupations, only 28 percent of the national adult population is employed in the professional-managerial occupations.[79] Yet, **Patterson** compares test scores from her sample to national averages. Because **Patterson** finds no differences in mean scores between the children of the lesbian mothers compared to national averages, one must raise the suppressor variable issue. Since **Patterson**'s sample is significantly higher in occupational status as compared to the national population, occupation is likely to be acting as a major suppressor variable. **Patterson** cannot respond scientifically to these criticisms of her studies regarding occupation (and other variables), because the study fails from the start to use a proper comparison group. Her study's conclusions, as they stand, are not valid.

Partner Status of Parent. All future homosexual studies should follow the design used by **Chan et al (1998)** to control for partner status. Studies should use two study groups and two comparison

groups: families headed by a lesbian couple, families headed by a single lesbian mother versus families headed by a heterosexual couple, and families headed by a single heterosexual mother. This four-group design allows simultaneous comparisons of groups based on the sexual identity of the parent and the number of parents in the household. Alternatively, studies should pair match subjects on whether they live with someone or live alone, and statistically adjust accordingly.

The traditional two-parent family whose child is biologically related gets the short shrift in these studies. Only four studies (**Brewaeys et al, 1997, Flaks, et al., 1995; Miller, 1982, and Mucklow, 1979**) compare lesbian parents to the traditional heterosexual household. **Miller (1982**) and **Mucklow (1979)** fail to distinguish lesbians living with and without partners.

Brewaeys et al (1997) create a good series of comparisons. They compared the lesbian parents and their child created through donor insemination, with a heterosexual couple and their child created through donor insemination and with a heterosexual couple and their child created the traditional way.

Regarding comparisons with the traditional family, **Chan et al (1998)** explicitly rule out making comparisons with the traditional family, where a married couple are raising their biological children. This rules out any possible generalizations based on **Chan et al's** findings to the larger population, since they only study donor-inseminated families.

Chan et al argue that the child's relationship to parents in the traditional family creates a situation of "ownness," and thus cannot be compared to the other types of family structure involving donor-inseminated children. This is a mistake. It would have been a better research strategy to have included the traditional family as another comparison group, as well as the single divorced mother and her traditionally conceived child.

These authors' approach implicitly concedes that the traditional family, a heterosexual (monogamous) husband and wife couple raising their own biological children, is the optimal form of family structure. This apparent concession contradicts the more general

paradigm of research studies carried out one of the above study's co-authors, Charlotte **Patterson**, who has argued elsewhere that the children of lesbian and gay parents "develop in a normal fashion."[80]

The other studies with control groups seriously compromised their studies by mixing lesbians living alone and lesbians living with partners in one study group, against heterosexual mothers living alone.[81] Another study compared a mixed group of lesbians with married heterosexual mothers (**Miller et al, 1982**), and two compared a mixed group of lesbian and heterosexual mothers (**Lewin, 1982; Lyons, 1982**).

The studies that used partnered and non-partnered parents in one group and partnered or non-partnered parents in the other group increased the likelihood of finding non-significant results. Controls should have been included either at the beginning of the study, while finding respondents, or subsequently in statistical analysis.

Age of Child. The age of the child as an extraneous variable is extremely important when the dependent variable is related to the child's development as most of them are in fact. For example, it would be developmentally appropriate to ask respondents for sexual preference if they were adolescents but not if they were young children. This variable should have been directly controlled via statistical testing, thus avoiding the problem of improper (i.e., imprecise) matching.

Only **Green et al, 1986** pair matched his subjects, then statistically controlled for the age of the child through subsequent multiple regression analysis along with a host of other possible extraneous variables.

Twenty studies controlled for the age of the child through group matching and/or statistical testing.[82] None found significant differences in the children's ages and therefore, none used the variable in subsequent statistical analysis.

Flaks et al, (1995) report the mean ages of the two groups of children, without statistical comparisons. Others deal with the age of children improperly. **Kirkpatrick et al, (1981)** report that the ages of children are similar, but provide no numbers. **Javaid, (1992)**

only presents the age distribution of the children. **Lewin and Lyons** report the age range of the children in the study as a whole, but nothing more.

Patterson (1994a, 1994b, 1997) also reports the children's mean ages when statistically comparing scores of Eder's sample of 60 children (5.5 years) with her sample of 35 (6 years, 2 months). She performs a t-test to see if the scores between Eder's sample and hers differ, but does not control for the child's age.

There is one other point to make regarding the treatment of the ages of the children. Only two studies look at adult children of male homosexuals. One, by **Bailey et al**, studies the adult sons of gay fathers.[83] **Bailey et al** found that 9 percent of the adult sons of homosexuals are gay, as reported by the sons themselves or by their fathers (when the sons would not respond to the survey). Another, by **Miller (1979),** looks at the adolescent and adult children of gay fathers, but like **Bailey et al, Miller** also fails to compare them to the proper heterosexual control group. Nevertheless, **Miller** still concludes, "Evidence in the children's biographies pointed to problems of sexual acting-out." This included premarital pregnancies, abortions, prostitution, etc. Because there is no comparison group, we cannot tell it if this acting out is a function of divorce, or an interaction of divorce and having a gay father, although the findings are somewhat suggestive.

The other group of studies looked at adult children of lesbian versus single heterosexual mothers, after initially studying these same children when they were much younger. **Tasker and Golombok** attempted to locate the initial group of children of lesbian versus heterosexual mothers that they had studied in 1976-1977. They re-interviewed them as young adults, on a wide series of topics including the adult children's sexual behavior, desires and sexual preference.[84]

Despite the ridiculously small samples they used, making it extremely difficult to obtain any statistically significant results, **Tasker and Golombok** managed to find statistically significant differences between the two groups. They find that adult children of lesbian mothers were significantly more likely to think about having homo-

sexual relations than were the adult children of heterosexual mothers. They also find that two women raised by lesbians were both in a lesbian relationship and identified themselves as lesbian, while no women raised by heterosexuals were either in a lesbian relationship or considered themselves to be lesbian.[85] These findings are not properly explained by the authors and contradict the "no-homosexual parent effects" view favored by these authors in their writings.

What Went Wrong and What Can Be Done About It?

Here are the lessons from Step 2: Controlling for Unrelated Effects. Recall that the question is, "Do these studies use methods that justify their asserted scientific conclusions about whether or not sexual orientation has any impact on childrearing?" To answer this, we have looked at the general methods that social science would require.

1) You must use a comparison group to draw valid conclusions about the possible effects of something on something else. Twenty-one of 49 studies had no control groups.

2) You must control for extraneous variables in order to eliminate false causes because correlation need not mean causation. Twenty-three of the 49 studies have some kind of control. But anytime you find a significant difference on some third variable, you must enter it as a control into subsequent analysis in order to obtain valid results. Of the 12 studies that found this, only one took this step.

3) You must control for suppressor variables in order to eliminate false (but true-looking) causes because non-correlation need not mean non-causation. Of the 49 studies, only one even came close to addressing this issue, and that study failed to even report which variables were dropped or added in its analysis.

4) You might use some form of matching. If you cannot use full-blown statistical analysis, you need an alternative basis for discerning what is or is not a significant difference. The

least accurate method, group matching, was used in 23 studies (two of which also used pair matching). The best method, pair matching, was used by only three studies. Twenty-seven studies used no form of matching at all.

5) You must supplement matching with multivariate statistical analysis, to test your matching and to deal with suppressor variables. Of the 23 studies that used matching, only 15 did this.

If a study does not use all five methods, it is badly flawed research. But our findings are that only one of the 49 studies used all five methods, **Green, et al (1986).** Even this study failed to explain its methods or justify its conclusions, as we discussed. In short, these 49 studies were conducted with control methods that are so inadequate that they cannot be relied upon for either scientific conclusions or public policy reforms.

Notes to Chapter 2

1. A useful discussion of operational definitions is given in Nachmias and Nachmais, 1996, pp. 30-32.

2. Ibid.

3. Needless to say, operationalization of variables is often far more complicated than this.

4. For instance, in their authoritative study of American sexuality, Laumann, et al. (1994) the term "homosexual" can be defined by means of desire, behavior, self-identification, or a mix of the three. Depending on the operationalization, the results vary. Laumann, et al. (1994), Chapter 8. Note also that this does not deal with the question of bisexuality.

5. Sometimes these are called the "experimental" or "treatment" group and the "control" group. Since neither homosexuality nor heterosexuality can or should be regarded as an "experiment" or "treatment," however, we use the term "study group" to refer to the primary population of the study.

6. These need not be explicitly selected groups, but can result from a random sample survey. With rare populations, however, selection of members of the rare group at a disproportionate rate is highly desirable. The issue is discussed at great length in Chapter 4.

7. Or confounding variables.

8. The studies lacking a heterosexual comparison group are: Bailey et al, 1995; Barret and Robinson, 1990; Bozett, 1980; Crosbie-Burnett and Helmsbrechty, 1993; Gartrell et al, 1996; Green, 1978; Hare, 1994; Koepke et al, 1992; Lewis, 1980; Lott-Whitehead and Tully, 1992; McCandlish, 1987; Miller, 1979; O'Connell, 1993; Pennington, 1987; Rand et al, 1982; Riddle and Arguelles, 1989; Ross, 1988, Turner et al, 1990; Weeks et al, 1975; West and Turner, 1995; and Wyers, 1987. Bailey et al (1995) compare adult sons of gay fathers with gay monozygotic and gay dyzygotic twins. Koepke et al (1992) compare lesbian mothers with childless lesbians. Riddle and Arguelles, 1989, Turner et al, 1990, West and Turner, 1995, and Wyers, 1987 compare gay versus lesbian parents.

9. p. 696.

10. Ibid.

11. p. 548.

12. p. 547.

13. Ibid.

14. Ibid.

15. Patterson, 1994a, 1994b, Children of the lesbian bably boom: Behavioral adjustment, self-concepts, and sex-role identity in Green, B.T., Herek, G.M. (eds.) *Lesbian and gay psychology: Theory, research, and clinical applications.* 156-275 1997, Lesbian mothers and their children: findings from the Bay Area Families Study in J. Laird and R.J. Green (ed.) *Lesbian and gays in couples and families: A handbook for therapists* (pp. 420-436). New York: Jossey-Bass.

16. Patterson, 1994a, p. 161. This statistical comparison, using a t-test, between Patterson's sample and Eder's original 60 child participants could be done, because Eder provides the means and standard deviations of his group of 60 children, Patterson, 1994a, p. 161.

17. The point is conceded by Chan et al. (1998) who use a comparison group of heterosexual donor-inseminated parents in their study. This study is discussed below.

18. These studies are: Bigner and Jacobsen, 1989a, 1989b, 1992; Brewaeys et al, 1997; Cameron and Cameron, 1996; Chan et al, 1998; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Javaid, 1992;

Kirkpatrick et al, 1981; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lyons, 1983; McNeill et al, 1998; Miller et al, 1982; Mucklow and Phelan, 1979; Pagelow, 1980; Patterson, 1994a, 1994b, 1997 (with the limitations discussed above); and Tasker and Golombok, 1995, 1997.

19. Cameron and Cameron, 1996. This study has other problems, however, which we note below.

20. Where feasible, a random sample taken from a population is the superior method of obtaining a comparison group. This is also discussed further in the discussion below.

21. In this case, the two groups are equated by randomly assigning individuals to the experimental group or the control group. This is the best available research design, because any statistically significant differences between the two groups can be plausibly attributed to the experimental manipulation (e.g. Campbell and Stanley, 1966).

22. Stated slightly differently, controlling for extraneous variables reduces the probability that the relationship between the independent variable and the dependent variable is spurious. Spurious relationships are those that are not true causal relationships, but are due to the presence of a third variable. The existence of spurious relationships, caused by extraneous variables, account for the substantial degree of truth in the old adage "correlation is not causation."

23. This tells us nothing, of course, about whether the relevant variables were controlled, or whether the controls were executed properly. These studies are: Bigner and Jacobsen, 1989a, 1989b, 1992; Brewaeys et al, 1997; Chan et al, 1998; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lyons, 1983; Miller et al, 1982; Pagelow, 1980; Riddle and Arguelles, 1989, Tasker and Golombok, 1995, 1997; and Turner et al, 1990. Koepke, Riddle and Turner lack heterosexual control groups.

24. Rossi and Freeman, 1994, p. 311, provide a standard list of extraneous variables that should be considered in social science studies that is discussed subsequently.

25. This also includes subsequent studies by Golombok and Tasker 1996 and Tasker and Golombok, 1995, 1997. As we shall see next, this applies to non-significant as well as significant results.

26. E.g., Hirschi and Selvin, 1973, Rosenberg, 1968, Davis, 1985.

27. Conversely, if one aims to avoid finding real causes, skipping this method will be helpful. We cannot say what the motives are of the researchers whose studies we are examining, but they have used (or failed to use) methods in a way that makes finding real causes very unlikely, and in some cases impossible.

28. The variables are child's age, child's age at separation from his or her father, the child's age when last adult male was living in house, mother's education, mother's feminist activism, whether mother was lesbian or heterosexual, and mother's lesbian political activism.

29. Multiple regression is a very common statistical technique for estimating the effects of a set of independent variables on a dependent variables. It allows examination of the effect of each independent variable while controlling for the effects of all the others.

30. p. 171. Also included were variables closely related to another variable (p. 171). For example, lesbians are more likely to be feminist and lesbian activists than are heterosexuals. This problem of multicollinearity, of which Green et al. are unaware, substantially complicates the interpretation of their results.

31. A measure derived from the multiple regression equation, which indicates how well the dependent variable is predicted by the full set of independent variables included in the equation.

32. See e.g., Chan et al, 1998, p. 452.

33. Those few findings that are reported are themselves quite interesting. They report a very large R-squared value of 0.21 (which is equal to a correlation of 0.46) in a regression equation predicting that the longer boys lived without a man in the house the more likely they would be to mention a women as a person they would like to be when grown (pp. 178-9). This statistically significant result ($p<.05$ level) is strong enough to count as what psychometrician Jacob Cohen (1988) calls a large effect. This concept is discussed below.

34. There is no a priori list of potential suppressors. Suppressors are a function of the research problem under investigation. We select these two because they are relevant to the studies we discuss.

35. Golombok et al, 1983, p. 556.

36. See below in the discussion of individual studies and in note 83.

37. Other statistical problems with Chan et al's study are discussed in the summaries of individual studies.

38. This often uses multiple regression mentioned above, or its close relative, the analysis of variance.

39. Another method of controlling for extraneous variables is test factor standardization, a form of reweighting which is used in demographic analysis (e.g., Davis, 1985) but not in any studies discussed here. We discuss sampling and statistical tests in greater detail in Chapters 4 and 5.

40. E.g., Campbell and Cook, 1979.

41. Experiments often have the problem of extrapolating their findings to the real world.

42. Bigner & Jacobson (1989a), Bigner and Jacobson (1989b) and Green, et al (1986) use some form of pair matching. The balance (there is some overlap) use group matching. See the list below at note 51.

43. We discuss group matching at length on pp.16-24. The reliance on group matching increases the propensity of finding non-significant results. This group-level matching control for extraneous effects is extremely imprecise and an inadequate substitute for statistically controlling for extraneous variables.

44. Id., p. 305.

45. Id., p. 167.

46. We infer this from their subsequent use of McNemar's chi-square test, which is used for paired cases. Green et al, p. 170.

47. Id., p. 169.

48. Rossi and Freeman (1994), p. 305.

49. See Bell and Weinberg, pp. 29-40.

50. p. 33; for the actual numbers in the recruitment pools, see Table 2-1, Appendix C.

51. Categories included, among others: white heterosexual male, with high school or less, and 25 years old or less; white heterosexual male, with high school or less, and 26 to 35 years old; white heterosexual male, with high school or less, and 36 to 45 years old; white heterosexual male, with high school or less, and 46 years old or more; black heterosexual male, with high school or less, and 25 years old or less; black heterosexual male, with high school or less, and 26 to 35 years old, and so on. Id.

52. Some percentages in the table may not add up to 100 percent because of rounding.

53. Despite their quota sampling regarding the comparison groups, Bell et al still

relied on subsequent statistical analyses to further control for extraneous variable effects.

54. These studies are: Bigner and Jacobsen, 1992; Brewaeys et al, 1997; Chan et al, 1998; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lyons, 1983; Miller et al, 1982; Pagelow, 1980; Riddle and Arguelles, 1989; Tasker and Golombok, 1995, 1997; Turner et al, 1990.

55. The investigators report no significant differences between the study and control groups on mother's age, child's age, and the number of children, but found significant differences in education levels and gender distribution of children. Brewaeys et al chose to report data separately for girls and boys rather than also control for the latter statistically. They did not include education in subsequent analyses when controlling for child's gender. Conversely, they did not include child's gender when controlling for education level of the parent. Both gender of child and education should have been entered into the analysis simultaneously as statistical controls, although this increases the probability of finding non-significant results as a function of sub-samples that are extremely small.

56. p. 107.

57. See Green et al, 1986 p. 170 for the correct statistical procedure if pair-matching is used.

58. p. 107.

59 Percentages do not add up to 100 percent due to rounding. 60. 1994, p. 303.

61. 1981, p. 134.

62. Bigner and Jacobsen, 1992; Flaks et al, 1995; Javaid, 1992; Kirkpatrick et al, 1981; Lewin and Lyons, 1982; Lyons, 1983; Miller et al, 1982; Pagelow, 1980.

63. Brewaeys et al, 1997; Chan et al, 1998; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Koepke et al, 1992; Kweskin and Cook, 1982; Riddle and Arguelles, 1989; Tasker and Golombok, 1995, 1997; and Turner et al, 1990.

64. Brewaeys et al, 1997; Chan et al, 1998; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Harris and Turner,

1985; Hoeffer, 1981; Huggins, 1989; Koepke et al, 1992; Kweskin and Cook, 1982; Riddle and Arguelles, 1989, Tasker and Golombok, 1995, 1997; and Turner et al, 1990.

65. More generally, it is a useful omnibus checklist of extraneous variables.

66. Rossi and Freeman (1994), p. 311

67. These studies are: Brewaeys et al, 1997; Chan et al, 1998; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lyons, 1983; Pagelow, 1980; Riddle and Arguelles, 1989; Tasker and Golombok, 1995, 1997; and Turner et al, 1990.

68. We discuss the probabilities of obtaining non-significant results in Ch. 5 (the logic of statistical testing).

69. The following studies report the educational level of the parents: Bigner and Jacobsen, 1989a, 1989b; Brewaeys et al, 1997; Chan et al, 1998; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Hoeffer, 1981; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Miller et al, 1982; Riddle and Arguelles, 1989, Tasker and Golombok, 1995, 1997; and Turner et al, 1990.

70. The studies are Chan et al, 1998; Golombok and Tasker, 1996; Golombok et al, 1983; Miller et al, 1982; and Tasker and Golombok, 1995; 1997. Hoeffer (1981) and Kweskin and Cook (1982) compare lesbian and heterosexual mothers with respect to education level but find no statistically significant differences. Turner et al (1990) and Koepke et al (1992) compare gay and lesbian parents but find no statistically significant differences regarding education. Flaks et al (1995), Javaid (1992), Kirkpatrick (1981), and Lewin and Lyons (1982) only block match for education.

71. See Golombok et al, 1983, Golombok and Tasker, 1996, Tasker and Golombok, 1995, 1997 for versions of the study.

72. See Golombok et al, 1983.

73. Golombok and Tasker; 1996; and Tasker and Golombok, 1995, 1997.

74. See the earlier footnote for a discussion of the findings on p. 107.

75. Standard measures of occupational prestige include O. D. Duncan's Socio-economic Index, Siegel's (NORC) Prestige Scores, and the Nam-Powers (U.S. Census) Score. See pp. 327-365 in Delbert C. Miller, Handbook of Research

Design and Social Measurement (1991) for a discussion of nine measures of socioeconomic status involving occupation.

76. These studies are: Bigner and Jacobsen, 1989a, 1989b; Chan et al, 1998; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Hoeffer, 1981; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Riddle and Arguelles, 1989, Tasker and Golombok, 1995, 1997; Turner et al, 1990.

77. E.g., 1994a.

78. p. 250.

79. Statistical Abstract, 1996, p. 405.

80. Patterson, 1997, p. 269.

81. These studies that mix lesbians living alone with lesbians living with partners versus heterosexual mothers living alone are: Golombok and Tasker, 1996, Golombok et al 1983; Green, 1982; Green et al., 1986; Harris, 1985; Javaid, 1992; Kirkpatrick, 1981; Kweskin, 1981; McNeill et al, 1998; Patterson, 1994a, 1994b, 1997; Tasker and Golombok, 1995, 1997.

82. Bigner and Jacobsen, 1992; Brewaeys et al, 1997; Chan et al, 1998; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Hoeffer, 1981; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lyons, 1983; Pagelow, 1980; Riddle and Arguelles, 1989, Tasker and Golombok, 1995, 1997; Turner et al, 1990.

83. The investigators fail to compare adult sons of homosexual fathers to adult sons of heterosexual fathers.

84. This study is the only one we looked at that used longitudinal design.

85. Tasker and Golombok, 1995, pp. 210-211. They also find that 36 percent of the adult children of lesbians but only 25 percent of the adult children of single heterosexual mothers experienced attraction to someone of the same gender. While this was not statistically significant, it is quite suggestive in light of the other findings and the extremely small sample sizes (Tasker and Golombok, 1995, pp. 210-211). In fact, a careful examination of Tasker and Golombok's (1997) data in Table 6.1, p. 107, suggests a very strong relationship between the sexual orientation of the mother and the child. Two of the four results are correctly presented as statistically significant, one is incorrectly presented as statistically insignificant, and the fourth is statistically significant (25 percent difference between the groups) if a one sided test is used, despite the fact that the number of cases in each test is only 45.

Chapter 3
Does it Measure Up?
Bias, Reliability and Validity

In Chapters 1 and 2 we have seen that a well-formulated hypothesis is critical, and that the researcher must use certain methods to control for unrelated effects that may skew a study's results. We have also seen that same-sex parenting studies are severely flawed in these regards.[1]

But what kind of measurements does a study use? This too is crucial. If one assumes that when a person lies, that person will perspire, then one might measure the sweat. But if the amount of sweat is a silly thing to gauge to uncover lying, then it's a poor measure. Take for example, astrology. Astrology claims to measure personality traits by aligning the stars and planets. Is the celestial alignment of stars and planets a silly measure of personality structure?

Regarding a study's measures, there are three questions that need to answered:

1) Is the measure self-constructed?
2) Is it reliable? and finally,
3) Is it valid?

We will provide a brief description of each topic, and look at how the parenting studies fare. The studies offer us no confidence in their results.

What Measures Are and Why They Matter

If variables are properly measured, we can say with greater confidence that the differences between the scores of two respondents are likely due to real differences. If the variables are wrongly measured,

Notes for this section begin on Page 67

we have a false impression based on errors of measurement. Since no measure is perfect, however, there will always be some error in the results. Sometimes the errors are based on the state of the respondent. The respondent may be sick, or tired, or inattentive in some other way. Or, the respondent may have previously answered a similar kind of survey and not be paying much attention. Of course, the respondent may not like the interviewer and give less-than-cooperative answers. The last feature is very important regarding homosexual parenting studies. The respondent may give what he or she perceives as the socially desirable answer. Respondents have been known to conceal their true feelings, actions and attitudes from the interviewer when they are undesirable. Respondents have also been known to exaggerate or even invent their actions and attitudes when respondents believe it may put them in a favorable light.[2] Even if the respondent's state is fine, the essential problem for evaluating the quality of measures is that one cannot really tell, given a respondent's answer, what proportion of that answer is true. Faced with the social desirability problem, checking for measurement errors, particularly in controversial areas, should be done, but should be done indirectly. To do so, an evaluator looks for indicators: Has the measure been used before? Does the measure work again and again? Does it really measure the thing it claims to measure?

In particular, one should be on the lookout for measurement errors that slant the responses consistently in one direction. After all, the researcher is supposed to eliminate, as much as possible, measurement errors that produce a systematic bias. Otherwise the measuring instruments will themselves increase the invalidity of the results.

Are the Measures Self-Constructed?

The key to accurate measurement is a scientific consensus that a measure works. This means that it has been subject to repeated use, and that the use has confirmed or revised the measure in such a way that it can be relied upon with confidence. This is even more important when the goal of a study is to impact public policy, not just remain as an interesting piece of academic research.

Therefore, while it is possible to construct one's own measures to

use in a study, such a strategy should be a last resort.[3] Self-constructed measures are generally a bad idea. At the very least, they offer no reason to trust that anything has yet been accurately and truly measured. The burden of proof is on researchers who design their own measures. The more self-constructed or unexplained the measure, the more dubious the study evaluator should be. For these reasons, we consider all self-constructed measures to be inadequate without direct, extensive evidence showing otherwise.

Looking at the 49 studies, we find that 23 studies appear to have created some of the measures used in their studies.[4] We say, "appear to," because they do not say. On this question, the studies are entirely silent. The watchwords here should be "Trust and Verify." Without further information, these studies should not be trusted.

Are the Measures Reliable?

Reliability is the extent to which repeated applications of the measure result in the same outcomes. No measuring instrument is perfectly reliable, but some measures are better than others. Established measures of physical variables such as height, weight, and body temperature are less prone to reliability errors than those in the social sciences.[5]

For example, if you use a ruler to measure a person's height at four different times during a week, the ruler should give you the same number of inches. In contrast, administering the Scholastic Aptitude Test (SAT) to the same subject four times will produce more varied results. The SAT is therefore a less reliable test, compared to using a ruler. On the other hand, if one compares SAT results to an individual's answers to a survey question, such as, "Should there be less regulation of the economy?" asked of the same subject four times. An individual's SAT score is far more reliable than an individual's responses in a survey. Reliability is a matter of degree.

This means, for better or for worse, there is no standard level of acceptability when testing for reliability, but there are three basic methods of assessing the reliability of a measuring instrument: test-retest, parallel forms, and split halves.[6] Of the three methods, experts

agree that the test-retest index is the best measure of reliability.[7] In other words, pick a measure already established in the area and carefully report upon and study its reliability. Well-known evaluation researchers and sociologists Peter Rossi and Howard Freeman's rule of thumb is that unless a measuring instrument yields the same results 75 to 80 percent of the time, it is not useful.[8]

Context is important. It is one thing to pioneer an exploratory study. It is another thing to set out to influence policymakers, including courts. When the goal is to affect the larger society with the findings obtained, researchers should be able to show that the measure has been widely used, in many studies, for a long period of time with good results.

This is not the case with most of the studies under examination here. Looking at the 49 studies, we find the following:

Twenty-three studies do not refer at all to tests for reliability[9] Five studies reported on the reliability of their measures[10]
Fifteen studies referenced measures previously used in other studies[11]
Six studies reported checks for reliability[12]

This is not to say that the measures are unreliable. We just cannot say that the measures are reliable. If we cannot say they are reliable, we cannot recommend public policies.

One good example of a carefully tested set of measures is in **McNeill et al (1998).** The investigators had each respondent complete four inventories to measure family relations and parental attitudes. These measures were 1) the Index of Family Relations,[13] 2) the Index of Parental Attitudes,[14] 3) the Family Awareness Scale,[15] and 4) the Dyadic Adjustment Scale.[16] Each measure was developed and tested in many previous studies over a long period of time. **McNeill** et al also report "test-retest reliability of .87 or higher."[17] That is, when an individual took the same test a second time after a reasonable passage of time, 87 percent of the answers were the same.

One final point about reliability has to do with the effect of unreliability on studies that seek to affirm the null hypothesis. Unreliable measures tend, as a rule, to lower the magnitude of correla-

tions and other statistics; this tends to make it easier to fail to reject the null hypothesis and bias results in favor of finding "no differences between homosexual and heterosexual parents."[18]

Are the Measures Valid?

Validity is the other major concern regarding measurement. Being able to replicate a measurement is essential but not sufficient. The measurement also needs to actually measure what it purports to measure. For example, do readings on your oven thermometer truly measure the temperature of your oven? Do readings of PH levels from a soil testing kit really measure the degree of acidity or alkalinity in your lawn? Does an individual's astrological birth sign really measure personality traits?

There are two kinds of validity: construct validity and empirical validity.[19]

Construct validity evaluates whether the measure (the reading on the oven thermometer) is a valid indicator of the underlying construct (the temperature).

Empirical validity (also called predictive validity) evaluates to what degree a measure correlates empirically with other independent measures of the same construct.

Here are two examples of how validity might be tested. The first concerns tests of mathematical ability. Standardized scores on Test X should be the same as those on other measures of math ability. If scores on Test X correlate better with a measure that is seemingly unrelated to math ability—such as church attendance—Text X is an invalid measure of mathematical ability.

Another example is the SAT. These tests are supposed to measure the theoretical construct, "academic ability." To a lesser extent, so do high school grade-point-averages. SATs are highly correlated with GPAs. The SATs in turn are highly correlated with first-year college grades. That these two measures are highly correlated with each other increases the validity of the SAT as a measure of academic ability. In contrast, suppose we used another measure, such as number of high school extra-curricular activities. This measure might

have no validity regarding academic ability. As such, we would not expect it to predict college performance as measured by a college GPA. In other words, the SAT would have high predictive validity, but participation in high school extra-curricular activities would have little or no such validity.

How did the 49 studies fare regarding measurement validity? Twenty studies provide no references or reports of calculations regarding validity.[20] Twenty-nine studies provided references or carried out calculations regarding validity Of the 29 studies, four reported and referenced the validity of their measures.[21] Three presented the reported validity of their measures.[22] The other 22 merely referenced the validity of their measures.[23]

What Went Wrong and What Can Be Done About It?

Here are the lessons from Step 3: Use Reliable and Valid Measures.

1) Avoid self-constructed measures. No one has a reason to trust them

2) Use reliable measures. Tools must be capable of being confirmed

3) Use valid measures. Tools must measure the object of your study

In a way, it is not a surprise that these studies perform poorly on the questions of measurement. Consider that assessing the reliability and validity of various IQ tests as measures of intelligence have been going on for more than 50 years, for large national samples all around the world (e.g., for different countries, for different age groups) yet questions are still sometimes raised concerning the validity and reliability of IQ scores. Because the study of homosexual parents and their children is so new and untested, claims of reliable and valid measurement are dubious without many careful and repeated studies of such measures. These are nonexistent. One should consider most measures as applied to homosexual parents and their children to be only exploratory. They should form no basis for public policy recommendations.

Notes to Chapter 3

1. The summaries of our comments can be found on pp. 21 (for Chapter 1) and 52-53 (for Chapter 2).

2. When the interviewer is the researcher also, which is the case in a number of these studies, the potential problem of response contamination is very great. In their path-breaking scientific study of sexual behavior in America, Laumann et al. (1995) used a respondent self-administered form to ask a number of extremely sensitive questions, so that the interviewers could not influence respondent's answers (1995, p. 60).

3. Miller goes so far as to state that creating one's own measure should be the action of last resort in any social science research. Miller, 1991, p. 580.

4. This is evidenced by the lack of references regarding their measures. The studies are: Barret and Robinson, 1990; Bozett, 1980; Cameron and Cameron, 1996; Gartrell et al, 1996; Hare, 1994; Harris and Turner, 1985; Javaid, 1992; Lewin and Lyons, 1982; Lewis, 1980; Lott-Whitehead and Tully, 1992; Lyons, 1983; McCandlish, 1987; Miller, 1979; O'Connell, 1993; Pagelow, 1980; Pennington, 1987; Rand et al, 1982; Riddle and Arguelles, 1989; Ross, 1988; Turner and Harris, 1990; Weeks et al, 1975; West and Turner, 1995; and Wyers, 1987.

5. Rossi and Freeman, p. 230; Nachmias and Nachmias, p. 170-171.

6. We will spare the reader the technical aspects of assessing reliability, since these are easily found in Nachmias and Nachmias, 170-175.

7. Miller, *Handbook of Research Design and Social Measurement*, 1991, p. 580.

8. Rossi and Freeman, p. 232.

9. This is evidenced by the lack of references regarding their measures. The studies are: Barret and Robinson, 1990; Bozett, 1980; Cameron and Cameron, 1996; Gartrell et al, 1996; Hare, 1994; Harris and Turner, 1985; Javaid, 1992; Lewin and Lyons, 1982; Lewis, 1980; Lott-Whitehead and Tully, 1992; Lyons, 1983; McCandlish, 1987; Miller, 1979; O'Connell, 1993; Pagelow, 1980; Pennington, 1987; Rand et al, 1982; Riddle and Arguelles, 1989; Ross, 1988; Turner and Harris, 1990; Weeks et al, 1975; West and Turner, 1995; and Wyers, 1987.

10. Chan et al, 1997; Kirkpatrick et al, 1981; McNeill et al, 1998; Tasker and Golombok, 1995, 1997. Referencing measures is clearly the best approach in a relatively new field.

11. The studies are: Bailey et al, 1995; Bigner and Jacobsen, 1989a, 1989b, 1992; Brewaeys et al, 1997; Flaks et al, 1995; Green, 1978; 1982; Green et al, 1986; Golombok and Tasker, 1996; Golombok et al, 1983; Kweskin and Cook, 1982; and Patterson, 1994a, 1994b, 1997. This makes them seem somewhat more reliable than those lacking any reported or referenced reliability

12. Crosbie-Burnett and Helmbrechty, 1993; Hoeffer, 1981; Huggins, 1989; Koepke and Moran, 1992; Miller et al, 1982; and Mucklow and Phelan, 1979.

13. In Hudson, 1992, *The Walmyr Assessment Scales, Scoring Manual.*

14. In Hudson, 1992, *The Walmyr Assessment Scales, Scoring Manual.*

15. In Green Kolevzon and Vosler, 1985.

16. In Spanier, 1976.

17. McNeill et al, 1988, p. 60.

18 For example, Cohen provides a brief discussion of this point (1987, p. 537).

19 There are many different labels and differentiated types of validity cited in the testing literature. Differentiating between construct and empirical validity is the minimal necessary distinction needed here.

20. Barret and Robinson, 1990; Bozett, 1980; Cameron and Cameron, 1996; Gartrell et al, 1996; Harris and Turner, 1985; Lewin and Lyons, 1982; Lewis, 1980; Lott-Whitehead and Tully, 1992; Lyons, 1983; McCandlish, 1987; O'Connell, 1993; Pagelow, 1980; Pennington, 1987; Rand et al, 1982; Riddle and Arguelles, 1989; Ross, 1988; Turner et al, 1990; Weeks et al, 1975; West and Turner, 1995; and Wyers, 1987.

21. Tasker and Golombok, 1995, 1997; Brewaeys et al, 1997; Kirkpatrick et al, 1981.

22. Huggins, 1989; Koepke, 1992; Miller et al, 1982.

23. Bailey et al, 1995; Bigner et al, 1989a, 1989b, 1992; Chan et al, 1997; Crosbie-Burnett and Helmbrechty, 1993; Flaks et al, 1995; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1978; 1982; Green et al, 1986; Hoeffer, 1981; Kweskin and Cook, 1982; and Patterson, 1994a; 1994b; 1997.

Chapter 4
Sampling:
It all Depends On Who You Ask

We have examined hypotheses, controls, and measurements. The next key issue is sampling. Sampling is a simple concept—choosing cases to include in your study. The question is: "Have you used a method from which you can reasonably generalize?"

What Sampling Is and Why It Matters

Sampling is the systematic means by which cases are selected for inclusion in a study. There are two basic types of samples: probability and non-probability samples. The distinction is critical because one cannot generalize from a non-probability sample.

Probability versus non-probability sampling is a fundamental distinction in research. The most important fact about the 49 studies we evaluated is that 48 of them used non-probability samples.[1] If we exclude the four clinical case studies with five or fewer subjects, we have 44 deeply flawed quantitative studies using non-probability samples.[2]

One cannot generalize from these studies. They may give us interesting leads, and suggest possible insights, but nothing reliable can be inferred from them outside the individuals studied.

In this chapter we explain the difference between probability and non-probability sampling, types of each, and give examples of faulty sampling from the parenting studies.

Notes for this section begin on Page 78

Probability Sampling: The Key to Valid Research

In a probability sample, each unit of the population studied has a known probability of being included in the sample. These studies use randomization methods to select the respondents for a study.[3]

There are three types of probability samples: the simple random sample, the stratified random sample, and the cluster sample. In the simple random sample, each unit in the population has an equal chance of being included in the study. In the stratified random sample, the population is divided into strata, and each stratum must be represented in known proportions within the sample. Independent samples are selected by random procedure within each stratum, and each unit must appear in one and only one strata.

The third type of probability sampling is the cluster sample. The population is divided into homogenous, geographically defined groups. A sample of these groups is next drawn by random procedure, and elements within each of these samples are in turn selected by random procedure. For example, cluster-sampling households in a major city might first involve separating the census tract into homogeneous geographical sections. The investigator would then randomly select a sample of these geographical sections (first-stage cluster sampling), then randomly select city blocks within the sample of geographical sections (second-stage cluster sampling), and finally select randomly the households within the sample blocks (third-stage cluster sampling).[4]

Modern sample surveys typically rely on both clustering and stratification to obtain probability samples. These complex designs require calculation of sampling weights and statistical software for statistical estimation and testing.[5]

Since the researcher can legitimately draw generalizations about the larger population from probability samples, the size of the sample is important. Larger probability samples yield more accurate results.

Non-Probability Sampling: It Doesn't Do the Job

As mentioned above, 44 studies used forms of non-probability sampling.[6] The researchers on homosexuals and homosexual parents

repeatedly blundered because they use non-probability samples to wrongly make population estimates.

The First General Problem with Non-Probability Samples. The first general problem with non-probability surveys has already been mentioned. One cannot generalize from a non-probability sample, although **Patterson** and others repeatedly try to do so. In a review of the literature, **Patterson** claims that the estimated American homosexual population is 10 percent of the adult population based on information gleaned from the dated Kinsey report.[7] Similarly **Patterson** states that the percentage of gays and lesbians that are parents as 10 percent and 20 percent respectively, based on "large-scale survey studies" such as those by Bell and Weinberg, (1978) and Saghir and Robins, (1973).[8] In the same review, **Patterson (1992)** also cites population estimates for the number of lesbians having borne children after coming out.[9]

Patterson claims that the figures under-represent the actual numbers.[10] However, she has no scientific basis for this claim. There is no way one can make population estimates on volunteer samples of anything. It has nothing to do with discrimination or stigmatization of homosexuals. It has everything to do with the basic distinction between a probability and non-probability sample.

The Second General Problem with Non-Probability Samples. The size of the sample is irrelevant for making estimates, because population estimates based on non-probability samples are not scientific, despite appearing to be so.[11] A large non-probability sample does not give you better population estimate than a small non-probability sample. The size of the sample is only relevant for probability samples, where larger samples allow greater precision of population estimates.

The Third General Problem with Non-Probability Samples. The third general problem with non-probability samples is their tendency toward bias. While probability samples systematically eliminate this problem of bias through random selection procedures, non-probability samples do not. Have the investigators overlooked any obvious biases in their process of sample selection? Where did the investigators get their participants? How were they found? Were there any predetermined restrictions on eligibility? How does this af-

fect the final sample used for data collection and analysis, and how might it affect the results?

For example, it is a sociological "truism" that college students become more liberal the longer they stay in college. The original study was done on a sample of college students at Bennington College. Even if the study managed to use the complete population of the college, the results would nevertheless contain biases associated with the selection of the individual school itself (which may be considerable). Magazine volunteer polls are common forms of non-probability sampling. In a typical magazine poll, the magazine reports the results of those who voluntarily respond to a questionnaire in a magazine. The respondents almost certainly differ in systematic ways from the non-volunteers, first by showing a strong interest in the subject of the questionnaire. There are also biases inherent in those reading the magazine itself, as compared to the general population. No matter how large the number of respondents, findings from such magazine, television, or Internet surveys cannot be generalized to the larger population.

The study of human sexual behavior is especially plagued by over-reliance on non-probability sampling. The famous estimates of the proportion of homosexuals the U.S. population in the Kinsey reports, for example, are based solely on non-probability samples, which rely substantially upon volunteers and on heavy sampling of highly unrepresentative locales such as prisons.[12]

There are times, however, when the researcher must rely on non-probability sampling. Social scientists sometimes do so if probability sampling is too expensive or difficult. We will now turn to four general types.

Specific Types of Non-Probability Samples.

There is no standardized classification of types of non-probability sampling. The three main types of non-probability sampling are convenience sampling, purposive sampling, quota sampling, and snowball sampling.[13]

Convenience sampling is just what it sounds like. One selects whoever is available, such as students in an introductory psychology class.

Purposive sampling involves selecting cases that the investigator believes are representative of the larger population. An example would be election forecasting. A small number of precincts in each state are selected, based on the extent to which those precincts mirror the overall state election returns for the previous election. Election forecasting rests on the assumption that these precincts still will mirror the state election returns.

With **quota sampling**, the investigator tries to select a sample as similar as possible to the sampling population. An investigator may seek an equal number of men and women, if the investigator thinks the population from which the sample is drawn will have an equal number of men and women. Quota sampling requires the investigator to use his or her judgment to identify all the important features that might affect the sampling.

Snowball sampling is another method of selecting cases that is not strictly speaking a form of sampling. The snowball method is sometimes used to study rare populations. It presumes that a network exists among members of the particular rare population. The investigator depends on one member of the network to identify others in the same network until the sufficient number of cases is reached.

As we said before, the researcher sometimes is forced to use non-probability samples. Because participants in any non-probability sample are not randomly selected, identifying where the participants come from is another critical component for evaluating the design of a study. Have the investigators overlooked any obvious biases in their process of sample selection? Where did the investigators get their participants? How were they found? Were there any predetermined restrictions on eligibility? How does this affect the final sample used for data collection and analysis, and how might it affect the results?

Non-Probability Studies. Potentially Biased Participants

Of the 44 quantitative studies using a non-probability sampling design, in five studies, the researchers established the initial list of potential subjects. In three of these, subjects were initially recruited from the investigators' own list of clinical cases.[14] In the other two,

the subjects were drawn from lists of sperm bank patients.[15] The remaining 39 studies relied on some form of self-selecting volunteers for their final pool of subjects.[16] The inherent problem with relying on self-selected volunteers is obvious: when either or both the study and comparison groups know the purpose of the study and have a large stake in its substantive outcome, one almost inevitably introduces very serious sample selection biases into a study. The participants have every incentive to paint themselves in the best possible light. Laumann et al. (1994), authors of a groundbreaking, scientifically rigorous study of sexual behavior in the United States, reviewed non-probability based studies of sexual behavior and found that generalizations "are very likely to be strongly biased in an upward direction (i.e., overestimating the incidence of certain behaviors) because the samples are highly self-selected on the very variables of interest."[17] It is Laumann et al.'s view that the public's perception of sexual behavior in the United States is formed by highly visible non-probability samples, which overestimate substantially the true frequency and types of sexual behavior.[18] Another well-known quantitative social scientist has come to a similar conclusion.[19]

These findings have serious implications regarding the studies under review. All the studies on homosexual parents and their children save one used some form of self-selecting non-probability sample. This means they cannot answer the question of whether there is "no difference" between homosexual versus heterosexual parents and their children in the larger population. The fact of self-selection bias and the problem of overestimation are inherent in these studies.

We find that investigators drew both heterosexual and homosexual parents from a socially active pool. This introduces biases, because the general public is inactive. Belonging to an organization is more active than subscribing to a newsletter; being a leader or regular participant is even more so. Ironically, being a study volunteer is associated with a host of demographic traits (higher levels of education and greater occupational prestige, for example) that makes one different from the general public.

Where did these studies draw subjects? Twenty of the studies relied on some form of snowball sampling, all in combination with

other methods. Study participants were asked to name or contact others that might fit the sample criteria and be interested in participating.[20]

Homosexual Participants. Publications and newsletters were also a major vehicle for recruiting homosexuals but not heterosexuals. Seventeen studies relied on gay-lesbian or feminist publications for the homosexual parent sample.[21] In contrast, one heterosexual sample was obtained from an advertisement in a feminist newsletter, which is likely to minimize rather than maximize differences between homosexual and heterosexual respondents.[22]

Ten studies also recruited homosexual parents from gay-lesbian parent support groups,[23] while 11 recruited gay-lesbian parents through gay and lesbian organizations.[24] Three studies relied also on feminist groups for lesbian mothers,[25] but none did so for the heterosexual samples. Three studies relied on day care centers and day care newsletters to recruit lesbian mothers.[26]

Since hard-to-find populations are expensive and time consuming to properly survey using some form of probability sampling, it is no wonder that investigators choose one or another variety of non-probability sampling. Unfortunately, drawing from a more activist pool of parents raises the problem of biased results. In particular, the parents in gay-lesbian parent support groups have every incentive to give the socially desirable answer to any questions about their children as a way of justifying their choices and lifestyles. These highly educated respondents are almost inevitably going to be biased toward giving the socially desirable answer that would put themselves and their children in the most positive light.

Heterosexual Participants. Three studies used random samples to obtain a heterosexual comparison group.[27] But the most common means of recruiting heterosexual mothers was the single parent support group or newsletter, used by 12 studies (or 80 percent) of those studies that used a heterosexual comparison group.[28] Two studies relied on day care centers and day care newsletters to recruit heterosexual mothers,[29] while two others drew heterosexual mothers from the local PTA.[30] These studies did not attempt to correct for

biases, even after acknowledging their existence.[31]

With the exceptions of **Cameron and Cameron (1996)** and **Bigner and Jacobsen (1989a and 1989b),** these studies draw heavily from sources that seem extremely unrepresentative of single parents. The most common source of finding heterosexual single mothers was through single parent organizations and newsletters (80 percent of the studies using a comparison group). This seems like an unusual source for single parents. While we do not know what percentage of single mothers participate in single parent organizations, we do know that 46 percent of single mothers participate moderately or often in school activities such as the PTA, school volunteer work, school committee work, or school event.[32] Single mothers from local PTAs may be a more representative source of single-mother respondents. In fact, the typical single mother, according to calculations from census data, is African-American or Latino, 33 years of age, with less than one year of college education.[33] Since the typical single mother respondent in these studies is Caucasian with some college education, the comparison sub-samples are already grossly non-representative of the general single-mother population.

Similarly, heterosexual participants drawn from institutional daycare centers and newsletters are also not representative of heterosexual mothers. The majority of children under age six are not in institutional daycare centers. Most children under age six during the workday are either cared for by their parents (40 percent), relatives (21 percent), or home-based daycare (18 percent). Only 31 percent are in day care centers, nursery schools, *Head Start*, or some form of institutional pre-Kindergarten program.[34]

This does not stop some from trying to salvage these studies. One review by **Patterson and Redding** goes so far as to totally invert the scientific standard regarding sampling and populations.[35] It argues that criticizing the non-probability sampling studies as non-representative is unjustified. Why? "[B]ecause researchers do not know the actual composition of lesbian mothers, gay fathers, or their children (many of whom choose to remain hidden), and hence cannot possibly evaluate the degree to which particular samples do or do not represent the population. At present, there is no more rea-

son to argue that samples do not represent the population of lesbian mothers, gay fathers, and their children than there is to argue that they do represent it (italics added)."[36]

Patterson and Redding's argument turns standard principles of sampling and statistics on their head. The criticism that these studies are non-representative is not based on the assumption that samples do or do not represent the larger population. Non-probability samples are not random by definition and therefore cannot be used to generalize to the larger population because there is no way of doing so. To assume that a sample is representative unless shown otherwise is simply absurd. Representativeness is a matter of science, not of advocacy.

What Went Wrong and What Can Be Done About It

Here are the lessons from Step 4: Use Valid Samples.

1) Use probability samples. There is no substitute. Only these offer any basis for scientific generalization to larger, representative populations.

2) Ignore studies based on non-probability samples when making policy decisions. They offer little basis for scientific generalization. Therefore they have no valid implications for general questions of public policy.

3) Especially ignore studies where participants recruit other participants. These are so subject to bias, that the limited results cannot be trusted. Patterson and Redding argue that, "In the long run, it is not the results obtained from any one specific sample but the accumulation of findings from many different samples that will be most meaningful."[37] This is a perfect illustration of the problem with these studies. In the long run, non-probability samples will yield non-generalizable and biased results-just as they have in the short run. Nothing plus nothing equals nothing. Meaningful and correct generalization to the population has nothing to do with the number of studies done on the subject, nor on the sizes

of their samples, nor on having "many different samples." It has to do with correct, that is to say probability, sampling. Whether researchers have the time and resources to properly survey populations such as homosexual parents is a question that only the researchers can answer. But proper research can be done, and there is no excuse in this age of interdisciplinary training and diffusion of elementary knowledge of the principles of modern scientific statistical methods for the low quality and inflated claims made by the studies we evaluate here. Laumann et al (1994) stands as an example of what can and should be done in for the study of the impact of homosexual parents on their children. They successfully used probability methods to obtain their study subjects, and then successfully minimized both bias on the part of the investigators (because the respondents who are sampled are not known individually to the researchers) and in the selection of subjects (because the sample is randomly selected, it does not consist of volunteers, and because they achieved a high response rate in obtaining information from the original sample).[38]

Any study seeking to survey a special population, especially for the purpose of influencing public policy decisions, ought to have available to its research team a professional sampling statistician. Brought in during the early stages of the planning process, the statistician would assist in creating a proper sampling design, and hopefully would prevent the repetition of the flaws present in the studies we have investigated, and might even prevent the inflated claims for their findings put out by their authors.

Notes to Chapter 4

1 The only exception is Cameron and Cameron (1996). This study, while using randomization, did not select a national sample and because they did not use stratification, obtained only 17 adults who reported having homosexual parents (Cameron and Cameron, 1996. p.764) Other problems are discussed in Chapter 5.

2. They are: Bailey et al, 1995; Bigner and Jacobsen, 1989a, 1989b, 1992;

Bozett, 1980; Brewaeys et al, 1997; Chan et al, 1998; Crosbie-Burnett and Helmbrechty, 1993; Flaks et al, 1995; Gartrell et al, 1996; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1978, 1982; Green et al, 1986; Hare, 1994; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lewis, 1980; Lott-Whitehead and Tully, 1992; Lyons, 1983; Miller, 1979; McNeill et al, 1998; Miller et al, 1982; Mucklow and Phelan, 1979; O'Connell, 1993; Pagelow, 1980; Patterson, 1994a, 1994b, 1997; Pennington, 1987; Rand et al, 1982; Riddle and Arguelles, 1989; Tasker and Golombok, 1995, 1997; Turner et al, 1990; West and Turner, 1995; and Wyers, 1987.

3. Nachmias and Nachmias, 1996, pp. 185-195 provide a textbook discussion of types of probability sampling, and examples of drawing a nation-wide sample.

4. See also Kish, 1965; Kalton, 1987; and Schaeffer et al., 1996 for detailed and technical discussions of modern techniques of probability sampling.

5. Kish, 1965; Kalton, 1987; and Schaeffer et al., 1996 op. cit for detailed discussion of the sophisticated sampling methods developed by modern sampling statisticians.

6. Five used no form of sampling at all.

7. 1992, p. 1026. Patterson obtained the figures from Kinsey, Pomeroy, and Martin, 1948.

8. Ibid.

9. Ibid.

10. Ibid. Estimates of lesbian mothers are provided by Falk (1989), Gottman (1990), Hoeffer (1981), and Pennington (1987), whil Bozette (1987), Gottman (1990), and Miller (1979) provide estimates of gay fathers. The number of children raised by homosexual parents is estimated by Bozette (1987), Peterson (1984), and Schulenberg (1985).

11. The size of the sample does affect the power of the statistical tests used to detect statistical significance (this is discussed in far more detail in subsequent sections on the logic of statistical testing).

12. Laumann et al. provide an explanation as to how the Kinsey 10 percent figure became accepted as the "right" proportion of homosexuals in the U.S. population in an insightful discussion entitled, "The Myth of 10 Percent and the Kinsey Research" (1994, pp. 287-290).

13. Nachmias and Nachmias, 1996, pp. 184-85.

14. Crosbie-Burnett, and Helmbrechty, 1993; Green, 1978; and Pennington, 1987.

15. Brewaeys et al, 1997; Chan et al, 1998.

16. Bailey et al, 1995; Bigner and Jacobsen, 1989a, 1989b, 1992; Bozett, 1980; Flaks et al, 1995; Gartrell et al, 1996; Golombok and Tasker, 1996; Golombok et al, 1983; Green, 1982; Green et al, 1986; Hare, 1994; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Lewis, 1980; Lott-Whitehead and Tully, 1992; Lyons, 1983; Miller, 1979; McNeill et al, 1998; Miller et al, 1982; Mucklow and Phelan, 1979; O'Connell, 1993; Pagelow, 1980; Patterson, 1994a, 1994b, 1997; Rand et al, 1982; Riddle and Arguelles, 1989; Tasker and Golombok, 1995, 1997; Turner et al, 1990; West and Turner, 1995; and Wyers, 1987.

17. p. 46.

18. 1994, p. 46.

19. Greeley, 1994. Greeley compared two surveys, one a popular non-probability survey, "The Janus Report" the other a nationally recognized, probability-based sample survey, "GSS," that is widely used in the quantitative social sciences. The first set of results was from "The Janus Report" (Janus and Janus, 1993). These were based on a non-probability sample of 8,000 respondents gathered unsystematically from various sources, including patients of sex therapists and their friends and acquaintances. The other set of results were from the General Social Survey (GSS). The GSS is based on a national household-based probability sample of adults. It is conducted by the National Opinion Research Center at the University of Chicago, nearly every year, for the past 20 years, funded by the National Science Foundation. *General Social Surveys, 1972-1996: Cumulative Codebook*, 1996. Greeley compared, among other data, the percentage of persons reporting they had sex at least once a week. Janus estimates were much higher than those in GSS. For men in the youngest age group (between the ages of 18 and 26), 72 percent in the Janus report claimed to have sex at least once a week versus 57 percent in the GSS. For women in the same age group, it was 68 percent in Janus versus 58 percent in the GSS. In another comparison, 83 percent of Janus men versus 56 percent of GSS men and 68 percent of Janus women versus 49 percent of GSS women between 39 and 50 years of age report having sex at least once a week. The overestimates are greatest among

the oldest respondents: 69 percent of Janus men versus 17 percent of GSS men, and 74 percent of Janus women versus 6 percent of GSS women over 65. The Janus estimates are four times that of the GSS estimates for men over 65, and 12 times that of the GSS estimates for women. That is, Janus respondents claim to have sex much more often than those surveyed in the GSS. The overall pattern is one of consistent overestimation by the non-scientific Janus Report.

20. The following studies rely on the snowball technique: Crosbie-Burnett and Helmbrechty, 1993; Flaks et al, 1995; Gartrell et al, 1996; Hare, 1994; Harris and Turner, 1985; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Koepke et al, 1992; Lewin and Lyons, 1982; Lott-Whitehead and Tully, 1992; Miller, 1979; O'Connell, 1993; Patterson, 1997; 1994a; 1994b; Rand et al, 1982; Riddle and Arguelles, 1989; Turner et al, 1990; and West and Turner, 1995.

21. Bailey et al, 1995; Crosbie-Burnett and Helmbrechty, 1993; Flaks et al, 1995; Gartrell et al, 1996; Golombok and Tasker, 1996; Golombok et al, 1983; Green et al, 1986; Hare, 1994; Harris and Turner, 1985; Kirkpatrick et al, 1981; Lewin and Lyons, 1982; Lewis, 1980; O'Connell 1993; Tasker and Golombok, 1995; Tasker and Golombok, 1997; Turner et al, 1990; West and Turner, 1995.

22. Kirkpatrick et al, 1981.

23. Bigner and Jacobsen, 1989a, 1989b, 1992; Crosbie-Burnett and Helmbrechty, 1993; Flaks et al, 1995; Gartrell et al, 1996; Golombok and Tasker, 1996; Golombok et al, 1983; Tasker and Golombok, 1995; 1997.

24. Flaks et al, 1995; Gartrell et al, 1996; Golombok and Tasker, 1996; Golombok et al, 1983; Harris and Turner, 1985; Lott-Whitehead and Tully, 1992; Pagelow, 1980; Tasker and Golombok, 1995, 1997; Turner et al, 1990; Wyers, 1987.

25. Green et al, 1986, Miller, 1982, O'Connell, 1993.

26. Harris and Turner, 1985; Lewin and Lyons, 1992; Turner et al, 1990.

27. Bigner and Jacobsen, 1989a, 1989b; Cameron and Cameron, 1996.

28. These studies are Bigner and Jacobsen, 1992; Golombok and Tasker, 1996; Golombok et al, 1983; Harris and Turner, 1985; Huggins, 1989; Javaid, 1992; Kirkpatrick et al, 1981; Kweskin and Cook, 1982; Lewin and Lyons, 1982; Pagelow, 1980; Tasker and Golombok, 1995, 1997.

29. Harris and Turner, 1985; Lewin and Lyons, 1992.

30. Miller et al, 1982; Mucklow and Phelan, 1979.

31. In contrast, Bell, Weinberg and Hammersmith (1981) attempted to correct for self-selection bias in their non-probability sample of homosexuals in the following way. They believed (correctly) that a homosexual respondent's familiarity with various theories of homosexuality might bias their responses. As a filter question, they asked whether the homosexual respondents had read books or articles, or attended lectures about homosexuality. Responses were subsequently controlled for this *a priori* familiarity. In some cases, it made a significant difference, in others, it did not, and reporting of results was adjusted accordingly. Bell, Weinberg, and Hammersmith, 1981, p. 20.

32. Zill and Nord, 1994, p. 46.

33. Zill and Nord, 1994, p. 14-15.

34. "Regular Child Care Arrangements for Children Under 6 years Old, by Type of Arrangement," 1995, *Statistical Abstract*, 1996, p. 386.

35. Patterson and Redding, 1996, p. 44.

36. Ibid.

37. Ibid.

38. Laumann et al. 1994, "Sampling Procedures and Data Quality," pp. 549-570.

Chapter 5
Just by Chance?
Statistical Testing

We come now to the culmination of the social-scientific process: statistical testing. For a non-social scientist this may sound like a valley rather than a mountaintop. But if hypotheses are properly conceptualized (Chapter 1); if extraneous variables are properly controlled (Chapter 2); if concepts are properly measured (Chapter 3); and if populations are properly defined and samples properly drawn (Chapter 4); then we are ready for the process of statistical hypothesis testing. It should be quite straightforward.

There is a lack of scientific rigor in the same-sex parenting studies in this step. Of the 49 published studies, we find the following results:

Four are case studies that do not carry out any statistical analysis of the data.[1]
Eighteen use only descriptive statistics, which offer no basis for generalization.[2]
Five use statistical tests, but fail to apply them to any kind of control group.[3]
Twenty-two use statistical tests in comparison with at least one control group.[4]
Forty-eight lack sufficient statistical power to validate their results.

This means that no scientific generalizations can be reliably made from these data. The researchers have not shown that the results are not a function of chance factors. This chapter begins by explaining

Notes for this section begin on Page 92

different kinds of statistics. It then takes a closer look at inferential statistics, the most important test for this research. It identifies the two types of statistical errors, and focuses on how to avoid Type II error. We conclude with a list of recommendations for good research methodology in this area.

What Statistical Tests Are and Why They Matter

Quantitative summaries of data come in two flavors: descriptive and inferential. Descriptive statistics are inherently limited. Descriptive statistics are used to organize and summarize data. Percentages are a type of descriptive statistics. They do not allow the researcher to scientifically generalize beyond the findings at hand. Eighteen of the studies use only descriptive statistics. Descriptive statistics include single variable (univariate) representations, such as the mean, median, range, standard deviation, interquartile range, and so on. They are statistics performed on one variable, and describe the variable mathematically.

Descriptive statistics may also appear in two-variable (bivariate) form. These are usually expressed as percentage differences between two groups. Less formally, they may take the form of comparing sets of percentages of one group to another.

For example, **Pagelow (1980)** compares the percentage of lesbian mothers and heterosexual mothers with regard to custody problems, living arrangements, home ownership, income after divorce and job discrimination. Similarly, **Javaid (1992)** compares 26 children raised by lesbian mothers with 28 children raised by heterosexual mothers, in terms of percentages. **Javaid** finds that 80 percent of girls in the heterosexual group (12 of 15) desired marriage and children for themselves compared to 55 percent of girls in the lesbian group (6 of 11). **Javaid** notes that a majority of both groups desired marriage and children.[5] But this is the wrong answer to the question of whether being raised by a lesbian mother makes girls less likely to want marriage and children for themselves. The proper scientific question is: "Are girls raised by lesbian mothers less likely to desire marriage and children than girls raised by heterosexual mothers?" Put in methodological terms, the question is whether there is a

statistical association between "being raised by a lesbian mother" and "wanting to be married with children when grown-up." It appears that such an association may exist, but we do not know for certain, because the samples were far too small and no inferential statistical tests were used.[6] Because of their limitations, descriptive statistics have been supplemented with more sophisticated inferential statistics. Inferential statistics enable the investigator to estimate the likelihood of whether or not the data support the research hypothesis. In other words, what is the probability that the differences found in the data are real or a function of chance? Results of inferential statistical testing are typically reported as statistical estimates. This is confusing to many people, but it is much more accurate.

Because descriptive statistics are irrelevant in questions of generalization, we will focus the rest of the chapter on inferential statistics.

Types of Inferential Statistics Used in the Studies

The simplest types of inferential tests are two-variable (bivariate) statistics, such as chi-square and t-tests that rely on one independent variable and one dependent variable. While these measures are a considerable improvement over descriptive statistics alone, they are not entirely satisfactory. They do not provide as complete a range of capabilities as is provided by the general linear model (to be discussed below). They are, however, more easily understood than the more complex statistics and allow a more useful degree of analytic simplification not present in analyzing descriptive statistics such as percentages alone. We will discuss these tools of data analysis used by the researchers.

One-Independent One-Dependent Variable Tests. We find that 18 studies use one-independent/one-dependent variable tests to compare homosexual and heterosexual subsamples.[7] These studies, already severely limited by small, self-selected volunteer samples, are further compromised by using a one-independent variable test. While they are a considerable improvement over percentages alone, statistical tests using only one independent variable are not satisfactory, because they provide no way of controlling statistically for extraneous variables. They do not provide as complete a range of capabilities as is provided by multivariate statistical testing. They are,

however, more easily understood, and allow a greater degree of analytic sophistication compared to presenting descriptive statistics such as percentages alone.

Multivariate Models. The preferred method of statistical testing is reliance on some kind of multivariate statistical test that can mathematically express the relationship of two or more variables on a third. These tests are variations of what statisticians call the general linear model, and are mathematically related. Multivariate statistics allow calculation of the effects of any one independent variable, while holding constant the effects of other variables. A researcher uses multivariate statistical analysis is to control for the relationship(s) between these other variables and the dependent variable. It is a way to avoid attributing the effects of these additional variables to the independent variable. In other words, the investigator should be ruling out spurious correlations or spurious non-correlations. In conceiving of a relationship, the normal approach is to avoid spurious correlation. "Spurious correlation" is the classic situation where one is cautioned against confusing correlation with causation. Here, the situation is that one has a statistically significant bivariate relationship that has been hypothesized to hold true.

The investigator should control (i.e., statistically adjust) for other potentially confounding variables to make sure the significant bivariate relationship is still significant. As noted in our earlier discussion on suppressor variables, spurious noncorrelation is a problem that takes the form of overlooking suppressor variables. Again, because the interest of the investigators is in trying to show that something is not the case, that there is no difference between heterosexual and homosexual parents, the search for potential suppressor variables is of critical importance in either establishing or rejecting their findings.

Flaws in The Studies Despite Using Inferential Tests

Only five studies use statistical techniques with more than two variables. **Brewaeys et al (1997)** use a two-way analysis of variance (ANOVA), **Flaks et al (1995) and Koepke et al (1992)** use multivariate analysis of variance (MANOVA), while **Chan et al (1998) and Green et al (1986)** use multiple regression. That is, they con-

trol for extraneous variables, which they treat statistically as the other independent variables along with parent sexual preference.[8]

Multivariate without Adequate Control Groups. **Koepke et al's** study, however, does not use a heterosexual comparison group. Instead, **Koepke et al (1992)** compare lesbian couples with and without children. Here, too, the lack of a proper comparison group cannot be compensated for by using sophisticated statistics. This study cannot be used as scientific evidence regarding homosexual parents and their children, although the authors conclude that the study can be used for such purposes.[9]

The use of MANOVA, which is a statistical test for situations employing multiple independent and multiple dependent variables in the same estimation procedure (e.g., **Flaks et al, 1995**), also introduces a bias in favor of the investigators' general hypothesis, of finding no differences between homosexual and heterosexual families. This is because the use of multiple dependent variables as well as multiple independent variables reduces the degrees of freedom available for statistical use, and thus reduces the power of the statistical test for correlations of a given size.[10] In short, with the same sample size, and the same overall correlation between independent and dependent variables, it is more difficult to reject the null hypothesis than it would be using a single dependent variable and multiple regressions.[11]

Inferential statistical testing is often misused, but scientific generalizations cannot be made without the use of inferential statistical testing. Use of inferential statistics alone does not however make a study minimally acceptable. Major defects in basic research design, measurement, and data collection cannot be corrected by using inferential statistical testing.

Multivariate Without Proper Comparison Groups. Five studies used inferential statistics within their homosexual samples, but did not use inferential statistics to compare homosexual and heterosexual samples-a critical design flaw, as discussed previously.[12] These studies ought not to be considered as providing scientific evidence regarding the quality of homosexual parenting and/or the condition of their children because they lack an appropriate comparison group.

For example, **Bailey et al (1995)** used inferential statistical tests to compare the rate of homosexuality among adult sons raised by gay fathers with the incidence of gay adoptive brothers, the incidence of gay monozygotic twins, and the incidence of gay dizygotic twins. The rate of homosexuality among adult sons of gay fathers was not statistically different from the rate of homosexuality among adoptive brothers. The incidence of homosexuality among adult sons of gay fathers was lower than either the rate of homosexuality among monozygotic twins or the rate of homosexuality among dizygotic twins. Moreover, the differences were statistically significant.

Bailey et al, however, failed to compare the homosexuality rate of adult sons of homosexual fathers to the rate of homosexuality of adult sons of heterosexual fathers-a serious limitation of their study. **Bailey et al** recognize this limitation, but nevertheless, argue that male sexual orientation is inheritable, not environmental. In their abstract, the authors state that their "results suggest that any environmental influence of gay fathers on their sons' sexual orientation is not large."[13] In fact, the absence of a heterosexual control group means that **Bailey et al**'s results are largely useless.[14]

In another case, **Koepke et al (1992)** used inferential statistics and compared lesbian couples with and without children in terms of the quality of their relationship. Their statistical analysis found lesbian couples with children had higher relationship satisfaction and felt better about their sexual relationship compared with lesbian couples without children. The differences were statistically significant. This led the investigators to the conclusion, among others, that since children thrive in stable, loving relationship between caregivers, the stable and loving quality of lesbian relationships should benefit children. "It is important that educators, clinicians, and practitioners who work with lesbian families understand that *lesbian relationships can provide a positive family environment for child rearing* (italics ours)."[15]

The investigators, however, failed to question the children in the study, failed to include heterosexual couples without children and, most important of all, failed to include a control group of heterosexual couples and their children. They should have compared the

views of the children of the lesbian couples with children of the heterosexual group if they are going to draw any remotely valid conclusions about positive family environments. The authors recognize the design limitations of their study, nevertheless they make policy pronouncements regarding lesbian couples and children anyway.[16]

This leaves us with 22 (or 45 percent) of the studies of homosexual parents and their children that use inferential statistics in any kind of comparison with at least one heterosexual control group.[17] Despite having a proper heterosexual control group, these studies are still flawed.

Lack of Published Estimates. The use of a control group and of statistical significance tests alone is no guarantee of a study's quality. The use of inferential statistics is greatly misleading if the investigators do not publish their statistical estimates, and only report whether the results were significant or not. For example, **Green et al (1986)** carry out several regression analyses. They report the variables entered into their statistical equations,[18] but noticeably fail to report the actual values. These results should be reported, even if they were not statistically significant. Such reporting is important, because the differences may generally be consistent even if not significant, and because the level of significance that is actually achieved may be of considerable relevance. For example, one's interpretation of the results is likely to quite different if the achieved probability value is 0.06 or 0.5. It is also necessary to publish detailed results because it gives independent observers a chance to see what was actually done and to ascertain if standard procedures for analysis of statistical data were followed or not.

Making It Too Easy. Investigators can also use an overly stringent statistical procedure, further increasing the probability of finding non-significant results. All of the studies by **Golombok and Tasker** use what are called Bonferroni corrections, as does **Chan et al (1998).** While often useful in studies that explicitly seek to reject the null hypothesis, they should not be used at all in studies that seek to show that no effect in the data, because they make it too easy for the investigator to prove his case and not, as the inventors of these techniques originally intended, to make it more challenging for the investigator to prove his case. We discuss why this is below.

Chan et al provides an example worth considering at some length. This relatively sophisticated piece of research studied children created via donor insemination and their parents.[19] The authors compare lesbian and heterosexual biological mothers, and the non-biological lesbian parent and the father (non-biological because of donor insemination). The investigators compare statistically the parents' self-reports on several demographic characteristics: age, educational attainment, employment (hours per week), annual individual income, and annual household income. Differences in educational level between the lesbian biological mother and heterosexual mother and between the lesbian non-biological mother and the father were statistically significant.

The critical level of significance needed to reject the null hypothesis was set extremely high. **Chan et al** (as well as **Golombok and Tasker's** studies) use the Bonferroni correction for their statistical tests. The ordinary level of statistical significance is 0.05, meaning that of 100 t-tests; one may obtain significant results five times due to chance. A p value of less of .01 means that significant results of one out of 100 test may be due to chance, and so forth.

Chan et al use t-tests extensively. The t-test is a statistical procedure that compares the means of two groups to see whether the difference between the two means are likely to be due to chance or statistically significant. Since **Chan et al** performed t-tests for several variables for their study and comparison groups, they adjusted their significance levels using the Bonferronni correction so that results were considered statistically significant when they were strong enough to achieve the significance level of .003, or 3 out of 1,000 t-tests rather than the usual .05 level of statistical significance.

There are several problems with the kind of study **Chan et al** carried out. This study on children resulting from donor insemination is exploratory, and it also suffers from small and imbalanced subsamples (e.g., 51 lesbian families versus 25 heterosexual families).[20] These conditions increase the likelihood of finding non-significant results and increase the probability of failing to reject the null hypothesis when it should be rejected.[21]

Moreover, there are differences between lesbian and heterosexual families in the study that favored the lesbian families. The lesbians

were older, had greater individual incomes (e.g., among the biological mothers, lesbians earned $46,000 and heterosexuals earned $31,700), and had greater household incomes ($82,000 versus $63,200) compared to their heterosexual counterparts. The lesbian parents, however, worked fewer hours per week than did their heterosexual counterparts (e.g., for lesbian non-biological mothers, 36.9 hours per week versus 40.2 hours per week for fathers).

The primary problem in using the Bonferroni correction in these studies, however, is that it introduces a bias in favor of the investigators' own hypotheses. The correction adjusts downward the significance level for each individual test so as to increase the probability of finding non-significant results. In the normal case, where the investigators wish to reject the null hypothesis, use of the Bonferroni correction can be very helpful because it counts *against* the investigators own affirmative research hypothesis. In other words, the investigator makes the situation more difficult for himself or herself and if the results warrant it, can have more confidence than otherwise in his or her conclusion.

In the circumstances here, however, the investigators intended to find, as stated in their initial research hypothesis,[22] that there is no difference between the homosexual and heterosexual parents. Thus, applying the Bonferroni corrections favor obtaining a finding of no difference, which is what the investigators are trying to achieve. Rather than making it more difficult for themselves, the investigators have misunderstood the logic of scientific procedure and have made it *easier* for the results to favor their view that there is no difference in child outcomes between homosexual and heterosexual donor inseminated parents. Researchers should carry out the following procedure instead. In situations where the investigators desire to show that no significant differences exist, we propose instead that the logical parallel of employing the Bonferroni correction in the normal situation is to use a less stringent significance level for rejecting the null hypothesis than is normally used. For example, a study might use a significance value of $p < .10$.[23] None of the investigators (including **Chan et al**) do this, even though **Chan et al** is the only research team to mention the problem of accepting the null hypothesis as a major logical problem that their study and all the other studies face. It is one thing to notice a problem, but another thing

to solve it. **Chan et al**, perhaps the most technically sophisticated of the studies, fails in this task.

What Went Wrong and What Needs to be Done

The lessons of Step 5, Using Statistical Tests, are as follows:

1) Do the prior four steps correctly so that your testing will not be a waste of time. Don't try to make up for poor design or controls by using fancy statistical techniques.

2) Use inferential statistics so that you may properly assess whether your findings are due to chance or not.

3) Properly correct for repeated use of statistical tests, keeping in mind that you are engaged in the difficult task of "trying to affirm the null."

4) Make sure statistical tests do not inadvertently favor the investigator's hypothesis. One should not use applications of conventional statistical procedure, developed for affirming the alternative research hypothesis, without extensive justification when trying to find "no difference." When a routine is developed to find "X," the investigators looking for "Not X" should be made to explain why the routine is applicable to "Not X" as well.

If statistical routines do not follow the steps above, the findings of "no difference" should be considered unreliable.

Notes to Chapter 5

1. Barret and Robinson, 1990; McCandlish, 1987; Ross, 1988; Weeks et al, 1975.

2. Bozett, 1980; Gartrell et al, 1996; Green, 1978, 1982; Hare, 1994; Javaid, 1992; Kirkpatrick et al, 1981; Lewin and Lyons, 1982; Lewis, 1980; Lott-Whitehead and Tully, 1992; Lyons, 1983; Miller, 1979; O'Connell, 1993; Pagelow, 1980; Pennington, 1987; Riddle and Arguelles, 1989; West and Turner, 1995; and Wyers, 1987.

3. These studies are: Bailey et al, 1995; Crosbie-Burnett and Helmbrechty, 1993; Rand et al, 1982; Turner et al, 1990; and Koepke et al, 1992.

4. The studies using inferential statistical test(s) between homosexual and heterosexual groups are: Bigner and Jacobsen, 1989a, 1989b, 1992; Brewaeys et al, 1997; Cameron and Cameron, 1996; Chan et al, 1998; Flaks et al, 1995; Green et al, 1986; Golombok and Tasker, 1996; Golombok et al, 1983; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Kweskin and Cook, 1982; McNeill, 1998; Miller et al, 1982; Mucklow and Phelan, 1979; Patterson, 1994a, 1994b, 1997; and Tasker and Golombok, 1995, 1997.

5. p. 241.

6. The data supplied here yield a percentage difference of 25 percent or an odds ratio of 3.33. Using the Fisher's exact test, the associated one sided p value for this sample is 0.0865, while the power of the test is only 22 percent! The far-reaching implications of these results are discussed below. It is likely that there is an effect.

7. These are: Bigner and Jacobsen, 1989a, 1989b, 1992; Cameron and Cameron, 1996; Golombok and Tasker, 1996; Golombok et al, 1983; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Kweskin and Cook, 1982; MacNeill et al, 1998; Miller et al, 1982; Mucklow and Phelan, 1979; Patterson, 1994a, 199b, 1997; and Tasker and Golombok, 1995, 1997.

8. Brewaeys et al (1997), Chan et al (1998), Flaks et al (1995), and Green et al (1986) are the only four studies that use a heterosexual control group and multivariate statistical testing. They still fall short of adequate analysis and proper statistical testing, however, because (as discussed below) their samples are much too small to yield statistically significant results. Cameron and Cameron (1996) fail to use any multivariate statistical methods in analyzing their data.

9. p. 228.

10. Cohen, 1988, p. 474.

11. Ibid.

12. These studies are: Bailey et al, 1995; Crosbie-Burnett and Helmbrechty, 1993; Rand et al, 1982; Turner et al, 1990; and Koepke et al, 1992.

13. 1995, p. 124.

14. A detailed analysis of these and other twin studies can be found in Satinover, 1999, pp. 24-50.

15. p. 228.

16. See "Lesbian Relationship as a Context for Raising Children," Koepke et al, 1992, p. 228.

17. The studies using inferential statistical test(s) between homosexual and heterosexual groups are: Bigner and Jacobsen, 1989a, 1989b, 1992; Brewaeys et al, 1997; Cameron and Cameron, 1996; Chan et al, 1998; Flaks et al, 1995; Green et al, 1986; Golombok and Tasker, 1996; Golombok et al, 1983; Harris and Turner, 1985; Hoeffer, 1981; Huggins, 1989; Kweskin and Cook, 1982; McNeill, 1998; Miller et al, 1982; Mucklow and Phelan, 1979; Patterson, 1994a, 1994b, 1997; and Tasker and Golombok, 1995, 1997.

18. p. 171.

19. 1998.

20. Couples that resort to donor insemination are not typical of families generally, so even if the results were valid, they could not be generalized to the larger American population.

21. In statistical terms they increase the probability of finding Type II error. We will discuss "Type II error" in Chapter 6.

22. Chan et al, 1998, p. 444.

23. A more sophisticated alternative would be to treat the research hypothesis as the null hypothesis and the null hypothesis as the research hypothesis, which, however, would require more statistical sophistication, than even most quantitative researchers posses, but see note 34 in Chapter I for some alternative suggestions.

Chapter 6
Give Me More Power:
How the Studies Find False Negatives

The Big Problem of Finding False Negatives

Since the 1960s, statisticians have developed practical means for researchers and evaluators to decide how large a sample is necessary to carry out certain kinds of research designs and statistical tests. The logic behind such testing is connected to the problem of finding false negatives. Every conclusion drawn from every test, including medical tests, face this problem. That is, what are the chances of a test claiming to find nothing-when, in fact, there is something. The popular press is replete with stories of the cancerous lump that showed up negative until it was too late, problems with an airplane that do not show up because the tests used had failed to warn manufacturers that something was wrong, and the prisoner released because of his positive psychological tests who then turns around and murders someone.

In most situations, since the test is designed to find something, the probability of making conclusions based on false negatives is not central when something is in fact found. However, when the research hypothesis and tests are used to "find nothing," the false-negative problem rears its head. Of course, the problem of false negatives is central to the lack of quality of the research in the homosexual parenting studies. The problem of the false negative is closely tied to the logic of inferential statistical testing. We will discuss these matters below in some detail. Our discussion includes such topics as: the role of the null hypothesis in inferential statistical testing; ad-

Notes for this section begin on Page 108

equate sample size; and adequate statistical power. This discussion ends with a critical evaluation of homosexual parenting studies, and generally concludes that the samples are so small as to reject the conclusions drawn in the studies. These studies, except for one, use samples that are much too small, and thus make it likely (e.g., more than 50 percent of the time) that they will find non-significant results and wrongly conclude that there is no difference between homosexual and heterosexual parents in their effects upon their children.

The Process of Statistical Testing: An Overview

Assume that a study "finds" something. Are these "findings" just the result of chance? It is important to know the answer to this question, so we do not make false claims. The goal of statistical testing is to rule out findings that are likely due to chance. The essential logic of these tests can be found in any applied statistics textbook.[1] Here is a brief synopsis. It involves three steps: (1) Formulating the "alternative hypothesis," (2) Formulating the "null hypothesis," and (3) Testing the hypotheses.

The Research Hypothesis. The proper way to use statistics in the analysis of data is to formulate a research hypothesis before collecting and analyzing data. This is what we discussed in Chapter 1. The statistical testing literature refers to this as the alternative hypothesis. The alternative hypothesis can take many different forms depending upon the statistical model tested.[2] In all cases, however, the alternative hypothesis must take the form of an affirmative research hypothesis. That is to say, the hypothesis must posit that genuine differences exist between the groups.

The Null Hypothesis. The researcher should then formulate the appropriate null hypothesis. The precise form of the null hypothesis depends on the statistical test used.[3] Statistically, the null hypothesis is the actual hypothesis tested on the data: "What is the likelihood that the results are due to chance?" Normally (unlike the situation for the studies being evaluated here (the goal of the test is to strengthen the affirmative hypothesis by showing that the null hypothesis may be validly rejected.

Testing the Hypotheses. Once the data is collected and the test hypotheses (affirmative versus null) are formulated, the actual process of testing is as follows. The investigators select a level of statistical significance. This is usually 0.05 or 0.01 in the typical research situation where the investigators seek to reject the null hypothesis.

Let us assume the significance level is the conventional 0.05. Under the null hypothesis, the investigator (or more typically, with the help of a computer's statistical software) calculates a test statistic and computes an accompanying p value for each variable in the statistical equation. If the p value is less than or equal to the preset level of significance (e.g., $p < 0.05$), the null hypothesis is rejected, and the results are said to be statistically significant. If it is not, then the investigator fails to reject the null hypothesis.

It must be stressed here that the significance level of 0.05 is a probability level (or an error rate.) The 0.05 level of statistical significance means that 5 percent of the results may be due to chance even when the "true" effect is in fact zero. (Alternatively put, the error rate is 5 percent.) If we generate 20 estimates and 20 corresponding p-values, 1 out of these 20, 5 percent of the results, will be statistically significant due to chance alone. We just do not know which of the 20 results is a function of chance, and which are true results. We can be more stringent, by setting our significance level at 0.01, whereby 1 percent of the results are due to chance, or at 0.001, whereby 0.01 percent (1 out of 1,000) are due to chance. Researchers can never be 100 percent sure that our results are not due to chance. We can be more certain of not making this kind of error if the significance level is set at .01 compared to .05, and 0.05 compared to 0.20, whereby 20 percent of the results generated (and we don't know which ones) are due to chance.

If we set the significance level at .20, more results (20 percent) would be found to be statistically significant. This means that we would have more results whereby we would reject the null hypothesis. Unfortunately, we would have also increased the chances to one in five where we reject the null hypothesis based on our calculations even if the null hypothesis is true. So, there is a limit to how high the significance level can be.

Minimizing False Positives and False Negatives: Type I and Type II Errors

The False Positive. The false positive is what statisticians refer to as Type I Error. That is, the test yields a positive result that is in reality false. As a result, the investigator rejects the null hypothesis despite the null hypothesis being true. The higher the significance level is set, the greater the chance is of a Type I error occurring. For example, a p level of .001 means that there is a 1 in 1,000 possibility that the results found are due to chance. A higher p level, say at 0.20, means that there is a 20 out of 100, or one in five chance that the results are due to chance. The higher the investigator sets the level of significance, the greater will be the probability of committing a Type I error, so normally it is set to be quite low.4

The False Negative. Critical to homosexual parenting studies is finding the false negative, or the problem of Type II Error. Type II Error occurs when one fails to reject the null hypothesis when it should be rejected and where in fact the null hypothesis (e.g., "no difference") is false. Finding false negatives means that the test gives you a negative result (a null result) that is in reality false. It applies to results found to be statistically non-significant-the exact sorts of results these studies claim to have found. There are research design and statistical techniques to decrease the probability (or rate) of Type II Error, however. We turn next to a discussion of how to reduce the probability of obtaining false negative findings.

Increasing Statistical Power to Avoid Type II Error

The probability of committing a Type II error, that is, of failing to reject the null hypothesis when it in fact should be rejected, is a function of the following factors:

The size of the sample,
The significance level chosen by the investigator,
The size of the effect anticipated or sought for (e.g., small, medium, or large), and
The power of the particular statistical test used.

These studies of homosexual parents and their children are particularly prone to Type II error. We will examine each of these four factors in turn.

Sample Size. All of the studies, except for **Cameron and Cameron (1996),** suffer from having samples that are in fact so small that the statistical tests used have extremely low power, and the investigators have a high probability of failing to reject the null hypothesis when they should in fact reject it. In other words, these studies on homosexual versus heterosexual parents and their children have such small samples that there is a high probability that the results are deemed non-significant, when in fact a difference between the homosexual and heterosexual groups exists. It is quite possible, then, that the failure to reject the null hypothesis as reported by many of the researchers is simply due to the fact that the samples they use are too small. We discuss this below.

Significance Level. We begin our evaluation of the statistical power of these studies by presetting the significance level at the conventional 0.05 level. When the basic research orientation is to affirm the null hypothesis, however, the case could be made that the investigators should preset the level of statistical significance at 0.10 or even higher. Levels of 0.05 or 0.01 were designed for research aiming to disprove an affirmative hypothesis. Those with other goals may easily manipulate the results. At the 0.05 level of significance, the researchers erroneously reject the null hypothesis 5 percent of the time, when they should not reject it. Conversely, 95 percent of the time, the researchers properly reject the null hypothesis. The research hypothesis, however, is conventionally phrased in the affirmative format. A significance level of 0.05 is not methodologically necessary when the researchers seek to find no difference, because the same level of significance makes it easier for these researchers to affirm their (null) hypothesis. In other words, a 0.05 level of significance makes it easier for researchers to find what they are looking for when they think they will find no differences. However, we keep to the usual convention by requiring the $p<.05$ level of statistical significance.

Effect Size. We preset the size of the effect we expect to find in the data to be "small" effect sizes, based on guidelines set by the late

noted pyschometrician Jacob Cohen.[5] Why do we select the small effect as the one to be detected? Because Cohen[6] and others have found that many effects in personality, clinical, sociological, political, and psychological research are small. This is due to various factors: greater unreliability of the data gathered in the field (as opposed to experimental) settings, because experimental controls are lacking, because measurement is imprecise, and because the effects sought after are often subtle.

One should also look for small effects when research is new,[7] because new research is more unreliable than "old" research (research that has been replicated again and again for many years). These studies of homosexual parenting are all "new" and original studies. None is a replication of other studies. No two studies select the samples in the identical way, set up the identical criteria for eligibility, define the independent, dependent, and possible extraneous variables in the identical way, measure the variables in the identical way, nor use the identical statistical tests. The proper size of effect to be detected for the studies on homosexual versus heterosexual parents and their children, then, is small.

A final yet extremely important reason that the effects that one might expect to have to detect should be small is because of the nature of what these studies are trying to show. As noted earlier, the authors of these studies are trying to find no differences between heterosexual and homosexual parents in terms of their effects on their children. They wrongly seek to confirm the null hypothesis. One appropriate means of trying to reach their research objective without mistakenly accepting the null hypothesis is to create a statistical test situation that seeks to detect a very small effect size. The smaller the effect size sought and the more consistent the results are in failing to reject the null hypothesis despite explicitly searching for the smallest possible effect, the more likely it is that the "no-difference" findings that are desired by all but one of the investigators describes the true situation accurately. As we will see, the actual procedures used in these studies are far from this ideal case.

Statistical Power. We must now set the value for sufficient statistical power, and thus the desired ratio for Type II error. Cohen sets

the value for sufficient statistical power at 0.80. Presetting the level for sufficient power is important in the following way. Statistical power is mathematically related to the probability of Type II error, whereby statistical power equals 1.00 minus the probability of Type II error. By presetting the statistical power level at 0.80, we arrive at the probability of Type II by subtracting 0.80 from 1.00. By presetting our power level at 0.80, we thus set the probability of Type II error as 0.20, or 20 percent. When one states there is a probability of Type II error of 0.20, it means that 20 percent of our non-significant results may, in reality, mask a true relationship. Stated somewhat differently, in 20 percent of the instances in which the researcher fails to reject the null hypothesis, in reality, he or she should have rejected the null hypothesis.

If through the researchers' calculations, statistical power is found to be less than 0.80, then the statistical testing has insufficient power to limit the probability of Type II errors. In other words, if we find through our calculations that the study has a degree of statistical power lower than 0.80, we should treat any non-significant results as inconclusive, despite a failure to reject the null hypothesis. No further conclusions should be drawn. Future studies should employ far larger samples to yield more power and better tests. [8]

Statistical Power: Evaluating the Studies

In evaluating published studies that claim to have non-significant findings, the size of the sample is related to the probability of finding the false negative. Smaller samples yield less power and an increasing probability of the Type II error described above.

We have compiled a table that shows the number of subjects needed for each statistical test, based on the tables of subjects provided by Cohen. As suggested by Cohen, to find the optimal number of subjects, we first had to set the effect size at small, the level of sufficient power at 0.80, and the significance level at 0.05.9 We report the adequate sample sizes for the statistical tests used in the homosexual parent-child studies. Looking up each individual statistical test in Cohen,[10] we find the following (see Table 5).

Table 5.
Determining the Minimum Number of Cases to Reduce the Likelihood of False Negatives.

Statistical Test	Number of Cases for an Adequate Sample
T-Test	393 [11]
Pearson Correlation Coefficient	783
Chi-Square	785 [12]
ANOVA	393 [13]
Regression	390 for 1 independent variable 485 for 2 independent variables 556 for 3 independent variables 615 for 4 independent variables

The t-test requires the smallest minimum sample size for a Type II error rate of 0.20 and a small effect size. Based on this table, a study using a t-test, which tests for the statistical significance of the differences between group means, needs a sample of 393 cases. If we are to have the same number of cases in the study and comparison group, this requires selecting sub-samples of 197 each. Even so, this would mean a power level of 0.80, or a Type II error rate of 0.20, or one in five findings.

The studies reviewed for this report all fall short of the minimum sample size for the test statistic used, with the exception of **Cameron and Cameron (1996).** These authors have a sample of more than 5,000 cases. The next largest sample size is a study of 164 children, as described by their gay and lesbian parents (**Riddle and Arguelles, 1989**). It is not included in Table 2 below because it lacks a heterosexual control group.

If we actually calculate the Type II error rates, we can see that the study's sample sizes are inadequate. As we can see in Table 6, all the studies of homosexual versus heterosexual groups have large problems with Type II error, except for **Cameron and Cameron**.

Table 6.

Statistical Power of Studies on Homosexual versus Heterosexual Parents and/or their Children

Authors	Year	Largest Sample for Test	Statistic Used	Statistical Power	Probability of Type II Error
Bigner and Jacobsen	1989a	66	T-test	.21	79%
Bigner and Jacobsen	1989b	66	T-test	.21	79%
Bigner and Jacobsen	1992	54	Chi -Square	.09	91%
Brewaeys et al	1997	72	ANOVA	.22	78%
Cameron and Cameron	1996	713	Chi-Square	.75	25%
Chan et al	1997	77	T-Test	.14*	86%
Flaks et al	1995	30	MANOVA	<.10 [14]	90%+
Golombok and Tasker	1996	45	Pearson	.10	90%
Golombok et al	1983	75	T-test	.23*	77%
Green et al	1986	104	M.Regression	.13	87%
Harris and Turner	1985	27	ANOVA	.08	92%
Hoeffer	1981	40	ANOVA	.14	86.5%
Huggins	1989	36	T-test	.13	87%
Kweskin and Cook	1982	44	Chi-Square	.08	92%
McNeill et al	1998	59	T-test	.11*	89%
Miller et al	1982	81	Chi-Square	.15	85%
Mucklow and Phelan	1979	81	Chi-Square	.15	85%
Patterson	1997	95	T-test	.15*	85%
Patterson	1994a	95	T-test	.15*	85%
Patterson	1994b	95	T-test	.15*	85%
Tasker and Golombok	1995	45	Pearson	.10	90%
Tasker and Golombok	1997	45	Pearson	.10	90%

*Unequal Groups

The implications of this table may be clearer if several examples are analyzed. For example, **Bigner et al (1989a, 1989b)** compare 33 gay fathers with 33 presumed heterosexual fathers. Using a t-test, they find no significant differences between the two groups regarding their attitudes toward children. A sample of 66 cases however, would yield a power value of 0.21.[15] This yields a Type II error of

0.79. In other words, the probability of failing to reject the null in **Bigner et al's** study is 79 percent.

Chan et al use t-tests and multiple regression analysis to study parents who have children from donor insemination, including 16 heterosexual couples, 34 lesbian couples, nine single heterosexual biological mothers, and 21 single lesbian mothers. **Chan et al** make reference to the problem of sufficient statistical power, but they misapply Cohen's techniques. **Chan et al**[16] state that their total sample of 80 donor insemination families is large enough to detect differences between groups for medium and large effect sizes. Despite citing Cohen's techniques, **Chan et al** downplay their study's lack of sufficient statistical power to detect small effects, by ignoring all of Cohen's recommendations regarding new and exploratory studies being investigated for small effects, as previously cited.

As important, the number of respondents varies, depending on **Chan et al**'s statistical analysis. For example, **Chan et al**'s largest sample is 55 lesbian biological mothers versus 25 heterosexual biological mothers, which they use for a t-test on various demographic data. Power calculations give us a power level of 0.13,17 or a probability of 87 percent for Type II error (failing to reject the null hypothesis). When comparing teacher's reports of child's behavior, **Chan et al** perform t-tests on 48 teacher's reports–32 children raised by lesbians versus 16 children raised by heterosexuals.[18] This gives us a power level of 0.10, or a Type II Error rate of 90 percent. In short, **Chan et al**'s results of non-significance should be treated as non-findings or inconclusive results, because the samples are too small and the statistical tests have too little power.

In another example, **Golombok et al (1983)** conducted a study with a sample of 75 British children of lesbian and heterosexual mothers, using a t-test in their statistical analysis. Assuming small effect sizes and a power level of 0.80, **Golombok et al**'s sample of 75 has a power value of only 0.23, which gives us a Type II error value of 0.77. That is, the probability of failing to reject the null hypothesis when they should have rejected it is 77 percent. Their findings cannot be considered to be anything but inconclusive. No conclusions or generalizations should be drawn from their work.[19]

McNeill et al are the only investigators to publish their statistical power values for each statistical test conducted.[20] **MacNeill et al** found no significant differences between lesbian and heterosexual mothers regarding family and relationship dysfunction, parent-child relationships, and knowledge of proper parenting role.21 Statistical power levels range from 0.33 to 0.72, although **McNeill et al** do not report the effect sizes they used in making these calculations. However, **McNeill et al's** power calculations do not match ours for small, medium or large effect sizes. The largest sample size for a t-test in **McNeill et al**'s study is 59 cases (24 lesbian and 35 heterosexual mothers), which gave us a power value of 0.11 for two unequal sub-samples and small effect sizes and a significance level of 0.05.[22] None of the four executed statistical tests rise to the level of having sufficient power, even by **McNeill et al**'s calculations. They make no comment regarding insufficient statistical power, although they did acknowledge that their overall sample was small. **McNeill et al** should have concluded that their sample was too small to draw statistical conclusions and redone their study with a much larger sample. Once again, non-significant results yield inconclusive findings.

In another example, **Patterson**[23] uses a t-test and a sample of 35 children born to lesbian mothers via donor insemination and 60 assumed heterosexual children from another study by R. A. Eder.[24] This gives us a power value of 0.15, or a probability of Type II error of 85 percent. Once again, the lack of power means that nothing can be usefully said about all of **Patterson**'s non-significant findings.

Other studies use chi-square as their primary test statistic. These studies need a minimum sample size of 785 to speak meaningfully of non-significant differences. This means a study and comparison group of 393. The samples for these studies are all too small, and non-significant results should be treated as inconclusive.

For example, **Bigner and Jacobsen (1992)** performed a chi-square on a sample of 53 homosexual and heterosexual fathers. They find no significant differences regarding attitudes towards their role as fathers, nor any regarding their parenting styles. The

calculated power value for this study is 0.09, or a probability of Type II error of 91 percent. **Harris and Turner** use a chi-square to compare 33 gay and lesbian parents with 16 single heterosexual mothers. For a sample of 39, this gives a power value of 0.10, or a 90 percent chance of Type II error. **Kweskin and Cook (1982)** compare 44 heterosexual and homosexual mothers regarding the sex-role behavior of their children and their ideal sex-role behavior for a child. The calculated power value for this study is 0.08, or a Type II error rate of 92 percent. **Mucklow and Phelan (1979)** compare 81 heterosexual and homosexual mothers, regarding their responses to child behavior and their self-concepts. They found no significant differences between the two groups, using a chi-square. Calculations show an 85 percent probability of falsely rejecting the null hypothesis.

Other statistics such as MANOVA[25] and multiple regression[26] impose even more stringent criteria, and require even larger samples, for achieving power levels of 0.80. **Flaks et al (1995)** for example compares 15 lesbian donor inseminated (DI) couples with 15 heterosexual DI couples (e.g., 30 biological mothers) and claim no significant results regarding parents' reports of their child's behavior and teacher's reports of the children. A simple t-test, which is the far less restrictive than is a MANOVA, on a sample of 30 subjects yields a power value of 0.12, and a Type II error rate of 88 percent. **Flaks et al**'s results of non-significant differences should be considered inconclusive. Power calculations on a multiple regression test on **Flaks et al**'s sample of 30 would give us a power value of less than 0.10, or a Type II error rate of greater than 90 percent. The same problems confront **Green et al (1986).** A multiple regression analysis on 104 children of lesbian and single heterosexual mothers results in a power value of 0.13, or a Type II error rate of 87 percent.

What Went Wrong and What Needs to be Done

The lessons of Step 6: Having Enough Power, are as follows:

1) Do the prior five steps correctly. When done correctly, your testing will not be a waste of time. Don't try to make

up for poor design or controls by using fancy statistical techniques. If you use fancy statistical techniques, make sure you use them properly-congruent with the kind of hypothesis to be tested. Don't mindlessly use conventions that were developed for affirmative research hypothesis testing when you're trying to "affirm the null."

2) Use a large enough sample. The studies of homosexual parenting that rely on inferential statistical testing have samples that are much too small to arrive at any genuine conclusions of "no significant difference" between the study and comparison groups. These studies must be replicated with significantly larger samples before their non-statistically significant findings can be taken seriously.[27] These calculations can be done before a study is executed and future research should include and report their power calculations as a matter of routine.

3) Set the significance level higher than 0.05. Appropriate conventions must be used regarding significance levels when the basic research orientation is to affirm, rather than reject, the null hypothesis. Perhaps the investigators should preset the level of statistical significance at 0.10 or even higher. Levels of 0.05 or 0.01 were designed for research aiming to disprove an affirmative hypothesis. They are too easily manipulated.[28]

4) Set the effect size so you can detect a "small" effect. Many effects in personality, clinical, sociological, political, and psychological research are small. Preset the size of the effect you expect to find to be "small," based on Cohen's guidelines.[29]

5) Set statistical power levels to find non-significance higher than Cohen's 0.80 level. The 0.80 level is standard for studies that set up the proper affirmative research hypothesis. Conversely, when the research is set up to seek no difference, the power level should be set at 0.90. This gives us a Type II error rate of 0.10, whereby there is only a 10 percent chance

of failing to reject the null hypothesis. This would make the procedure of "seeking to not reject the null hypothesis" more similar to the standard procedure when seeking to reject the null hypothesis.[30]

If these five steps were followed, and then replicated over time, we may learn something about homosexual parenting, even something relevant to our public policies.

As things currently stand, however, these studies display an unreflective, rote-like application of statistical methods. The researchers seem to have spent no time reflecting upon what these statistical tests and methods mean. The critical methodological issue is how to "test" for non-significance, i.e., the task of establishing the appropriate standards. The standards for testing affirmative research hypotheses are well established and a part of a standard social science statistics textbook. Until similar standards for testing research hypotheses of non-significance are established and widely accepted in the social sciences, these small studies claiming non-significant results must be treated as entirely inconclusive.

Notes to Chapter 6

1. See for example, Agresti and Findlay, 1996 and Cohen, 1988.

2. For example, a t-test statistically compares the difference between the mean of the study group versus the mean of the comparison group. A simple ANOVA involves a test of the hypothesis that the means of all the groups are not equal to zero. Multiple regression involves a test of the hypothesis that all the regression coefficients are not equal to zero or that one or more of them is either positive or negative.

3. The null hypothesis for a t-test is that the difference between the means of the two groups is equal to zero. The null hypothesis for the ANOVA is that all means are equal to each other. Other common null hypotheses are that the sum of the differences between the expected and actual values in the cells of a cross-tabulation or table is zero (the chi-square test); or that the correlation coefficient, which measures the linear association between two variables, is equal to zero, or that one or more of the regression coefficients is equal to zero. Of course it is very unlikely that the value of these coefficients will be exactly zero when they are computed using an actual data set. The statistical testing question

is whether they are statistically significantly different from zero.

4. The usual convention is the .05 level of statistical significance.

5. 1988, p. 83.

6. 1988, p. 13.

7. Cohen, 1988, p. 25.

8 Statistically significant results, however, indicate that the power level was adequate. This is why such power analysis is absolutely vital for studies that seek to "accept" the null hypothesis.

9. In fact, Cohen suggests that studies that seek to "demonstrate" the null hypothesis should use the even higher power level of .90 (1988, p. 104).

10. 1988.

11. The needed sample sizes for t-tests and ANOVA is relatively small because the calculation in the table is based on two assumptions: 1) that the size of the two groups are equal and 2) the variances of the two groups with respect to the dependent variable are the same. When this is not the case, the needed number of cases is about the same as needed for the correlation coefficient and the regression equation respectively.

12. For 1, 2, and 3 degrees of freedom (df) respectively, which covers the chi-square tests for all the studies reviewed here.

13. For df=1, which covers all the studies reviewed here.

14. Calculated using multiple regression which has more power than MANOVA.

15. See Cohen's table of power values for a sample of 66 cases, pp. 36-37.

16. 1998, p. 454.

17. Chan et al. have unequally sized sub-samples, therefore we use the formula provided by Cohen for t-tests for unequal groups (1988, p. 42).

18. Table 3, Chan et al, 1998, p. 450.

19. Since, as we have noted earlier, some of Golombok's data suggest definite effects of parental sexual orientation on adult's children's orientation, reanalysis of their data is probably required in order to determine exactly what is going on.

20. McNeill et al, 1998, p. 61.

21. p. 60.

22. Cohen, p. 36, p. 42.

23. 1994a, p. 168.

24. As reported in Patterson, 1994a, p. 167.

25. Flaks et al, 1995.

26. Green et al, 1986; Chan et al, 1998.

27. Cohen's power values show, at a minimum, samples of 393 subjects. Of course, this assumes sub-samples of equal sizes, meaning a minimum of 197 in the study group and 197 in the comparison group. None of the studies except Cameron and Cameron (1996) come close to these numbers.

28. At the 0.05 level of significance, the researchers erroneously rejects the null hypothesis 5 percent of the time, when they should not reject it, or conversely, 95 percent of the time, the researchers properly reject the null hypothesis. Methodologically, however, the research hypothesis is conventionally in the affirmative format. A significance level of 0.05 is less methodologically sound when the researchers seek to find no difference, because the same level of significance makes it easier for this researcher to affirm their (null) hypothesis. In other words, a 0.05 level of significance makes it easier for researchers to find what they are looking for; when they think they will find no differences, compared to when a research hypothesis is conventionally set up in the proper form.

29. 1988, p. 83.

30. Using a t-test, for two equal groups, if we assume a 0.20 level of significance, and a 0.90 power value, and a small effect size, we would need a sample of only 338 respondents. Of course, with a t-test, we would have no way of testing for the effects of extraneous variables.

Appendix 1
Bibliography

A. Studies Evaluated

Bailey, J.M., Bobrow, D., Wolfe, M., and Mikach, S. (1995). Sexual orientation of adult sons of gay fathers. *Developmental Psychology*, 31, 124-129.

Barret, R. L., and B.E. Robinson, 1990, Children of gay fathers, in R.L. Barret and B.E. Robinson, *Gay fathers.* Lexington, MA: Lexington Books.

Bigner, J.J. and Jacobsen, R.B. (1992). Adult responses to child behavior and attitudes toward fathering: Gay and nongay fathers. *Journal of Homosexuality*, 23 (3), 99-112.

Bigner, J.J., and Jacobsen, R.B. (1989a). The value of children to gay and heterosexual fathers. *Journal of Homosexuality*, 19 (1/2), 163-172.

Bigner, J.J., and Jacobsen, R.B. (1989b). Parenting behaviors of homosexual and heterosexual fathers. *Journal of Homosexuality*, 18 (1/2), 173-186.

Bozett, F. 1980, Gay fathers: how and why they disclose their homosexuality to their children. *Family Relations*, 29, 173-179.

Brewaeys, A., I. Ponjaert, E.V. Van Hail, and S. Golombok, 1997, Donor insemination: child development and family functioning in lesbian mother families with 4 to 8 year old children. *Human Reproduction* 12, 1349-1359.

Cameron, P. and Cameron, K. (1996). Homosexual parents. *Adolescence*, 31(124), 757-776.

Chan, R.W., Raboy, B., and Patterson, C.J. (1998). Psychosocial adjustment among children conceived via donor insemination by lesbian and heterosexual mothers. *Child Development* 69(2), 443-457.

Crosbie-Burnett, M., and Helmbrechty, L. (1993). A descriptive empirical study of gay male stepfamilies. *Family Relations* 42 (July), 256-.

Flaks, D.K., Ficher, I., Masterpasqua, F. and Joseph, G. (1995). Lesbians choos-

ing motherhood: A comparative study of lesbians and heterosexual parents and their children. *Developmental Psychology* 31, 105-114.

Gartrell, N., Hamilton, J., Banks, A., Mosbacher, D., Reed, N., Sparks, C.H., and Bishop, H. (1996). The national lesbian family study: Interviews with prospective mothers. *American Journal of Orthopsychiatry* 66 (2), 272-281.

Golombok, S. and Tasker, F. (1996). Do parents influence the homosexual orientation of their children? Findings from a longitudinal study of lesbian families. *Developmental Psychology*, 32, 3-11.

Golombok, S., Spencer, A., and Rutter, M. (1983). Children in lesbian and single-parent households: Psychosexual and psychiatric appraisal. *Journal of Child Psychology and Psychiatry* 24 (4), 551-572.

Green, Richard, 1978, Sexual identity of 37 children raised by homosexual or transsexual parents. *American Journal of Psychiatry*, 135, 692-697.

Green, R. (1982). The best interests of the child with a lesbian mother. *Bulletin of the AAPL* 10 (1), 7-15.

Green, R., Mandell, J.B., Hotvedt, M.E., Gray, J., and Sarnith, L. (1986). Lesbian mothers and their children: A comparison of solo parent heterosexual mothers and their children. *Archives of Sexual Behavior*, 15. 167-184.

Hare, J. (1994). Concerns and issues faced by families headed by a lesbian couple. *Families in Society: The Journal of Contemporary Human Services.* 75 (1), 27-35.

Harris, M., and Turner, P. (1986). Gay and lesbian parents. *Journal of Homosexuality*, 12, 101-113.

Hoeffer, B. (1981). Children's acquisition of sex-role behavior in lesbian-mother families. *American Journal of Orthopsychiatry*, 51. (3), 536-544.

Huggins, S.L. (1989). A comparative study of self-esteem of adolescent children of divorced lesbian mothers and divorced heterosexual mothers. *Journal of Homosexuality*, 18(1-2), 123-135.

Javaid, G.A. (1993). The children of homosexual and heterosexual single mothers. *Child Psychiatry and Human Development*, 23 (4), 235-248.

Kirkpatrick, M., Smith, C., and Roy, R. (1981) Lesbian mothers and their children: a comparative survey. *American Journal of Orthopsychiatry*, 51, 545-551.

Koepke, L., Hare, J., and Moran, P.B. (1992). Relationship quality in a sample of lesbian couples with children and child-free lesbian couples. *Family Relations*, 41, 224-229.

Kweskin, S.L., and Cook, A.S. (1982). Heterosexual and homosexual mothers' self-described sex-role behavior and ideal sex-role behavior in children. *Sex Roles* 8, 967-975.

Lewin, Ellen, and Terrie A. Lyons, 1982, Everything in its place: the coexistence of lesbianism and motherhood. In W. Paul, J. Weinrich, J. Gonsiorek and M. Hotvedt (ed.), *Homosexuality—Social, Psychological and Biological Issues.*

Lewis, Karen Gail, 1980, Children of lesbians: their point of view. *Social work* (May), 198-203.

Lott-Whitehead, L., and Tully, C. (1992). The family lives of lesbian mothers. *Smith College Studies in Social Work*, 63.265-280.

Lyons, Terry A., 1983, Lesbian mothers' custody fears. *Women and Therapy* 2, 231-240.

McCandish, B. (1987). Against all odds: Lesbian mother family dynamics. In F. Bozett, ed., *Gay and lesbian parents* (pp. 23-38), New York: Praeger.

McNeill, Kevin F., Beth M. Rienzi, and Augustine Kposowa, 1998, Families and parenting: a comparison of lesbian and heterosexual mothers. *Psychological Reports* 82, 59-62.

Miller, B. (1979). Gay fathers and their children. *The family coordinator* 28 (4), 544-552.

Miller, J.A., Jacobsen, R.B., and Bigner, J.J. (1982), The child's home environment for lesbian versus heterosexual mothers: a neglected area of research. *Journal of Homosexuality* 7(1), 49-56.

Mucklow, B.M., and Phelan, G.K. (1979), Lesbian and traditional mothers' responses to adult response to child behavior and self-concept. *Psychological Reports* 44, 880-882.

O'Connell, A., 1993, Voices form the heart: the developmental impact of a mother's lesbianism on her adolescent children. *Smith College Studies in Social Work*, 63, 281-299.

Pagelow, M.D., 1980, Heterosexual and lesbian single mothers: a comparison of problems, coping, and solutions, *Journal of Homosexuality* 5(3), 189-204.

Patterson, C.J. (1994a) Children of the lesbian baby boom: behavioral adjustment, self-concepts, and sex-role identity in Greene, B.T., Herek, G.M. (eds.) *Lesbian and gay psychology: Theory. research, and clinical applications.* 156-175.

Patterson, C. J. (1996) Lesbian mothers and their children: findings from the Bay Area Families Study in J. Laird and R.J. Green (ed.) Lesbians a*nd gays in couples and families: A handbook for therapists* (pp. 420-436). New York: Jossey-Bass.

Patterson, C.J. (1997). Children of lesbian and gay parents in T.H. Ollendick and R. J. Prinz, *Advances in clinical child psychology* 19 (pp. 235-282). New York: Plenum Press.

Pennington, S. B. (1987). Children of lesbian mothers. In F.W. Bozett (ed.), *Gay and lesbian parents* (pp. 58-74). New York: Praeger.

Rand, C., Graham, D.L.R., and Rawlings, E.I. (1982). Psychological health and factors the court seeks to control in lesbian mother custody trials. *Journal of Homosexuality* 8, 27-39.

Riddle, D.I., and Arguelles, M. (1989). Children of gay parents: Homophobia's victims. In I. Stuart and L. Abt (eds.) *Children of separation and divorce*. New York: Van Nostrand Reinhold.

Ross, J., 1988, Challenging boundaries: an adolescent in a homosexual family. *Journal of Family Psychology*, 2(2), 227-240.

Tasker, F., and Golombok, 5. (1995). Adults raised as children in lesbian families. *American Journal of Orthopsychiatry* 65(2), 203-215.

Tasker, F. and Golombok, S. (1997). *Growing up in a lesbian family: effects on child development*. New York: Guilford Press.

Turner, P.H., Scadden, L., and Harris, M.B. (1990). Parenting in gay and lesbian families. *Journal of Gay and Lesbian Psychotherapy* 1(3), 55-66.

Weeks, R.B., Derdeyn, A.P. and Langman, M. 1975, Two cases of children of homosexuals. *Child Psychiatry and Human Development*, 6(1):26-32.

West, R. and Turner, L.H. (1995). Communication in lesbian and gay families. T. J. Socha and G. H. Stamp (ed.), *Parents, children and communication: frontiers of theory and research.* Mehwah, NJ: Lawrence Erlbaum.

Wyers, N.L. (1987). Homosexuality in the family: Lesbian and gay spouses. *Social Work*, 32(2), 143-148.

B. Other References

Agresti, A., and Findlay, B. (1996). *Statistical methods for the social sciences*, third edition. San Francisco: Dellen Publishing.

American Psychological Association (1995). *Lesbian and Gay Parenting: A Resource for Pyschologists.* Washington, D.C.: American Psychological Association.

Amato, P.R., and Booth, A. (1997). *A generation at risk: Growing up in an era of family upheaval.* Cambridge, MA: Harvard University Press.

Angell, M. (1996). *Science on Trial: The Clash of Medical Evidence and the Law in the Breast Implant Case.* NY: W.W. Norton and Company.

Bell, A.P., and Weinberg, M.S. (1978). *Homosexualities: A study of diversity among men and women.* New York: Simon and Schuster.

Bell, A.P., Weinberg, M.S., and Hammersmith, S.K. (1981). *Sexual preference: Its development in men and women.* Bloomington: Indiana University Press.

Blumenfeld, M.J. and Raymond, D. (1988). *Looking at gay and lesbian life.* Boston: Beacon.

Bozett, F.W. (1987). Children of gay fathers. In F.W. Bozett (ed.), *Gay and lesbian parents* (pp. 39-57). New York: Praeger.

Bureau of the Census, U.S. Department of Commerce. (1996) *Statistical abstract of the United States, 1996: The national data book.* Washington, DC: U.S. Government Printing Office.

Campbell, D.T. and Stanley, J.C. (1966) *Experimental and quasi-experimental designs for research.* Chicago, Il: Rand McNally.

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences, second edition.* Hillsdale, NJ: Erlbaum.

Cook, T., and Campbell, D.T. (1979). Q*uasi-experimentation, design and analysis issues for field settings.* New York: Houghton-Mifflin.

Davis, J.A. (1985). *The logic of causal order.* Beverly Hills, CA: Sage.

Editors of the Harvard Law Review (1990). *Sexual orientation and the law.* Cambridge, MA: Harvard University Press.

Falk, P.J. (1989). *Lesbian mothers: psychosocial assumptions in family law.* American Psychologist 44, 941-947.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions, second edition.* New York: John Wiley and Sons.

Foster, K.R., and Huber, P.W. (1997)*Judging Science: Scientific Knowledge and the Federal Courts* Cambridge, MA: The MIT Press.

Gottman, J.S. (1990). *Children of gay and lesbian parents.* In F.W. Bozette and M.B. Sussman (eds.), *Homosexuality and family relations* (177-196). New York:

Harrington Park Press.

Greeley, A. (1994). *Marital infidelity.* Society 31(4), 9-13.

Green, R.G., Kolevzon, M.S. and Vosler, N.R. (1985). The Beavers-Timerline model of family competence and the circumplex model of family adaptability and cohesion: separate, but equal? *Family Process*, 24, 385-398.

Harris, J.R. (1998). *The nurture assumption: why children turn out the way they do*. New York: Free Press.

Hirschi, T., and Selvin, H.C. (1973). *Principles of survey analysis.* New York: Free Press.

Hudson, W.W. (1992). *The Walmyr Assessment Scales, Scoring Manual.* Tempe, AZ: Walmyr.

Kalton, G. (1983). *Introduction to survey sampling.* Beverly Hills, CA: Sage.

Kinsey, A.C., Pomeroy, W.B., and Martin, C.E. (1948). Sexual behavior in the human male. Philadelphia: Saunders.

Kish, L. 1995. *Survey Sampling*. New York: John Wiley & Sons.

Laumann, E.O., Gagnon, J.H., Michael, R.T., and Michaels, S. (1994). *The social organization of sexuality: sexual practices in the United States.* Chicago: The University of Chicago Press.

McLanahan, S. and Sandefur, G. (1994). *Growing up with a single parent.* Cambridge, MA: Harvard University Press.

Miller, D. (1991). *Handbook of research design and social measurement, fifth edition*. Newbury Park: Sage.

Nachmias, C.F., and Nachmias, D. 1996. *Research methods in the social sciences.* New York: St. Martin's Press.

National Opinion Research Center at the University of Chicago, *General Social Surveys, 1972-1996: Cumulative Codebook*, 1996. Chicago: National Opinion Research Center.

Patterson, C.J., 1992. *Children of lesbian and gay parents.* Child Development 63, 1025-1042.

Patterson, C. J., and Redding, R.. 1996. *Lesbian and gay families with children: implications of social science research for policy.* Journal of Social Issues 52, 29-50.

Pedhazur, E.J. (1982) *Multiple regression in behavioral research: Explanation and prediction*. New York: Holt, Rinehart and Winston.

Popenoe, D. (1998). *Life without father: Compelling new evidence that fatherhood and marriage are indispensable for the good of children and society*. Cambridge, MA: Harvard University Press.

Rosenberg, M. (1968). *The logic of survey analysis.* New York: Basic.

Rossi, P., and Freeman, H.E. (1994). *Evaluation: a systematic approach, fifth edition*. Newbury Park, CA: Sage.

Satinover, Jeffrey. *The Biology of Homosexuality: Science or Politics?* (1999) Pp. 3-61 in Homosexuality and American Public Life. Christopher Wolf (ed.) Dallas: Spence Publishing Company.

Saghir, M.T., and Robins, E. (1973). *Male and female homosexuality: A comprehensive investigation*. Baltimore: Williams and Wilkins.

Scheaffer, R.L., Mendenhall, W., and Ott, R.L. (1996). *Elementary survey sampling, fifth edition*. Belmont, CA: Wadsworth.

Schulenberg, J. (1985). *Gay parenting: A complete guide for gay men and lesbians with children*. New York: Anchor.

Spanier, G.B. (1976). Measuring dyadic adjustment: new scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family*, 38, 15-28.

Stacey, J. (1999). Virtual truth with a vengeance. *Contemporary sociology: A journal of reviews* 28(1):18-23.

Waite, L.J. (1995). Does marriage matter? *Demography* 32(4), 483-507.

Whitehead, B.D., (1998). *The Divorce Culture: Rethinking our commitments to marriage and family.* New York: Vintage Books.

Zill, N., and Nord, C.W. 1994. *Running in Place: How American Families Are Faring In a Changing Economy and an Individualistic Society*. Washington, DC: Child Trends Inc.

Appendix 2
Evaluation of the Studies

A. By Catagory

If the answer is no, reject the study.

Study	Quant. Study	Hypothesis: Null, Affirm Neither	Heterosexual Comparison Group?	Control for Extraneous Variables?	Measures Reliable?
Bailey et al., 1995	Yes	Neither; Dubious	No	No	Yes
Barret and Robinson, 1990	No	Neither; Dubious	No	No	No
Bigner and Jacobsen, 1989a	Yes	Neither; Dubious	Yes	Yes	Yes
Bigner and Jacobsen, 1989b	Yes	Neither; Dubious	Yes	Yes	Yes
Bigner and Jacobsen, 1992	Yes	Neither; Dubious	Yes	Yes	Yes
Bozett, F. 1980	Yes	Neither; Dubious	No	No	No
Brewaeys et al., 1997	Yes	Neither; Dubious	Yes	Yes	Yes
Cameron and Cameron 1996	Yes	Neither; Dubious	Yes	No	No
Chan et al., 1998	Yes	Null; Dubious	Yes	Yes	Yes
Crosbie-Burnett and Helmbrechty, 1993	Yes	Null; Dubious	No	No	Yes
Flaks et al., 1995	Yes	Null; Dubious	Yes	Yes	Yes
Gartrell et al., 1996	Yes	Neither; Dubious	No	No	No
Golombok and Tasker 1996	Yes	Neither; Dubious	Yes	Yes	Yes
Golombok et al., 1983	Yes	Neither; Dubious	Yes	Yes	Yes
Green R. 1978	Yes	Neither; Dubious	No	No	Yes
Green R. 1982	Yes	Null; Dubious	Yes	Yes	No

Study	Quant. Study	Hypothesis: Null, Affirm Neither	Heterosexual Comparison Group?	Control for Extraneous Variables?	Measures Reliable?
Green et al., 1986	Yes	Null; Dubious	Yes	Yes	No
Hare, J. 1994	Yes	Neither; Dubious	No	No	No
Harris and Turner, 1986	Yes	Null; Dubious	Yes	Yes	No
Hoeffer, B. 1981	Yes	Neither; Dubious	Yes	Yes	Yes
Huggins, S.L., 1989	Yes	Null; Dubious	Yes	Yes	Yes
Javaid, G.A., 1993	Yes	Null; Dubious	Yes	Yes	No
Kirkpatrick et al., 1981	Yes	Null; Dubious	Yes	Yes	Yes
Koepke et al., 1992	Yes	Neither; Dubious	No	Yes	Yes
Kweskin and Cook, 1982	Yes	Neither; Dubious	Yes	Yes	Yes
Lewin and Lyons 1982	Yes	Neither; Dubious	Yes	Yes	No
Lewis, K.G., 1980	Yes	Neither; Dubious	No	No	No
Lott-Whitehead and Tully, 1992	Yes	Neither; Dubious	No	No	Yes
Lyons, T.A., 1983	Yes	Neither; Dubious	Yes	Yes	No
McCandish, B., 1987	No	Neither; Dubious	No	No	No
McNeill et al., 1998	Yes	Null; Dubious	Yes	Yes	Yes
Miller B., 1979	Yes	Null; Dubious	No	No	No
Miller et al., 1982	Yes	Null; Dubious	Yes	Yes	Yes
Mucklow and Phelan, 1979	Yes	Neither; Dubious	Yes	No	Yes
O'Connell, A., 1993	Yes	Neither; Dubious	No	No	No
Pagelow, M.D., 1980	Yes	Affirmative; OK	Yes	Yes	No
Patterson, C.J. 1994a	Yes	Null; Dubious	Yes	No	Yes
Patterson, C.J. 1996	Yes	Null; Dubious	Yes	No	Yes
Patterson, C.J. 1997	Yes	Null; Dubious	Yes	Yes	Yes
Pennington, S.B. 1987	Yes	Neither; Dubious	No	No	No

Study	Quant. Study	Hypothesis: Null, Affirm Neither	Heterosexual Comparison Group?	Control for Extraneous Variables?	Measures Reliable?
Rand et al., 1982	Yes	Neither; Dubious	No	No	Yes
Riddle and Arguelles, 1989	Yes	Neither; Dubious	No	No	No
Ross, J., 1988	No	Neither; Dubious	No	No	No
Tasker and Golombok, 1995	Yes	Neither; Dubious	Yes	Yes	Yes
Tasker and Golombok, 1997	Yes	Neither; Dubious	Yes	Yes	Yes
Turner et al., 1990	Yes	Null; Dubious	No	Yes	No
Weeks et al., 1975	No	Neither; Dubious	No	No	No
West and Turner, 1995	Yes	Neither; Dubious	No	No	No
Wyers, N.L., 1987	Yes	Neither; Dubious	No	No	No

Study	Self-Constructed Measures?	NPS or PS Sample	Inferential Statistics?	Adequate Sample Size?	Adequate Power?
Bailey et al., 1995	Used by Others	NPS Dubious	Yes	No	No
Barret and Robinson, 1990	Self-Constructed; Dubious	NA	No	No	No
Bigner and Jacobsen, 1989a	Self-Constructed; Dubious	NPS; Dubious	Yes	No	No
Bigner and Jacobsen, 1989b	Self-Constructed; Dubious	NPS; Dubious	Yes	No	No
Bigner and Jacobsen, 1992	Self-Constructed; Dubious	NPS; Dubious	Yes	No	No
Bozett, F. 1980	Self-Constructed; Dubious	NPS; Dubious	No	No	No
Brewaeys et al., 1997	Self-Constructed; Dubious	NPS; Dubious	Yes	No	No
Cameron and Cameron, 1996	Self-Constructed; Dubious	PS; OK	Yes	No	No
Chan et al., 1998	Used by Others	NPS; Dubious	Yes	No	No
Crosbie-Burnett and Helmbrechty, 1993	Self-Constructed; Dubious	NPS; Dubious	No	No	No
Flaks et al., 1995	Used by Others	NPS; Dubious	Yes	No	No
Gartrell et al., 1996	Self-Constructed; Dubious	NPS; Dubious	No	No	No

Study	Self-Constructed Measures?	NPS or PS Sample	Inferential Statistics?	Adequate Sample Size?	Adequate Power?
Golombok and Tasker 1996	Self-Constructed; Dubious	NPS Dubious	Yes	No	No
Golombok et al., 1983	Used by Others	NPS Dubious	Yes	No	No
Green R. 1978	Used by Others	NPS; Dubious	No	No	No
Green R. 1982	Used by Others	NPS; Dubious	No	No	No
Green et al., 1986	Used by Others	NPS; Dubious	Yes	No	No
Hare, J. 1994	Self-Constructed; Dubious	NPS; Dubious	No	No	No
Harris and Turner, 1986	Self-Constructed; Dubious	NPS; Dubious	Yes	No	No
Hoeffer, B. 1981	Self-Constructed; Dubious	NPS Dubious	No	No	No
Huggins, S.L., 1989	Used by Others	NPS; Dubious	Yes	No	No
Javaid, G.A., 1993	Self-Constructed; Dubious	NPS; Dubious	No	No	No
Kirkpatrick et al., 1981	Used by Others	NPS; Dubious	No	No	No
Koepke et al., 1992	Used by Others	NPS; Dubious	Yes	No	No
Kweskin and Cook, 1982	Used by Others	NPS Dubious	Yes	No	No
Lewin and Lyons 1982	Self-Constructed; Dubious	NPS Dubious	No	No	No
Lewis, K.G., 1980	Self-Constructed; Dubious	NPS Dubious	No	No	No
Lott-Whitehead and Tully, 1992	Used by Others	NPS Dubious	No	No	No
Lyons, T.A., 1983	Self-Constructed; Dubious	NPS Dubious	No	No	No
McCandish, B., 1987	Self-Constructed; Dubious	N.A.	No	No	No
McNeill et al., 1998	Used by Others	NPS Dubious	Yes	No	No
Miller B., 1979	Self-Constructed; Dubious	NPS Dubious	No	No	No
Miller et al., 1982	Self-Constructed; Dubious	NPS Dubious	Yes	No	No
Mucklow and Phelan, 1979	Used by Others	NPS Dubious	Yes	No	No
O'Connell, A., 1993	Self-Constructed; Dubious	NPS Dubious	No	No	No
Pagelow, M.D., 1980	Self-Constructed; Dubious	NPS Dubious	No	No	No

Study	Self-Constructed Measures?	NPS or PS Sample	Inferential Statistics?	Adequate Sample Size?	Adequate Power?
Patterson, C.J. 1994a	Used by Others; Dubious	NPS	Yes	No	No
Patterson, C.J. 1996	Used by Others	NPS Dubious	Yes	No	No
Patterson, C.J. 1997	Used by Others	NPS; Dubious	Yes	No	No
Pennington, S.B. 1987	Self-Constructed; Dubious	NPS; Dubious	No	No	No
Rand et al., 1982	Used by Others	NPS; Dubious	No	No	No
Riddle and Arguelles, 1989	Self-Constructed; Dubious	NPS; Dubious	Yes	No	No
Ross, J., 1988	Self-Constructed; Dubious	N.A.	No	No	No
Tasker and Golombok, 1995	Used by Others	NPS; Dubious	Yes	No	No
Tasker and Golombok, 1997	Used by Others	NPS; Dubious	Yes	No	No
Turner et al., 1990	Self-Constructed; Dubious	NPS; Dubious	Yes	No	No
Weeks et al., 1975	Self-Constructed; Dubious	N.A.	No	No	No
West and Turner, 1995	Self-Constructed; Dubious	NPS; Dubious	No	No	No
Wyers, N.L., 1987	Self-Constructed; Dubious	NPS Dubious	No	No	No

B. Conclusions

Bailey, J.M., Bobrow, D., Wolfe, M., and Mikach, S. 1995. Reject

Barret, R. L., and Robinson, B.E. 1990. Reject

Bigner, J.J., and Jacobsen, R.B. 1989a. . Reject

Bigner, J.J., and Jacobsen, R.B. 1989b. Reject

Bigner, J.J. and Jacobsen, R.B. 1992. Reject

Bozett, F. 1980. Reject

Brewaeys, A., I. Ponjaert, E.V. Van Hail, and S. Golombok, 1997. Reject

Cameron, P. and Cameron, K. 1996. Reject

Chan, R.W., Raboy, B., and Patterson, C.J. 1998. Reject

Crosbie-Burnett, M., and Helmbrechty, L. 1993. Reject

Flaks, D.K., Ficher, I., Masterpasqua, F. and Joseph, G. 1995. Reject

Gartrell, N., Hamilton, J., Banks, A., Mosbacher, D., Reed, N., Sparks,

C.H., and Bishop, H. 1996. Reject

Golombok, S. and Tasker, F. 1996. Reject

Golombok, S., Spencer, A., and Rutter, M. 1983. Reject

Green, Richard, 1978. Reject

Green, R. 1982. Reject

Green, R., Mandell, J.B., Hotvedt, M.E., Gray, J., and Sarnith, L. 1986. Reject

Hare, J. 1994. Reject

Harris, M., and Turner, P. 1986. Reject

Hoeffer, B. 1981. Reject

Huggins, S.L. 1989. Reject

Javaid, G.A. 1993. Reject

Kirkpatrick, M., Smith, C., and Roy, R. 1981. Reject

Koepke, L., Hare, J., and Moran, P.B. 1992. Reject

Kweskin, S.L., and Cook, A.S. 1982. Reject

Lewin, Ellen, and Terrie A. Lyons, 1982. Reject

Lewis, Karen Gail, 1980, Reject

Lott-Whitehead, L., and Tully, C. 1992. Reject

Lyons, Terry A., 1983, Reject

McCandish, B. 1987. Reject

McNeill, Kevin F., Beth M. Rienzi, and Augustine Kposowa, 1998. Reject

Miller, B. 1979. Reject

Miller, J.A., Jacobsen, R.B., and Bigner, J.J. 1982. Reject

Mucklow, B.M., and Phelan, G.K. 1979. Reject

O'Connell, A., 1993. Reject

Pagelow, M.D., 1980. Reject

Patterson, C.J. 1994a. Reject

Patterson, C. J. 1996. Reject

Patterson, C.J. 1997. Reject

Pennington, S. B. 1987. Reject

Rand, C., Graham, D.L.R., and Rawlings, E.I. 1982. Reject

Riddle, D.I., and Arguelles, M. 1989. Reject

Ross, J., 1988. Reject

Tasker, F., and Golombok, 5. 1995. Reject

Tasker, F. and Golombok, S. 1997. Reject

Turner, P.H., Scadden, L., and Harris, M.B. 1990. Reject

Weeks, R.B., Derdeyn, A.P. and Langman, M, 1975. Reject

West, R. and Turner, L.H. 1995. Reject

Wyers, N.L. 1987. Reject

Appendix 3
Same-Sex Parenting Studies and the Law

By William C. Duncan
Assistant Director, Marriage Law Project

As the Lerners have noted, studies of homosexual parenting have been published and put to use in a variety of highly contentious settings. One of these settings is the legal profession. In this Appendix we demonstrate how these unreliable studies have made their way into judicial decisions, expert witness testimony by study authors in cases, briefs filed by advocacy groups, and law review articles, including some of the most prestigious law journals in the country. With only one exception, these studies have been treated with a complete absence of criticism.

Citation of Studies in Judicial Opinions

The most commonly cited study is *Children of Lesbian and Gay Parents* by **Charlotte Patterson**.[1] In an important 1992 case where a biological mother's same-sex partner was allowed to adopt her child without terminating the mother's parental rights,[2] this study was cited in a footnote for the proposition that "[c]oncern that a child would be disadvantaged by growing up in a single-sex household is not borne out by the professional literature examined by this Court."[3] The next year, a New Jersey court quoted the study to establish that current research, though limited, indicates that children raised by homosexual parents will turn out "normal."[4] In that case, the court held that a same-sex couple could jointly adopt the child of one of the partners. Later in 1993, the Patterson study again surfaced in an opinion, but this time it was merely mentioned without

Notes for this section begin on Page 139

substantial comment, because the court was ruling on the constitutionality of Florida's adoption law, which excludes homosexuals as prospective adoptive parents.[5] In another New York case granting two petitions to adopt by same-sex partners of a child's biological mother, the Patterson study was again cited as evidence that children raised by homosexual parents do not experience significant differences from children raised by heterosexual parents.[6] The court discussed the Patterson study at length, and summarized its finding that concerns about the development of children raised by homosexual parents were unfounded.[7] The court felt that this study was persuasive enough to put to rest developmental concerns about the children involved in the case.

Another study mentioned in the cases was conducted by **Green**, et. al.: *Lesbian Mothers and their Children: A Comparison with Solo Parent Heterosexual Mothers and their Children*.[8] This study is cited in two cases. In the first instance, which involved the New Hampshire Supreme Court's holding that excluding homosexuals from adoption and foster parenting would be constitutional, the dissent referred to the study to support the argument that "apparently the overwhelming weight of professional study on the subject concludes that no difference in psychological and psychosexual development can be discerned between children raised by heterosexual parents and children raised by homosexual parents."[9] The **Green** study was also cited in a Pennsylvania case, where it informed the view of the majority that a lesbian mother should have child visitation in the presence of her partner. This time, the study was cited in a footnote, for the proposition that "a variety of psychological studies indicate that lesbianism does not correlate negatively with the ability to raise a healthy, normal child."[10] Another study by **Green**, *Sexual Identity of 37 Children Raised by Homosexual or Transsexual Parents*,[11] was cited by a dissenting judge for a definition of transsexuality.[12] In another case, Dr. **Green** was described as "an expert on gender identity in children" and was quoted as saying "[n]o theory in the developmental psychology literature suggests that having homosexual parents leads to a homosexual outcome."[13] Interestingly, this same study was rejected as not credible in a Missouri case that same year, deferring to the trial court's judgment as to credibility and because the court's

concerns with the "moral growth and best interests" of the children involved was paramount.[14] Other studies have also been noted in judicial opinions. One by **Mildred Pagelow**[15] was cited by a witness supporting a mother whose open homosexuality was found to be a sufficient change in circumstance to warrant a change in custody of her child from the mother to the father.[16] The court quoted from the study as follows: "Dr. W., in support of her opinion, testified to a study involving homosexual and heterosexual mothers which found 'essentially no difference in the development of the children or the relationship between mothers and their children or generally the problems that the mothers were having in raising their children.'"[17] The court, however, concluded that the potential negative impact on the child by the conflict between societal and moral norms and the mother's lifestyle supported the decision to modify custody.

Expert Witness Testimony by Study Authors in Cases

A second way that studies on homosexuality and parenting can affect outcomes of cases occurs when the authors of the studies testify in individual cases. It is hard to gauge how often this happens, because the court's final published opinion in a case may or may not mention particular witnesses.

One author who has offered testimony is Dr. **Richard Green**, who has published three studies on homosexual parenting.[18] In a New Jersey case in which a homosexual father was granted visitation rights (though visitation was restricted to daylight hours), Dr. **Green** contradicted another expert (Dr. Richard Gardner) who argued that exposure to the father's lifestyle could be detrimental to the children involved in the case.[19] Dr. **Charlotte Patterson** has also had extensive experience testifying in cases. For instance, in a Virginia case that held that a maternal grandmother should be given custody instead of the lesbian mother of a child, Dr. Patterson testified as an expert witness.[20] Dr. **Patterson** testified on behalf of the guardian ad litem[21] regarding her study on child development of children with homosexual parents.[22] According to a commentator, Dr. **Patterson** testified that:

1) children of lesbian parents are as secure in their gender identity as are children of heterosexual parents;
2) children of lesbian mothers are equally likely to exhibit the expected gender role behavior as children of heterosexual parents;
3) there is no difference in sexual orientation between children of lesbian and gay parents and children of heterosexual parents;
4) there is no difference in personal development between children of lesbian parents and children of heterosexual parents; and
5) there are no differences in relationships with peers between children of lesbian parents and children of heterosexual parents.
Finally, Dr. Patterson pointed out that there is not a single study that would suggest that children of lesbian parents would be any less well-off developmentally than children of heterosexual parents.[23]

Dr. **Patterson** was also asked about the possibility that children of homosexual parents will be teased. Dr. **Patterson** responded that every child is teased, but the issue is whether the child has a good parent to teach him or her how to deal with it.[24] In a 1997 case challenging Florida's adoption statute, which prohibits adoption by "homosexual" persons, Dr. **Patterson** testified on behalf of the plaintiffs. In a note in the final opinion issued by the circuit court judge dismissing the case, the judge described Dr. **Patterson**'s evidence as "questionable."[25] He noted that she had testified that the research literature on the issue of same-sex and homosexual parenting was "substantial," but took issue with this conclusion, given the fact that no studies addressed parenting by homosexual men.[26] The judge further noted: Dr. **Patterson**'s ability also came into question when prior to trial she refused to turn over to her own attorneys copies of documentation utilized by her in her studies. The court then ordered her to do so (both sides having stipulated to the Order), yet she unilaterally refused despite the continued efforts on the part of her attorneys to have her do so. Both sides stipulated that Dr. **Patterson**'s conduct was a clear violation of this Court's order. Her attorneys requested that sanctions be limited to the exclusion of her personal studies at trial and the court agreed to do so.[27]

Dr. **Patterson** also testified in the important Hawaii same-sex "marriage" case in 1996.[28] In the face of a Hawaii Supreme Court finding that the existing law constituted sex discrimination under the Hawaii Constitution, the Hawaii trial court held the state of Hawaii did not show a compelling state interest for limiting marriage to opposite sex couples.[29] In its opinion, the trial court summarized Dr. **Patterson**'s testimony. The court noted that Dr. **Patterson** testified regarding the findings of her Bay Area Family Study and concluded from that study "that the particular group of children [of homosexual parents], when compared to available norms, appeared to be developing in a normal fashion."[30] The court also noted, however, that Dr. **Patterson** testified that the children of the lesbian mothers sampled reportedly expressed feeling more symptoms of stress in their lives and that she admitted that the sample studied was not representative of the nation.[31] Dr. **Patterson** also testified regarding her Contemporary Family Study, noting her findings that "sexual orientation of the parents was not a good predictor of how well children do in terms of a child's well-being and adjustment" and that "irrespective of their parents' sexual orientation, children who live in a harmonious family environment had better reports from parents and teachers."[32] Finally, based on her general expertise, Dr. **Patterson** concluded that:

> 1) "A biological relationship between parent and child is not essential to raising a healthy child. The quality of parenting that a child receives is more important than a biological connection or the gender of a parent." (2) "[T] here is no data or research which establishes that gay fathers and lesbian mothers are less capable of being good parents than non-gay people and which supports denying gay people the ability to adopt and raise children." 3) "[G] ay and lesbian people and same-sex couples are as fit and loving parents as non-gay people and different-sex couples. Further, sexual orientation is not an indicator of parental fitness." 4) "[S] ame-sex couples can, and do, have successful, loving and committed relationships." 5) In her opinion, "[t] here is no reason related to the promotion of the development of children why same-sex couples should not be permitted to marry."[33]

In the Hawaii case, the state did call a witness to rebut the asse tions based on the social science studies of homosexual parenting,

Dr. Richard Williams.[34] Dr. Williams had reviewed a number of these studies and critiqued nine in particular. He testified that the studies' general flaws included "non-representative sampling of heterosexual, gay and lesbian parents," "inadequate sample size" and "comparison groups [which] were not comparable in terms of household make up."[35] Dr. Williams also noted specific flaws in individual studies.[36] However, the trial court judge, Judge Chang, dismissed Dr. Williams' testimony out-of-hand. Judge Chang's basis for discounting his testimony did not rely on the substance of Dr. Williams' testimony, but on what the judge called Dr. Williams "expressed bias against the social sciences," his belief "that there is no scientific proof that evolution occurred" and the fact that Dr. Williams' testimony reflected a "minority position" on disputed issues. [37]

Citations to Studies in Legal Briefs

Like testimony offered in court, legal briefs submitted to courts can have great influence on the final outcome of a case. While it is hard to gauge the extent of that effect (since the final opinion of the court rarely discusses the relative persuasiveness of legal briefs), it is clear from available information that briefs submitted on behalf of social scientists or social science organizations, and briefs which use information from social science research and writing, are routinely used in cases.

For example, a Westlaw search of reported state cases since 1944 found five cases of child custody or visitation disputes involving homosexual parents where the National Association of Social Workers and/or one of its state affiliates filed an amicus brief.[38] Likewise, the American Psychological Association or a State Psychological Association also filed briefs in four of such cases.[39] The National Association of Social Workers also filed a brief in a case involving an adoption by the same-sex partner of a child's biological mother[40] and a case challenging the State of Florida's policy prohibiting adoption by homosexuals (in which a brief was also filed by the Florida Psychological Association).[41] In addition to these findings, the briefs filed in recent same-sex "marriage" cases in Hawaii and Vermont are particularly instructive.

A. Hawaii Same-Sex "Marriage" Case

In the Hawaii same-sex "marriage" case, many briefs were filed by social scientists in support of the couples seeking marriage licenses. Here we will mention only the most recent ones. The first brief was filed by five sociologists who worked on issues of marriage and the family.[42] This brief approved the trial court's finding that the social science studies "consistently show that children raised by gay and lesbian parents and same-sex couples develop in a normal fashion and are not detrimentally affected by their parents sexual orientation."[43]

The second brief, filed on behalf of 11 "scholars and researchers of child and family issues," including **Susan Golombok, Richard Green, Martha Kirkpatrick, Lawrence Kurdek and Fiona Tasker** (all authors of social science studies on homosexual parenting) discussed the studies in more depth.[44] This brief argued that the studies on homosexual parenting establish that there is equivalency in the psychological development of children raised by homosexual parents and those raised by heterosexual parents.[45] The brief also asserted that the studies indicate that homosexuals can be good, capable parents.[46] On the other hand, a brief filed by the National Association for Research and Therapy of Homosexuality (NARTH) found the studies relied on by these briefs to be unpersuasive.[47] The NARTH brief noted that the evidence was "far from being 'clear' and 'conclusive'" and that merely because the studies cited had not found harm did not mean that there is no harm, to children raised by homosexual parents or same-sex couples, or that such harm could not be shown later.[48]

B. Vermont Same-Sex "Marriage" Case

In the Vermont same-sex "marriage" case in which three couples challenged the state's denial of marriage licenses to same-sex couples,[49] the plaintiffs submitted a brief to the Vermont Supreme Court which discussed (among other things) the social science research on homosexual parenting.[50] Under the title, "Social Science Research Belies the Assumption that Children are Better Off in a Home with a Father and a Mother," the brief states: "[T]he unspo-

ken but implicit assumption that children are better off when raised by a father and a mother as opposed to two fathers or two mothers is unsupported by contemporary social science research."[51] To support this proposition, the brief quotes a report of the American Psychological Association: "Not a single study has found children of gay or lesbian parents to be disadvantaged in any significant respect relative to children of heterosexual parents. Indeed, the evidence to date suggests that home environments provided by gay and lesbian parents are as likely as those provided by heterosexual parents to support and enable children's psychosocial growth."[52]

The brief also argues "researchers have debunked the myth that children raised by gay or lesbian parents have unhealthy or even atypical gender identities or sexual orientations,"[53] and that "studies have consistently confirmed that the parenting skills of gay and lesbian parents are essentially the same as those of their heterosexual counterparts."[54] In support of the couples seeking a marriage license, the Vermont Psychiatric Association, the Vermont Psychological Association, the Vermont chapter of the National Association of Social Workers, a private practice mental health group from Burlington, Vermont and three University of Vermont psychologists joined in filing a brief that discussed social science evidence regarding homosexual parenting.[55] In their brief, these organizations took the position that "social science research supports the conclusion that legal recognition of marriages between partners of the same sex would benefit the children of gay and lesbian parents."[56]

This conclusion was based on a three arguments: (1) Same-sex couples are raising children already and doing well at it; (2) Recognition of same-sex "marriage" would "create a legal relationship between the children's parents;" (3) Recognizing same-sex "marriage" would end "state-sanctioned stigmatization" of children raised by same-sex parents.[57] The social science studies of homosexual parenting were prominent in providing support for the first argument. In discussing this argument, the brief asserted that "researchers who have studied the parenting skills and values of gay and lesbian parents have found them to be essentially the same as those of their heterosexual counterparts."[58] While noting that small sample sizes and non-random samples affect the validity of individual stud-

ies, the brief argues that taken together, "the collective weight of the research unquestionably reinforces the conclusion that the children of gay or lesbian, or same-sex parents are as happy, healthy and well adjusted as their counterparts with heterosexual or different-sex parents."[59] To support this statement, the brief reviewed study findings regarding the psychological development, social development, moral development, intelligence, adult relations, gender identity, and sexual orientation of children raised by homosexual parents, and concluded that "in their capacity as parents . . . committed gay and lesbian couples are functionally equivalent to their heterosexual counterparts"[60]

A brief was filed in support of the existing marriage law, which included a discussion of homosexual parenting. Authored by the Massachusetts Family Institute and the National Association for the Research and Therapy of Homosexuality (NARTH), the brief aimed to "alert the Court to flaws" in the studies relied on by the couples seeking marriage licenses.[61]

First this brief argued that although their opponents briefs claim that social science research has shown no developmental difference between children raised by homosexual parents and children raised by heterosexual parents, some of the studies cited in the briefs actually contradict the claims of those briefs.[62]

The brief then turned to substantive defects in the social science studies relied on by plaintiffs and their supporters. The brief discussed four major problems with the studies: (1) defective research design, (2) defective sampling techniques, (3) erroneous data analysis and interpretation, and (4) indications in the studies that there are significant differences between same-sex and opposite-sex parenting.[63] In regards to research design, the brief mentioned that some of the studies did not have a systematic design, that others did not incorporate appropriate comparison groups in their design, and others did not include adequate control groups of heterosexual parents.[64] Then, the brief discussed a number of individual studies with these types of flaws.[65] The brief's discussion of sampling techniques critiqued a number of studies for such flaws as small sample-size, non-representative samples, non-random samples, or inappropriate samples for the research questions involved in the study.[66] The sec-

tion discussing data analysis and interpretation identified flaws including outright inaccuracy in reporting of the findings, "fail[ure] to incorporate any inferential statistical testing of hypotheses, and engag[ing] in erroneous scientific reasoning by making the illegitimate claim of affirming the null hypothesis and/or making illegitimate generalizations of conclusions from the study's data."[67] Finally, the brief asserted that some studies may actually indicate significant differences between homosexual and heterosexual parenting (despite their own claims), such as the likelihood that children of homosexual parents may become involved in homosexual behavior themselves.[68] After the NARTH brief was filed, the Vermont Psychiatric Association and the other organizations joining its original brief attempted to file an additional brief attacking the NARTH brief.[69] The Vermont Supreme Court rejected the brief.

The attempted response to the NARTH brief was twofold.[70] First, the reply brief argued that most studies of any kind are non-representative, so that flaw should not invalidate the studies relied on by the plaintiffs and their supporters.[71] Second, the reply brief argued that by combining the findings of all of the studies, regardless of their flaws, an accurate picture of the effects on children of homosexual parenting was still possible and that "the cumulative consistency in the results of the individual studies, as well as the collective weight of the studies as assessed in the meta-analysis, provide the most accurate information to date on this issue, and indicate that children are not psychologically harmed by having gay or lesbian parents."[72]

Use of Studies in Law Reviews

A more indirect way these social science studies may affect the state of the law occurs as they are introduced into the legal discourse through scholarly journals.

Typical Articles

The most common use of social science studies in law journal articles is in footnotes, as evidence of assertions of the harmlessness to children of homosexual parenting. Some articles do discuss these studies in some depth, though. One of the earliest articles to discuss

the body of social science research on the subject of homosexual parenting was published in 1990 in the *Georgetown Law Journal*.[73] That article cited a number of social science studies to back the assertions that "little difference exists in the overall mental health of children raised in lesbian-mother households and children raised in heterosexual-mother households" and "the quality of mothering, not the mother's sexual orientation, is the most crucial factor for a child's healthy growth and development."[74] The articles then sought to establish support for these arguments by discussing studies by **Kirkpatrick, Golombok, McCandish, Hotvedt, and Green** in some detail.[75] The author then made an assessment of the value of the literature for courts dealing with issues regarding homosexual parenting:

> "Because the existing psychological literature uniformly agrees that children raised by lesbians are as psychologically healthy as children raised by heterosexual parents, *courts influenced solely by this literature would have to agree that raising a child in a lesbian-mother family is not against a child's best interest*."[76]

It is important to note the implication of this assertion—that judges who do not agree with the proposition must be influenced by something else (almost assuredly something inappropriate).

Then, the author responds to a potential criticism of this remarkable assertion. The author says: "Criticisms aimed at the research methodology are misplaced. Compared with other studies on child rearing that courts and legislatures use to support a variety of custody rules, much of the lesbian-mother custody research is exemplary."[77] After this, the author argues that the only reasons judges or legislatures would disagree with the findings of these studies are "homophobia and heterosexism."[78]

In 1995, another law journal article addressed the use of social science research in adoption cases.[79] This article is even more uncritical. It states: "No study has shown any harm to children raised by lesbian or gay parents."[80] In briefly reviewing the studies, the author says that the research establishes: (1) "[L]esbian and gay parents have parenting skills that are at least equivalent to those of heterosexual parents."[81] (2) "[U]nanimously that there are no significant differences in the psychological health of [children raised by homo-

sexual parents] compared with children raised by heterosexual parents."[82] (3) "[T]he gender identity of children raised by a lesbian mother does not differ from the gender identity of children raised by a heterosexual mother."[83] (4) "[B]eing raised by a lesbian or gay parent does not increase the likelihood that a child will become lesbian or gay."[84]

The author also makes a number of other assertions regarding homosexual parenting, citing social science studies as support. The importance of these studies is clear to the author: "This new research . . . must be used to fight the existing statutory bans [on adoption by homosexual prospective parents], to fight new proposed statutory bans, and, perhaps the biggest challenge, to prevent the use of non-statutory means to discourage or disallow adoption by lesbian and gay people."[85]

Another article from 1996 argues "there is no justification for special rules or particular scrutiny for lesbian and gay parents" because "extensive social science research" has shown that concerns with homosexual parents are "unwarranted."[86] To bolster this contention, the author examines some of the studies. According to the author these studies "emphatically demonstrate[] that there is no basis for any generalized concern about harm to children from being raised by lesbian and gay parents."[87] Then, the author specifically talks about studies regarding gender roles, children's sexual orientation, self-esteem and the effects of stigmatization on children, arguing that in each case, the social science evidence indicates that children of homosexual parents have no significant problems in these areas.[88]

Law Review Articles by Social Scientists

In the past few years, legal journals have also provided a forum for two authors of same-sex parenting studies, **David Flaks** and **Charlotte Patterson**. Given the fact that the legal profession is often highly dependent on professionals from other fields, these types of articles can be very significant in two ways. First, they make the arguments of these studies more readily available to the legal profession. Second, they allow the authors to advertise their work to the legal profession. Dr. **Flaks**' article summarizes the social science re-

search on homosexuality, homosexual parenting, same sex relationships, and community stigma.[89] Dr. **Flaks** then discusses studies regarding psychological health of children of homosexual parents, their potential for being molested or contracting AIDS and the potential effects of social stigma on the children and concludes, "The social science literature reviewed here demonstrates clearly that lesbians and gay men can and do raise psychologically healthy children. In fact, no evidence has emerged to date that homosexual parents are inferior to their heterosexual counterparts, or that their children are in any way compromised."[90] To Dr. **Flaks**, the research provides "sufficient evidence to support judicial reassessment of the disparate treatment sometimes afforded lesbian and gay families within the American legal system."[91]

Dr. **Patterson**'s article advances the argument that "children of lesbian mothers have been found to develop normally."[92] In her article, Dr. **Patterson** discusses research on homosexual parenting as it relates to gender role development of children, psychological disorders, child abuse and problems with social interaction.[93] She also describes her own study of 37 lesbian mother families with pre-school age children, and concludes that it showed no significant negative differences between children raised by lesbian mothers and children raised by heterosexual parents, except that the children of lesbian mothers reported more symptoms of stress.[94] She argues that the stress finding may merely indicate that the children of lesbian mothers are more used to talking about their feelings.[95] Dr. **Patterson** also writes that future research should focus on the strengths of "lesbian families" which she thinks might include an increased appreciation for diversity, expanded views of gender roles, a strong feeling of being wanted, and the opportunity to observe "a model of justice, especially in terms of the division of labor at home."[96] Dr. **Patterson** concludes, "in decisions regarding the adoption of minor children, parental sexual orientation should be considered irrelevant."[97]

Professor Wardle

The only law journal article to significantly criticize the same-sex parenting social science research was published in the *University of Illinois Law Review* by Professor Lynn D. Wardle.[98] Professor Wardle summarizes his review of the research as follows:

> Most of this new literature is supportive, much of it self-affirming, of homosexual parenting. Advocates of legalized homosexual parenting have cited this social science literature to show that homosexual parenting is just as good as heterosexual parenting or that there is no more detriment to children from homosexual parenting than from heterosexual parenting. However, on close examination, the evidence does not prove that homosexual parenting is equivalent to heterosexual parenting or that it is not harmful in significant ways to children. Most of the studies of homosexual parenting are based on very unreliable quantitative research, flawed methodologically and analytically (some of little more than anecdotal quality), and provide a very tenuous empirical basis for setting public policy. And even from this body of research there are indications of some significant potential detriments to children from homosexual parenting.[99]

Professor Wardle notes a wide range of methodological and analytical flaws in the studies, including: small sample size, self-selection of participants, using single heterosexual parents as a control group rather than male-female couples, inadequate control for other variables, bias of researchers and respondents, reliance on faulty methods of response retrieval, absence of longitudinal studies, use of unreliable criteria for comparison, inadequate data compilation, overgeneralizations from data, flawed analysis of sexual orientation of children, and misinterpretation of statistical analysis.[100]

Given these problems, Professor Wardle concludes, "Until these concerns are conclusively dispelled, it would not be rational to adopt a public policy endorsing or legitimating homosexual parenting."[101] Professor Wardle does note that even with their flaws the studies may indicate the opposite of what they are characterized as showing—that there may be potential problems for children raised by homosexual parents, including increased risk of "homosexual interests and behaviors," problems in the parent-child relationship, lower self-image regarding masculinity for boys, and others.[102] He concludes "although the social science research is not conclusive; it does suggest that there are some particular and unique potential risks to children raised by active homosexual parents."[103]

The year after Professor Wardle's article was published, the *Uni-*

versity of Illinois Law Review published a response by Professor Carlos A. Ball and Janice Farrell Pea. In their response, Professor Ball and Ms. Pea conceded that there were methodological flaws in the studies but argue that the studies themselves noted this. By taking all of the studies together in a "meta-analysis," Ball and Pea claim, an accurate picture of the effect of homosexual parenting on children could emerge.[104] In a response published in the same journal, Professor Wardle responded to the "meta-analysis" argument: "They suggest that meta-analysis based on aggregating the flawed studies validates the conclusions of the flawed studies, but they do not explain how combining 15 or 20 methodologically flawed studies produces a result that is not also methodologically impaired."[105]

Their article then reviews the studies in some detail and takes issue with some of the concerns that Professor Wardle noted were raised by the existing social science literature. Ball and Pea argue that the social problems experienced by homosexuals are produced largely by "hostility and harassment . . . from society in general and . . . peers in particular." They conclude that restrictions on homosexual parenting would not solve the problems.[106]

The authors also addressed the fact that while the studies indicated that there may be differences between children raised by homosexual and heterosexual parents the differences are not necessarily bad.[107]

The fact that Ball and Pea accept the argument that there are differences between the two groups of children is surprising, since it is extremely common for the findings of the studies to be described as showing no differences. To show that the differences are not necessarily bad, they examine some of the studies that indicate a difference and conclude that in the development of traditional gender roles, the difference doesn't indicate a problem because there is "no societal interest in perpetuating rigid gender roles."[108] In his response to Professor Ball and Ms. Pea's criticisms, Professor Wardle notes that many of the key points of his article are conceded in the response, including his charge that the existing research is badly flawed.[109]

Conclusions

There is evidence that courts are relying on the studies and their authors for information that is used to determine adoption, custody, visitation and same-sex "marriage" cases. At least 13 reported court opinions have cited to the studies or their authors and at least two briefs submitted in same-sex "marriage" cases reference the studies. Of the opinions citing the studies, six have made decisions favorable to the homosexual parent or same-sex couple, and three have yielded a result negative to the homosexual parent or couple. Twice the studies are cited in dissenting opinions, and in one case study was cited but the decision was related to a different issue. Meanwhile, discussions of the articles in academic journals are overwhelmingly accepting of the assertions made in the studies. It is likely that many more legal briefs submitted in cases rely on these studies, and that other courts are influenced by the studies or by the testimony of the studies' authors in their decisions. This makes it extremely important that the flaws in these studies be exposed so that courts can weigh their relative value when making decisions that will dramatically impact both individual children and society at large.

Notes to Appendix 3

1. *Child Development* 1025 (1992).

2. Adoption of Evan, 583 N.Y.S.2d 997 (N.Y. Cty. Surrogate Ct. 1992).

3. Id. at 1002.

4. Adoption of J.M.G., 632 A.2d 550, 554 (N.J. Super. Ct. 1993).

5. Florida Dept. Of Health and Rehabilitative Services v. Cox, 627 So.2d 1210, 1213 (Fla. Ct. App. 1993) (upholding state's adoption law).

6. Adoption of Caitlin, 622 N.Y.S.2d 835, 840 (N.Y. Fam. Ct., Monroe Cty. 1994).

7. Id. at 840-841.

8. 15 *Archives of Sexual Behavior* 167 (1986).

9. Opinion of the Justices, 530 A.2d 21, 28 (N.H. 1987) (Batchelder, J., dissenting). Justice Batchelder also cites Golombok, Spencer & Rutter, Children in Lesbian and Single-Parent Households: Psychosexual and Psychiatric Appraisal 24 J. *Child Psychology & Psychiatry* 4:551 (1983); Harris & Turner, Gay and 12 J of H 101 (1986).

10. Blew v. Verta, 617 A.2d 31, 36 n. 2 (Pa. Super. 1992). The footnote also cites M. Kirkpatrick, C. Smith, & R. Roy, Lesbian Mothers and their Children: A Comparative Survey, *American Journal of Orthopsychiatry* (July 1981).

11.135 *Am. J. Psych.* 692 (1978).

12. Daly v. Daly, 715 P.2d 56, 60 n. 1 (Nev. 1986) (Gunderson, J., dissenting)(terminating parental rights of transsexual father).

13. Conkel v. Conkel, 509 N.E.2d 983, 986-987 (Ohio Ct. App. 1987) (granting overnight visitation to homosexual father)(citing Richard Green, The Best Interests of the Child with a Lesbian Mother 10 *Bulletin of the Am. Acad. of Psychiatry and the Law* 7 (1982).
14S.E.G. v. R.A.G., 735 S.W.2d 164, 166 n. 1 (Mo. Ct. App. 1987) (awarding father custody of children in preference to homosexual mother).

15. Heterosexual and Lesbian Single Mothers: A Comparison of Problems, Coping, and Solutions 5 J. *Homosexuality* 189 (Spring 1981).

16. M.J.P. v. J.G.P., 640 P.2d 966, 968 (Ok. 1982).

17. Id.
18. See Richard Green, *Sexual Science and the Law* 18-49 (1992) (discussing Dr. Green's experience testifying in child custody cases involving homosexual parents).

19. In the Matter of J.S. & C., 324 A.2d 90, 96 (N.J. Super. 1974).

20. Bottoms v. Bottoms, 457 S.E.2d 102 (Va. 1995).

21. Barry M. Parsons, Note, Bottoms v. Bottoms: Erasing the Presumption Favoring a Natural Parent Over Third Parties—What Makes This Mother Unfit? 2 *George Mason Indep. L. Rev.* 457, 477 (1994).

22. Charlotte Patterson, Children of Lesbian and Gay Parents 63 *Child Development* 1025 (1992).

23. Parsons, Note, *George Mason Indep. L. Rev.* 457, 477 (1994).

24. See Peter Nash Swisher & Nancy Douglas Cook, Bottoms v. Bottoms: In Whose Best Interest? Analysis of a Lesbian Mother Child Custody Dispute, 34 *U. Louisville J. Fam. L.* 843, 859 (1995-1996).

25. Amer v. Johnson, Case No. 92-14370 at 14, note 18 (Broward County Circuit Ct., 1997).

26. Id.

27. Id.

28. Baehr v. Miike, 1996 WL 694235 (Haw. Cir. Ct. 1996).

29. Id.

30. Id. at *12.

31. Id. at *12-13.

32. Id. at *13.

33. Id.

34. Id. at *8.

35. Id.

36. Id.

37. Id. at *8-9.

38. Boswell v. Boswell, 721 A.2d 662 (Md. Ct. App. 1998); Pulliam v. Pulliam, 501 S.E.2d 898 (N.C. 1998); Inscoe v. Inscoe, 700 N.E.2d 70 (Ohio Ct. App. 1997); Bottoms v. Bottoms, 444 S.E.2d 276 (Va. Ct. App. 1994); Marriage of Diehl, 582 N.E.2d 281 (Ill. App. 1991).

39. Boswell v. Boswell, 721 A.2d 662 (Md. Ct. App. 1998); Inscoe v. Inscoe, 700 N.E.2d 70 (Ohio Ct. App. 1997); Hertzler v. Hertzler, 908 P.2d 946 (Wyo. 1996); Bottoms v. Bottoms, 444 S.E.2d 276 (Va. Ct. App. 1994).

40. Adoption of Baby Z., 724 A.2d 1035 (Conn. 1999).

41. Cox v. Florida Department of Health and Rehabilitative Services, 656 So.2d 902 (Fla. 1995).

42. Baehr v. Miike, Civ. No. 91-1394-05, Brief of Amici Curiae Andrew E. Cherlin, et al. (Haw. S.Ct., June 2, 1997).

43. Id. at 6.

44. Baehr v. Miike, Civ. No. 91-1394-05, Brief of Amici Curiae Urie Bronfenbrenner, et al., (Haw. S.Ct., May 23, 1997).

45. Id. at 4.

46. Id. at 5.

47. Baehr v. Miike, Civ. No. 91-1394-05, Brief of Amicus Curiae National Association for Research and Therapy of Homosexuality, Inc., (Haw. S.Ct., March 24, 1997).

48. Id. at 8.

49. Baker v. Vermont, Superior Ct. Docket No. S1009-97 Cnc., Opinion and Order (filed Dec. 17, 1997), reversal by Baker v. State, 744 A.2d864 (Vt. 1999).

50. Mary Bonauto, Susan M. Murray & Beth Robinson, Brief, The Freedom to

Marry for Same-Sex Couples: The Opening Appellate Brief of Plaintiffs Stan Baker Et. Al. v. State of Vermont, 5 *Mich. J. Gender & L.* 409 (1999).

51. Id. at 443.

52. Id. (citing Charlotte J. Patterson and Richard E. Redding *Lesbian and Gay Parenting: A Resource for Psychologists* 8 (1995), which includes a chapter by Charlotte J. Patterson).

53. Id. at 444 (citing Charlotte J. Patterson and Richard E. Redding, *Lesbian and Gay Parenting: A Resource for Psychologists* 8 (1995), which includes a chapter by Charlotte J. Patterson). *J. of Soc. Issues* 29, 41 (Fall 1996)).

54. Id. at 444.

55. Baker v. Vermont, Brief of Amici Curiae Vermont Psychiatric Association, et al. (Vt. S.Ct. March 1998).

56. Id. at 13.

57. Id. at 14-43.

58. Id. at 17.

59. Id. at 24.

60. Id. at 36.

61. Baker v. Vermont, Brief of Amici Curiae Massachusetts Family Institute and National Association for the Research and Therapy of Homosexuality (Vt. S.Ct. 1998).

62. Id. at 3.

63. Id. at 5-22.

64. Id. at 6.

65. Id. at 7-10.

66. Id. at 10-17.

67. Id. at 17.

68. Id. at 21-22.

69. Baker v. Vermont, Reply Brief of Amici Curiae Vermont Psychiatric Association, et al. (Vt. S.Ct. June 1998). Submitted, but not accepted.

70. Id. at 6-16.

71. Id. at 3.

72. Id. at 6.

73. Nancy D. Polikoff, This Child Does Have Two Mothers: Redefining Parenthood to Meet the Needs of Children in Lesbian-Mother and Other Nontra-

ditional Families, 78 *Geo. L.J.* 459 (1990).

74. Id. at 561-562 (citing Kleber, Howell & Tibbits-Kleber, The Impact of Parental Homosexuality in Child Custody Cases: A Review of the Literature, 14 *Bull. Am. Acad. Psychology & L.* 81, 86 (1986); Pennington, Children of Lesbian Mothers in *Gay and Lesbian Parents* 58-59 (F. Bozett, ed. 1987); Bozett, Children of Gay Fathers, in *Gay and Lesbian Parents* 39-57 (F. Bozett, ed. 1987)).

75. Id. at 562-565 (citing Kirkpatrick, Smith & Roy, Lesbian Mothers and Their Children: A Comparative Survey. 51 *Am J. of Orthopsychiatry* 545 (1981); Golombok, Spencer & Rutter, Children in Lesbian and Single Parent Households: Psychosexual and Psychiatric Appraisal 24 J. *Child Psychology & Psychiatry* 551(1983); Hotvedt, Green & Mandel, The Lesbian Parent: Comparison of Heterosexual and Homosexual Mothers and Their Children, Presentation at the Annual Meeting, American Psychological Association (Sep. 4, 1979); Green, The Best Interests of the Child with a Lesbian Mother 10 *Am. Acad. Psych. & L.* 7, 191(1982)).

76. Id. at 566.

77. Id.

78 Id. at 567.

79. Marc E. Elovitz, Adoption by Lesbian and Gay People: The Use and Mis-Use of Social Science Research, 2 *Duke J. Gender L. & Pol'y* 207 (1995).

80. Id. at 211.

81. Id. (citing David K. Flaks et al., Lesbians Choosing Motherhood: A Comparative Study of Lesbian and Heterosexual Parents and their Children 31 *Dev. Psychol.* 105, 111 (1995); Beverly Hoeffer, Children's Acquisition of Sex-Role Behavior in Lesbian-Mother Families, 51 *Am J. of Orthopsychiatry* 536, 543 (1981); Martha Kirkpatrick et al., Lesbian Mothers and Their Children: A Comparative Study, 51 *Am J. of Orthopsychiatry* 545, 550 (1981); Judith Ann Miller et al., The Children's Home Environment for Lesbian vs. Heterosexual Mothers: A Neglected Area of Research 1 *J. Homosexuality* 49, 55 (1981); Jerry J. Bigner & R. Brooke Jacobsen, Adult Responses to Child Behavior and Attitudes Toward Fathering: Gay and Nongay Fathers 3 *J. Homosexuality* 99, 109 (1992)).

82. Id. at 211.

83. Id. at 212.

84. Id. at 213.

85. Id. at 224-225.

86. Julie Shapiro, Custody and Conduct: How the Law Fails Lesbian and Gay Parents and Their Children, 71 *Ind. L.J.* 623 (1996).

87. Id. at 650.

88. Id. at 651-654.

89. David K. Flaks, Gay and Lesbian Families: Judicial Assumptions, Scientific Realities, 3 *Wm. & Mary Bill Rts. J.* 345 (1994).

90. Id. at 371.

91. Id. at 372.

92. Charlotte J. Patterson, Adoption of Minor Children by Lesbian and Gay Adults: A Social Science Perspective 2 *Duke J. Gender L. & Pol'y* 191 (1995).

93. Id. at 198-200.

94. Id. at 202-203.

95. Id. at 203.

96. Id. at 204.

97. Id. at 205.

98. Lynn D. Wardle, The Potential Impact of Homosexual Parenting on Children, 1997 *U. Ill. Law Rev.* 833 (1997).

99. Id. at 844.

100. Id. at 845-851.

101. Id. at 852.

102. Id. at 852-855.

103. Id. at 857.

104. Carlos A. Ball & Janice Farrell Pea, Warring with Wardle: Morality, Social Science, and Gay and Lesbian Parents, 1998 *U. Ill L. Rev.* 253 at 272 & 275 (1998).

105. Lynn D. Wardle, Fighting with Phantoms: A Reply to Warring with Wardle, 1998 *U. Ill. Law Rev.* 629, 636 (1998).

106. Ball & Pea at 291.

107. Id. at 294.

108. Id. at 296-297 (quoting David K. Flaks, Gay and Lesbian Families: Judicial Assumptions, Scientific Realities 3 *Wm. & Mary Bill Rts. J.* 345, 365 (1994)).

109. Lynn D. Wardle, Fighting with Phantoms: A Reply to Warring with Wardle 1998 *U. Ill. Law Rev.* 629, 636-637 (1998).

Appendix 4
No Balance:
Same-Sex Parenting Studies in the News

By Kristina Mirus
Marriage Law Project

The news media has an impact on public opinion. Often the reporter's own viewpoint is promoted by the tone of the article. Through the citation and presentation of evidence, a reporter can help to legitimize and build support for a specific position in the eyes of the public.

Has this been going on in the coverage of the studies evaluated in this book? A Lexis-Nexis search of current and archived news articles retrieved 143 stories between January 1, 1979 and December 31, 1999 that mention homosexual parenting studies.[1] Overall, these articles present a biased view of the studies and their results. The vast majority of the articles generalized that all studies prove conclusively that children raised by homosexuals are no different than children raised by heterosexuals. Often this is done without naming a specific study. The coverage these same-sex parenting studies have received in the news media has contributed to their acceptance by the general public, including policymakers.

A substantial amount of this influence comes from the slanted portrayal of the studies found in most of these articles. Though expressed in various ways, an overwhelming majority of the articles report the studies as proving that there is no significant difference in

Notes for this section begin on Page 148

child outcomes between children raised by homosexuals and children raised by heterosexuals. One-hundred-twenty-one of the 143 articles (85 percent) presented the studies in this manner. The language used often indicates that these studies are to be considered conclusive and indisputable. For example, an article in Memphis' Commercial Appeal on a same-sex "marriage" court decision states, "A mounting body of research by sociologists and psychologists has contributed to a scientific consensus that children of homosexual parents are no more likely to grow up troubled or homosexual than other children."[1] Some go further: "the research suggests that any presumption of unfitness [of homosexual parents] rests solely on prejudice and false stereotypes."[2] All of these 121 articles suggest that the studies done to date prove that homosexuals make good parents.

The other 22 of the 143 articles (15 percent) present a more balanced view of the studies: though the studies exist, they should not necessarily be viewed as conclusive. The majority of these 22 articles cite crucial methodological flaws, such as small sample sizes, which, as this evaluation has shown, make the results of these studies scientifically useless. Some of these 22 articles identify other psychologists or family studies that suggest that children are better off when raised by both a man and a woman. A few explain that the statement "no significant difference" does not necessarily mean that homosexuals have been proven to be good parents, but rather that where differences were found between homosexual and heterosexual parents, these variances were not considered "statistically significant" by the researchers. Not only do most of the 143 articles present these studies from a single viewpoint, the majority of the articles do so with all-inclusive statements, seldom actually mentioning any specific study.

One-hundred-six of the 143 articles (74 percent) do not cite or mention a single same-sex parenting study. While a few of these articles present their facts on the authority of an individual psychologist or of another essay written on the subject, they do not refer to any particular study when claiming that "studies have proven" the end goal the studies were to support. For example, an article on gay families in the Seattle Post-Intelligencer simply states, "Studies thus

far have shown them [children of gay parents] to be no better or worse off than other children, and no more likely to be gay."[3] Similar statements may be made in a negative fashion, such as this one in a New Jersey paper: "There is no research establishing the lesser capacity of gay and lesbian parents over other parents."[4] But all 106 articles have this in common: they never name a specific study. Conversely, the other 37 of the 143 articles (26 percent) did point out a particular study (or studies). The most commonly cited author was University of Virginia psychologist Charlotte Patterson, a strong supporter of lesbian parenting. But even in these 37 articles, eight of the specific citations are accompanied by a general statement to the effect that all other studies have found the same results. Why were these articles mentioning same-sex parenting studies written? Fifty of the 143 articles were written in response to a court case or piece of legislation dealing with same-sex parenting issues such as custody or adoption laws. Nine of them were written about the studies themselves, and 21 were editorials or opinion pieces.

But the majority (83) of these articles were written simply as public interest pieces. They focused on gay rights issues, such as adoption or foster care, and same-sex marriage, or offered written portrayals of alternative families. The answer to the question of how the media has portrayed same-sex parenting studies can be summed up in the following terms: an overwhelming majority (85 percent) of the articles have an overall tone suggesting, and often stating, that the results of these studies are beyond criticism. Only a small minority (15 percent) suggests that the studies could be flawed. Not only do most of the articles present only one viewpoint, the majority of them (74 percent) do so with broad generalizations, rarely referring to any specific studies. Of those articles that do cite a particular study (26 percent), almost a third follow it with a general claim that all other studies are in agreement. Further, these articles were most often written simply to discuss gay and lesbian issues and used the studies as evidence that gay and lesbian families are the same as heterosexual families. It is ironic that a media that prides itself on its critical acumen would treat same-sex parenting studies with such extraordinary deference. We leave the reader to speculate as to why this is the case.

Notes to Appendix 4

1. Two searches were done, one in the current news library and one in the archived news library. Both were worded as follows: "gay or lesbian or homosexual w/para study and parenting or adoption." "Judge on gay marriage: The kids are all right?" The Commercial Appeal, 8 December 1996, p. 6A.

2. Mike McKee, "Farella Weighs In on Florida Gay Rights Case," The Recorder, 18 April 1995, p. 2.

3. Lynn Steinberg, "Not So Different After All: Gay and Lesbian Families Put Down Roots," Seattle Post-Intelligencer, 20 April 1998, p. D1.

4. "Spun by Intolerance: Denial of Same-sex marriages Have Terrible Consequences," Asbury Park Press, 9 January 1997, p. 15.

About the Authors

Robert Lerner and **Althea K. Nagai** received their Ph.D.s from the University of Chicago in sociology and political science, respectively. They have coauthored three books: *Giving for Social Change* (Praeger), *Molding the Good Citizen* (Praeger), and *American Elites* (Yale). They are currently partners in Lerner and Nagai Quantitative Consulting, a social-science research consulting firm.

David Orgon Coolidge is the director of the Marriage Law Project based in Washington, D.C. at the Ethics and Public Policy Center and the Columbus School of Law, The Catholic University of America. He is a graduate of the Georgetown University's Law Center, and is licensed to practice law.

William C. Duncan is the Associate Director of the Marriage Law Project. He is a graduate of the J. Reuben Clark Law School at Brigham Young University, and is licensed to practice law.

Kristina Mirus, a graduate of Christendom College, worked with the Marriage Law Project from June 1999 to July 2000.