Applied Economics, 2006, **38**, 1–13

Routledge
Taylor & Francis Group

Same-sex sexual behaviour: US frequency estimates from survey data with simultaneous misreporting and non-response

Nathan Berg[a,*] and Donald Lien[b]
[a]*School of Social Sciences, University of Texas at Dallas, Richardson, Texas, USA*
[b]*University of Texas at San Antonio, San Antonio, Texas, USA*

Survey-based research concerning sexual behaviour almost inevitably confronts the simultaneous problems of misreporting and non-response. These problems lead to disparities among estimates of the number and characteristics of those who engage in same-sex sexual behaviour. This paper proposes a statistical model to consistently estimate the frequency of same-sex sexual behaviour in the presence of non-ignorable misreporting and non-response. The model is fitted using 1991–2000 General Social Survey data. Frequency estimates corrected for simultaneous misreporting and non-response are reported. According to the model, 7.1% of US males and 4.1% of females – 15.8 million individuals – are not exclusively heterosexual. Allowing for misreporting and non-response increases the estimated same-sex frequency by more than four million. The model reveals new patterns between misreporting and non-response probabilities and standard demographic variables such as age and income.

I. Introduction

Do we know what fraction of people engage in same-sex sexual behaviour? Existing estimates in the US span a perplexingly large range, from 1% to 10% of the adult population and beyond (Lauman *et al*., 1994; Michael *et al*., 1994; Badgett, 1995; Murray, 1999). Unfortunately, divergent definitions of sexual orientation and disagreements about the meaning of the label gay (Lauman *et al*., 1994; Michael *et al*., 1994; Black *et al*., 2000; Murray, 1999), result in empirical treatments that do not precisely track the population of interest in this paper – Americans who have had sex with at least one same-sex partner within the last five years.

1

The population of those who are not exclusively heterosexual plays an important role in a number of policy questions. Forecasts of the spread of AIDS

*Corresponding author. E-mail: nberg@utdallas.edu

Applied Economics ISSN 0003–6846 print/ISSN 1466–4283 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00036840500427114

and cost-benefit analyses of initiatives to prevent the spread of sexually transmitted disease require as inputs the size of high-risk populations such as sexually active men with same-sex partners (Bloom and Glied, 1992; Thomas, 2001). Tabulations of projected costs and benefits for legislative proposals to protect non-heterosexuals against workplace discrimination, such as the Employment Non-Discrimination Act which was proposed in the US Congress multiple times before being defeated in 1996 (Badgett, 2001), require size estimates of the potentially affected population. And anti-sodomy laws, traditionally used to prosecute same-sex sexual behaviour (e.g., *Lawrence and Garner* v *Texas*), and movements to rescind those laws provide yet another example in which quantification of the affected population requires estimates for which there exists little reliable data.

Related, yet distinct, is the more narrowly defined gay population, definitions of which often require exclusively homosexual behaviour or public self-identification of a gay identity (i.e., being out). Previous studies have suggested that sexual orientation both at the individual and macro levels are indeed economically consequential. Florida (2002) argues that cities' gay populations help stimulate economic growth. Marketers, religious activists, and gay rights advocates who debate the degree to which personal income and sexual orientation are correlated often agree that gays are economically distinct as consumers and workers (Allegretto and Arthur, 2001; Arabsheibani *et al.*, 2001; Berg and Lien, 2002). Demographers with opposing views about the extent and causes of geographic concentration of gays in particular cities (D'Emilio, 1989; Chauncey, 1994; Black *et al.*, 2002) seem to agree that differential responses to economic incentives by sexual orientation play a role. Those studies also inspire questions about whether unconventional definitions used to measure gay identity combined with systematic misreporting and non-response may generate spurious spatial effects. In debates surrounding issues such as workplace discrimination (Plug and Berkhout, 2004; Black *et al.*, forthcoming), gay marriage (Alm *et al.*, 2000), and regulations concerning child adoption (Collum, 1993), divergent estimates of the gay population's size and demographic characteristics clearly contribute to persistence of disagreement over policy.

Unfortunately, the difficulty of accurately measuring incidence rates of sexual behaviour and sexual orientation raises doubt about the validity of existing empirical characterizations of populations defined by sexual behaviour and orientation (Lauman, *et al.*, 1994; Badgett, 1997). There is a well-known tendency for survey questions about sexual behaviour to elicit non-random patterns of misclassification and non-response (Pearl and Fairley, 1985; Kupek, 1998; Marquie and Baracat, 2000). It is strongly suspected that naive estimates (e.g., those that handle non-response by invoking an unverified missing-at-random assumption) systematically undercount same-sex and gay populations because they have special incentives to misreport and non-respond in many survey settings. Similarly, it is suspected that the empirical distribution of demographic variables among self-reported homosexuals is distorted because those variables are correlated with propensities to misreport and non-respond.

The goal of this paper is to develop a probability model that simultaneously deals with misreporting and non-response and is capable of producing superior estimates of the size and characteristics of the same-sex-partner population. The parametric probability model introduced here encompasses the missing-at-random and missing-completely-at-random hypotheses as testable parameter restrictions. In addition to providing improved point estimates of the incidence of non-heterosexuality, the model yields explicit expressions for misreporting and non-response probabilities as functions of observable individual characteristics such as income, age, residential city size, marital and parental status.

The plan of the paper is as follows. Section II reviews related methodological studies of non-response and misreporting. Section III specifies the statistical model and shows how to derive estimates from it. Section IV describes the data, reports estimated probabilities of misreporting and non-response, and provides revised estimates of the non-heterosexual population's frequency and size. Section V concludes with a discussion of the main results and their implications.

II. Methodological Background

Within the non-response literature, Little and Rubin (1987), Rubin (1987), and Little (1993) distinguish between selection- and pattern-mixture approaches. This paper's model follows the pattern-mixture approach, which assumes that the decisions of respondents and non-responders arise from completely different conditional distributions. The argument in favour of this approach is the intuitive appeal of the notion that responders and non-responders are two wholly different groups with covariates that follow entirely different joint distributions. Ekholm and Skinner (1998), Forster and Smith (1998), and Lee and Marsh (2000) provide applications in which the pattern-mixture approach yields clear advantages. In contrast, the selection approach assumes that a single regression model applies to the entire (hypothetically complete) data set and appends to it additional equations intended to capture the process of selection, most often a probability model of the chance of being missing from the sample. Applications that successfully adopt the selection approach include Heckman (1979), Lien and Rearden (1990), Stasny (1991), Conaway (1992), Lipsitz *et al*. (1994), Baker (1995), Roy and Lin (2002), and Nandram and Choi (2002).

In the misclassification literature, economists and statisticians have demonstrated that it is not necessary to directly observe misreporting in order to estimate its frequency (Hausman *et al*., 1998; Black *et al*., 2000a). This is counterintuitive and prompts the question of how rates of misreporting can be estimated without observing the phenomenon directly. The underlying idea is to exploit information contained in the right-hand side variables that are correlated with non-response and misreporting while including information from incomplete observations that are typically discarded in the estimation of naive models.

The model used here handles misreporting and non-response simultaneously whereas previous work usually treats them as separate problems. An exception is Wu (2002) who models a complex error structure generated by a combination of imperfectly-measured patient outcomes and premature patient dropouts using data from a medical study of an AIDS treatment.

Survey design methodologists concerned with misreporting and non-response have developed techniques tailored to situations where researchers can influence sample design or collect additional data. For example, Embree and Whitehead (1991) compare self-reported alcohol consumption with aggregate state alcohol sales statistics to produce quantitative corrections for raw survey frequencies. Referred to as cross validation, this technique has revealed significant bias in self-reported church-going, charitable giving, and high-risk sexual contact (Kupek, 1998; Turner, 1999; Berg, 2005). By conducting multiple surveys of the same population using different survey instruments, useful predictions of the chances of misreporting and non-response can be demonstrated using straightforward techniques (Whitehead *et al*., 1993). In contrast to situations in which researchers plan for misreporting and non-response at the stage of survey design and have access to multiple sources of data, the present model is specialized for secondary analysis of a single set of survey data without requiring further data collection.

III. The Model

It is assumed that every individual can be categorized as either heterosexual or non-heterosexual. Non-heterosexuals are defined as those who have had at least one same-sex sexual partner, while the heterosexual category includes everyone else, including those who have had no sex partners of either gender. The rationale for this simplifying categorization scheme is to focus information in the data on the quantity of interest, the incidence of same-sex sexual behaviour.

The probability that a randomly selected individual's true behavioural state is non-heterosexual is represented by the function $g(\phi' x)$, which is assumed to depend on a $K \times 1$ vector of covariates x, not including a constant, and a vector of parameters ϕ. The probability function g is assumed to be smooth and non-linear in the linear index $\phi' x$, as is the case with a logistic or normal pdf. If same-sex behaviour were easy to observe, standard statistical techniques could be used to estimate ϕ and the empirical relationship between observable traits (x) and sexual behaviour could be established.

The gender(s) of an individual's sexual partners are difficult to observe directly, however. Surveys asking respondents for such information yield observations of a different variable, self-reported sexual behaviour, taking on one of three possible values: 'heterosexual,' 'non-heterosexual' or 'no response.' Rather than assuming that survey responses map transparently into true behavioural states, the model allocates positive probability to all six pairs of the two true behavioural states and three self-reported sexual orientations.

By assumption, misreporting and non-response probabilities for individuals whose true behavioural state is non-heterosexual depend on x. The probability $m(\beta' x)$ represents the chance that a non-heterosexual with demographic characteristics x misreports his or her behaviour as exclusively heterosexual. The non-heterosexual individual's chance of non-response is represented by the probability function $n(\gamma' x)$. The vectors of slope parameters, β and γ, transform x into two linear indexes whose weights indicate the relative importance of different covariates.

The probability of non-response conditional on an exclusively heterosexual true behavioural state is given by the function $r(\rho' x)$. Non-response among heterosexuals also depends on x. The linear weights that enter the probability function are not constrained to coincide with those of non-heterosexuals, however. Similar to g, the functions m, n and r are pdfs on linear indexes in x.[1]

Economic and non-economic incentives motivating the hypothesized relationship between x and probabilities of misreporting and non-response draw upon a rich, although partially anecdotal, empirical backdrop. Horrific stories of violence against homosexuals (e.g., the murders of Brandon Teena in Falls City, Nebraska [1993], Matthew Shepard in Laramie, Wyoming [1998], and Danny Overstreet in Roanoke, Virginia [2000]) illustrate a basic motive – averting physical threat – to be less than fully open about same-sex sexual activity. Regarding possible labour market incentives, legal precedent is noted in some states supporting employers who refuse to hire homosexuals (*England* v *the City of Dallas*). Moreover, outspoken criticism of homosexuality by prominent political leaders (e.g., Pat Buchanan's speech at the 1992 Republican National Convention) suggests links between characteristics such as geographic location and socioeconomic status and non-heterosexuals' propensities to misreport and non-respond.

There is comparatively little theoretical support for asserting a stable relationship between heterosexual misreporting and survey respondents' demographic characteristics. Heterosexual misreporting occurs when an individual who has never had a same-sex partner incorrectly reports same-sex partners in his or her sexual history. Heterosexual misreporting is assumed to occur for highly idiosyncratic reasons and, therefore, the model assumes independence between x and the heterosexual misreporting frequency M.

Heterosexual non-response is different. There is, for example, prior evidence that older heterosexuals and other demographically defined subsets are more likely to refuse to answer survey questions about sexual behaviour (Kupek, 1998). Thus, the heterosexual non-response probability $r(\rho' x)$ is specified as a function of x.

In this paper individual i's true behavioural state is represented with the symbol $G_i \in$ {hetero, non-hetero} and self-reported sexual behaviour as $Y_i \in$ {report hetero, report non-hetero, no response}. The joint pdf of true states and self-reports can be expressed as:

	report hetero	report non-hetero	no response
true hetero	$(1-g(\phi' x)) \times (1-M-r(\rho' x))$	$(1-g(\phi' x))M$	$(1-g(\phi' x)) \times r(\rho' x)$
true non-hetero	$g(\phi' x)m(\beta' x)$	$g(\phi' x)(1-n(\gamma' x)-m(\beta' x))$	$g(\phi' x)n(\gamma' x)$

The marginal probability of truly having at least one same-sex partner is obtained by summing horizontally. The marginal probabilities of reporting heterosexual, reporting non-heterosexual, and non-responding, are obtained by summing vertically.

Additional structure must be imposed on the functions g, m, n and r in order for the slope parameters ϕ, β, γ and ρ to be identified. To see what can go wrong without additional constraints, consider the neighbourhood in parameter space about the point $\phi = \beta = \gamma = \rho = 0$. This point corresponds to the situation in which x is completely uninformative in estimating misreporting and non-response probabilities, in which case variation in x is irrelevant and there are only two independent pieces of information, the number of self-reported heterosexuals N_s and the number of self-reported non-heterosexuals N_g. (Denoting the number of non-responses as N_n and the overall sample size as N, the equation $N = N_s + N_g + N_n$ holds trivially.) The problem is that the model requires five parameter estimates, the constants M, $g(0), m(0), n(0)$ and $r(0)$, using only two observable pieces of information, N_s and N_g. The model is under-identified. Referring to the point in parameter space at which all slope parameters are zero as the case in which x is completely uninformative, the following restrictions are imposed:

Assumption 1: If x is completely uninformative, rates of non-response among non-heterosexuals and heterosexuals are equal: $n(0) = r(0)$.

Assumption 2: If x is completely uninformative, rates of misreporting among non-heterosexuals and heterosexuals are equal: $m(0) = M$.

Assumption 3: If x is completely uninformative, the marginal probability of non-response is equal

[1]It is possible to let g, m, n, and r depend on different sets of covariates. This can be implemented by concatenating all independent variables into the vector x and imposing zero restrictions on particular elements of φ, β, γ and ρ. Since there is no reason to rule out possible connections between x and reporting probabilities *a priori*, and because identification of the model (discussed subsequently) does not require it, no such restrictions are imposed.

to the empirical non-response rate: $(1-g(0))r(0)+g(0)n(0)=(N_n/N)$.

Assumption 4: If x is completely uninformative, the marginal non-heterosexual probability is equal to the empirical frequency of self-reported non-heterosexuality among those who are not non-responders: $(1-g(0))M+g(0)(1-m(0)-n(0))=(N_g/(N_g+N_s))$.

In assessing how reasonable these assumptions are, it is important to point out what they allow and precisely what they rule out. A key feature of the model is that misreporting and non-response probability functions may differ according to whether an individual's true behavioural state is heterosexual or otherwise. Assumptions 1 and 2 require that whenever the covariates x contain no information about misreporting and non-response, then no difference across unobserved underlying behavioural states can be claimed. This reflects an agnostic prior that begins search in parameter space using symmetric guesses based on unconditional frequencies. As long as x helps predict misreporting and non-response, then misreporting and non-response probabilities are free to vary across heterosexual and non-heterosexual behavioural states. However, when x contains no information, the default misreporting and non-response probabilities are symmetric. Thus, the model allows the data to decide the extent to which there is variation across types without building in differential rates of misreporting and non-response.

Similarly, Assumption 3 centres estimated non-response probabilities on the empirical non-response rate while allowing the data to guide search elsewhere using a likelihood criterion. If x provides no basis for adjusting an individual's estimated probability of non-response by the likelihood criterion, then non-response estimates do not move away from the unconditional frequency estimates which are based on face-value interpretations of the data. Assumption 4 establishes an analogous prior for the probability that an individual's true behavioural state is non-heterosexual. Guesses are centred at face-value empirical frequencies, based on self-reported sexual behaviour, and adjusted away from those priors only when x is informative in the sense that at least one of the slope parameters in the non-heterosexual probability function is non-zero.

Assumptions 1–4 imply four functional relationships between the constants $g(0)$, $m(0)$, $n(0)$, $r(0)$ and the parameter M conditional on the observed number of self-reported heterosexuals and non-heterosexuals, N_s and N_g. This suggests the possibility of line search on M and, at each prospective value, maximum likelihood estimation of the slope parameters. The likelihood function then decides which value of M and corresponding MLE slope estimates are best. This technique is used to derive the estimates reported below. Additional details of the estimation procedure are relegated to the Appendix. The explicit functional dependence of $g(0), m(0), n(0)$, and $r(0)$ on M, N_s and N_g is presented there, along with a modified logistic specification of the probability functions g, m, n and r.

To develop the likelihood function, it is convenient to represent the three possible realizations of self-reported sexual behaviour as a set of three indicator variables:

$$z_{i0} = \begin{cases} 1 & \text{if i reports heterosexual} \\ 0 & \text{otherwise} \end{cases}$$

$$z_{i1} = \begin{cases} 1 & \text{if i reports non-heterosexual} \\ 0 & \text{otherwise} \end{cases}$$

and

$$z_{i2} \equiv 1 - z_{i0} - z_{i1} = \begin{cases} 1 & \text{if } i \text{ refuses to respond} \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function is:

$$\begin{aligned} &L(\phi, \beta, \gamma, \rho, M|x) \\ &= \textstyle\prod_{i=1}^{N}[(1-g(\phi' x_i))(1-M-r(\rho' x_i)) \\ &\quad + g(\phi' x_i)m(\beta' x_i)]^{z_{0i}} \times [(1-g(\phi' x_i))M + g(\phi' x_i) \\ &\quad \times (1-m(\beta' x_i)-n(\gamma' x_i))]^{z_{1i}} \times [(1-g(\phi' x_i))r(\rho' x_i) \\ &\quad + g(\phi' x_i)n(\gamma' x_i)]^{z_{2i}}. \qquad (1) \end{aligned}$$

Although it is not explicit in Equation 1, the functions g and m depend on M through the constants $g(0)$ and $m(0)$.

The strategy for numerically optimizing Equation 1 combines line search on M with steepest descent and Newton–Raphson refinement on the $4K \times 1$ vector $[\phi' \ \beta' \ \gamma' \ \rho']'$. For each proposed value of M, a numerical solution is computed to solve first order conditions (Equations A24–A27 in the Appendix). Among all pairs of proposed M values and associated slope estimates, the pair that maximizes Equation 1 provides a bias-corrected estimate of misreporting and non-response probabilities that vary according to x. Taking the average estimated probability or the estimated probability evaluated at the average value of x, an estimated rate of same-sex sexual behaviour follows. Multiplying the estimated rate times the total adult US population provides an estimate of the number of Americans who engage in same-sex sexual behaviour. Standard errors and t statistics for these quantities are computed based on Taylor expansions of the non-linear probability functions with respect to $[\phi' \ \beta' \ \gamma' \ \rho']'$ and M and the expressions for asymptotic variance they yield.

At the point in parameter space where x is uninformative, the model is, by construction, locally identified as a result of assumptions 1–4. At other points in parameter space, non-linearity of the probability functions succeeds in locally identifying the model. This can be verified by checking the rank of the second derivative matrix of the log-likelihood function, an expression for which appears in the Appendix. After unsuccessfully attempting to prove global identification, numerical rank tests at a variety of points in parameter space were settled for, including all maximum likelihood estimates reported here.

IV. Data and Estimation

The model is estimated using GSS data from 1991 through 2000.[2] Out of a potential pool of 10 458 survey responses in the GSS from 1991–2000, over 20% contain one or more missing items among the variables in the model. The variables most frequently missing are those that code sexual behaviour variable and income.

Respondents who non-respond to items pertaining to sexual behaviour while providing valid responses to other items are, of course, included in the sample. Survey respondents who non-respond on other items aside from those concerning sexual behaviour (giving rise to incomplete observations of the independent variables x) do, however, create a potential selection problem. The variables in x correspond to survey items that seek to elicit standard, non-stigmatizing demographic information. Therefore, no attempt is made to augment what is already a highly parameterized model by accounting for other possible forms of selection bias.

After dropping partial responses with missing items other than self-reported sexual behaviour, the sample size drops to 8446, with 4263 women and 4183 men. Respondents with missing self-reported sexual behaviour but otherwise complete item responses remain in the sample and are coded with a dependent variable distinguishing non-response, $(z_{0i}, z_{1i}, z_{2i}) = (0, 0, 1)$, from the other two possible self-reports of heterosexual $(z_{0i}, z_{1i}, z_{2i}) = (1, 0, 0)$ and non-heterosexual $(z_{0i}, z_{1i}, z_{2i}) = (0, 1, 0)$. Only 157 women (3.9% excluding non-responders) self-report non-heterosexual, and 214 (5.0% of all women in the sample) non-respond. A total of 176 men (4.4% excluding non-responders) self-report as non-heterosexual and 180 (4.3% all men in the sample) non-respond.

Table 1 presents mean values of the independent variables x broken out by reporting decision and own gender. Standard errors appear in parentheses below each mean. According to Table 1, men who report having same-sex partners earn less than self-reported heterosexual men. In contrast, self-reported non-heterosexual women earn more than self-reported heterosexual women. The average self-reported non-heterosexual is younger than the sample average and lives in a larger city than the average respondent does. Self-reported non-heterosexuals are less likely to have children, less likely to be married, and are better educated (as measured by the number of degrees held, where 0 indicates no high school, 1 indicates completion of high school, 2 indicates completion of junior college, 3 indicates college graduate status, and 4 indicates the completion of at least one graduate degree). Non-responders earn less than average, with non-whites disproportionately represented among them. The average non-responder is less well educated and is significantly older than average.

Table 2 reports model estimates based on separate male and female samples. Raw parameter estimates and estimated changes in probabilities are reported expressed in percentage points as the effect of a one-unit change in x on the chances, respectively, of misreporting and non-response. Raw parameter estimates and transformed probability effects give somewhat different impressions about the relative importance of different regressors. The magnitudes of the effects are variable, and parameters with small t statistics sometimes have large and statistically significant probability effects.

Estimates under the headings Δp in Table 2 are defined in such a way so as to compare the probabilities m, n, g and r for two individuals who have opposite values of one characteristic but are

[2]It is possible to extend the sample backward in time to 1972. There are at least two reasons for restricting the sample period to 1991–2000, however. First, the income variable in the GSS is coded by income brackets and those brackets (both the thresholds and the number of categories) have changed periodically from 1972 through 2000. Inflation and changes in the real income distribution make it difficult to link bracketed income data over long periods of time. Choosing a sample that begins in 1991 avoids all but one of the bracket changes, which ocurred between GSS coding schemes in 1996 and 1998. Income data are adjusted in the analysis by a simple linear transformation of the initial 21 categories onto the range of the lates 23 categories. Shifting attitudes toward homosexuality (Newport, 2001) provide a second rationale for sampling over relatively short time frames. Changes in the anticipated consequences and self-identification habits of non-heterosexuals would result in unstable or time-dependent functions m, n and r.

Table 1. Sample means among self-reported heterosexuals, non-heterosexuals and non-responders (GSS 1991–2000)

	Men				Women			
Variables	All	Report hetero	Report non-hetero	Non-respond	All	Report hetero	Report non-hetero	Non-respond
Income[a]	14.92	15.01	13.91	14.07	11.91	11.91	12.47	11.52
	(0.08)	(0.08)	(0.40)	(0.42)	(0.09)	(0.09)	(0.46)	(0.41)
Parent	0.66	0.67	0.32	0.65	0.73	0.74	0.50	0.73
	(0.01)	(0.01)	(0.04)	(0.04)	(0.01)	(0.01)	(0.04)	(0.03)
White	0.85	0.85	0.80	0.81	0.80	0.81	0.83	0.66
	(0.01)	(0.01)	(0.03)	(0.03)	(0.01)	(0.01)	(0.03)	(0.03)
Married	0.58	0.60	0.23	0.60	0.56	0.57	0.31	0.55
	(0.01)	(0.01)	(0.03)	(0.04)	(0.01)	(0.01)	(0.04)	(0.03)
Degrees[b]	1.64	1.64	1.85	1.47	1.62	1.63	1.69	1.40
	(0.02)	(0.02)	(0.09)	(0.09)	(0.02)	(0.02)	(0.10)	(0.07)
Age	40.19	40.07	39.01	43.78	39.49	39.30	36.18	45.33
	(0.19)	(0.20)	(0.85)	(0.98)	(0.18)	(0.18)	(0.84)	(1.05)
City size[c]	0.34	0.32	0.85	0.33	0.36	0.34	0.58	0.43
	(0.02)	(0.02)	(0.14)	(0.09)	(0.02)	(0.02)	(0.13)	(0.10)
N	4183	3827	176	180	4263	3892	157	214

Notes:
[a] Income is measured categorically on a 21-category scale in which the top bracket is US $75 000 and above, for 1991, 1993, 1994 and 1996. In 1998 and 2000, income is measured on a 23-category scale in which the top bracket is US $110 000 and above. Surveys were not conducted in 1992, 1995, 1997 and 1999. The categorical variable from the earlier years was transformed to the 23-category scale by assigning top bracket individuals from earlier years to the top bracket for later years, and by adjusting midpoints for inflation (using the CPI-U index), recording the adjusted values on a 23-point scale. The average male's income (in the 13–15 range) corresponds to US $20 000–US $30 000 in 1998 dollars. The average female income (in the 11–12 range) corresponds to US $15 000–US $20,000.
[b] The variable Degrees is a count variable ranging from zero to four that indicates the number of degrees each respondent has earned.
[c] The variable City size is adjusted to cover the unit interval. City size in its unadjusted form ranges from 1000 to 7.3 million people. The overall average adjusted city size of 0.35 corresponds to a city population of 2.6 million. The average self-reported homosexual male's city size of 0.85 corresponds to a city population 6.2 million. The median city size for all males is approximately 28 000, and 76 000 among self-reported gay males.

otherwise average. Denote the arithmetic mean of x with its kth component replaced by the maximum sample value $\max(x_k)$ as $\bar{x}_k^{\max}$. Similarly, denote average x with its kth component replaced by the minimum sample value as $\bar{x}_k^{\min}$.[3] For a generic probability function $p \in \{m(\cdot), n(\cdot), g(\cdot), r(\cdot)\}$, and generic slope parameter $\eta \in \{\beta, \gamma, \phi, \rho\}$, define

$$\Delta p_k \equiv p(\eta' \bar{x}_k^{\max}) - p(\eta' \bar{x}_k^{\min}) \qquad (2)$$

The estimated asymptotic standard error $\mathrm{se}_{\Delta p_k}$ used in computing the t statistics labeled $t_{\Delta p}$ in Table 2 are computed as the square root of:

$$\left[\partial p(\hat{\eta}'\bar{x}_k^{\max})/\partial\eta - \partial p(\hat{\eta}'\bar{x}_k^{\min})/\partial\eta\right]' \mathrm{var}(\hat{\eta}) \times \left[\partial p(\hat{\eta}'\bar{x}_k^{\max})/\partial\eta - \partial p(\hat{\eta}'\bar{x}_k^{\min})/\partial\eta\right], \qquad (3)$$

where the theoretical value of the matrix $(\hat{\eta})$ is replaced by an outer product estimator.

The top panel of Table 2 shows that moving from the lowest to the highest income category increases a behaviourally homosexual man's misreporting probability by 42 percentage points. Among behaviourally homosexual men, the model suggests that whites are significantly more likely to misreport, as are young males and those who live in small towns. Among female non-heterosexuals, only city size appears to have noticeable effects, with lower chances of misreporting in large cities.

The non-response probability for female non-heterosexuals is generally more sensitive to x than for males. Moving from the lowest to the highest income bracket, or from the highest to the lowest age group (72 to 18), reduces the average

[3] To improve the numerical stability of computer routines used in likelihood maximization, all independent variables were scaled to range over the unit interval. Because most are 0/1 variables, rescaling changed only three variables: degrees, age and city size. Thus, $\bar{x}_k^{\max}$ and $\bar{x}_k^{\max}$ are equal to $\bar{x}$ with the kth component replaced by 1 or 0.

Table 2. Estimated parameters and changes in probability 2

	Men ($N=4183$)				Women ($N=4263$)			
	θ	t	Δp	$t_{\Delta p}$	θ	t	Δp	$t_{\Delta p}$
Effect on Pr (Misclassification|Homosexual), $m(\beta' x)$								
Income	15.60	0.6	0.42	0.5	−83.67	−0.1	0.00	−0.1
Parent	9.37	0.5	0.05	0.2	76.47	0.1	0.00	1.1
White	18.79	0.7	0.47	144.0	−42.37	−0.1	−0.00	−0.1
Married	5.05	0.4	0.00	0.1	97.45	0.1	0.00	0.2
Degrees	−12.94	−0.6	−0.16	−0.1	−9.73	−0.1	−0.00	−0.1
Age	−34.03	−0.6	−0.48	−1283.8	84.45	0.1	0.00	0.1
City size	−50.98	−0.5	−0.50	−2577.7	−95.77	−0.1	−0.46	−31.1
Effect on Pr (Nonresponse|Homosexual), $n(\gamma' x)$								
Income	0.63	0.2	0.07	0.2	−11.31	−0.4	−0.49	−3.8
Parent	0.61	0.3	0.06	0.3	−2.04	−0.2	−0.19	−0.4
White	−1.49	−0.5	−0.11	−0.5	−23.25	−0.4	−0.47	−34.3
Married	2.11	0.6	0.22	0.8	2.90	0.3	0.20	0.0
Degrees	1.88	0.7	0.17	0.6	6.97	0.4	0.46	1.5
Age	4.80	0.8	0.37	1.2	53.18	0.4	0.50	1338.2
City size	−8.64	−0.5	−0.40	−2.0	−63.88	−0.3	−0.42	−0.8
Effect on Pr (True behavioral state is Homosexual), $g(\phi' x)$								
Income	0.44	0.9	0.02	0.9	−0.16	−0.3	0.00	−0.4
Parent	−1.35	−3.9	−0.07	−3.1	−0.66	−1.9	−0.02	−1.9
White	0.15	0.5	0.01	0.5	−0.11	−0.3	−0.00	−0.5
Married	−1.51	−3.4	−0.08	−3.6	−1.49	−2.5	−0.05	−3.4
Degrees	0.63	1.5	0.03	1.4	0.24	0.6	0.01	0.6
Age	3.25	4.0	0.21	2.3	3.48	3.1	0.15	2.1
City size	0.09	0.2	0.00	0.2	−0.02	0.0	−0.01	−0.0
Effect on Pr (Nonresponse|Heterosexual), $r(\rho' x)$								
Income	−1.20	−1.0	−0.04	−0.7	−0.59	−0.7	−0.02	−0.3
Parent	−0.19	−0.3	0.00	−0.1	−0.47	−0.9	−0.02	−0.2
White	−0.06	−0.1	−0.00	−0.0	−1.12	−2.8	−0.05	−1.1
Married	0.42	0.5	0.01	0.2	−0.04	−0.1	−0.00	−0.2
Degrees	−1.79	−1.3	−0.04	−0.9	−1.02	−1.2	−0.03	−0.3
Age	2.02	1.2	0.06	0.3	3.38	2.6	0.15	0.6
City size	0.41	0.4	0.01	0.1	0.50	0.6	0.02	0.1
Estimated Pr (Misreport|Heterosexual)								
M	0.013	1.7			0.016	1.4		
Pseudo R^2	0.08				0.07			
Log likli.	−1364.4				−1414.2			
LL ratio	240.19				197.17			

non-heterosexual woman's non-response probability by more than 45 percentage points. There is also a strong education effect, where non-heterosexual females with more degrees are more likely to non-respond. Among non-heterosexual males, the magnitudes of the effects indicate suggest that only city size has much of an effect on the probability of non-response, a decreasing function of city size.

The third panel of Table 2 presents estimated relationships between the components of x and the probability of being behaviourally non-heterosexual controlling for misreporting and non-response. As one might expect, being a parent or married in a heterosexual marriage decreases the chance of homosexual behaviour. Recall, however, from Table 1 that more than 30% of self-reported non-heterosexuals (50% among female non-heterosexuals) are parents, and more than 20% are married. Therefore, it is incorrect to assume that same-sex sexual behaviour and being a parent, or being in a heterosexual marriage, are mutually exclusive. Also noteworthy is that, after correcting for reporting bias, non-heterosexuals do not appear more numerous in large cities. Knowing that an individual lives in a large city appears to increase the chance that he or she will truthfully disclose non-heterosexuality conditional on non-heterosexual behaviour. But there is no evidence that big city residents are actually more likely to engage in same-sex behaviour, contrary to numerous claims in the demographic literature that rely on survey frequencies while making no allowance for misreporting and non-response.

The fourth panel of Table 2 applies to heterosexuals and the chance of non-response. Age appears to be the single most important factor, with older heterosexuals disproportionately refusing to answer questions about sex partners and the gender of those partners. Small magnitudes and low t statistics among the other variables suggest that non-response among heterosexuals is difficult to predict. The final estimate presented in Table 2 is M, the probability that heterosexuals misreport, which turns out to be less than 2% for both men and women.

Table 3 presents likelihood ratio test statistics for several additional parameter restrictions corresponding to various notions of random misreporting and non-response. The first row of Table 3 suggests that x is a useful predictor for misreporting among non-heterosexuals. The second row tests the missing-at-random assumption (i.e., the hypothesis that non-response mechanisms among heterosexuals and non-heterosexuals depend on x in the same way) leading to rejection. Reported in the next row is that the missing-completely-at-random hypothesis (i.e., the proposition that non-response is not a function of x) is rejected as well. The last two rows demonstrate that non-heterosexual behavioural parameters are jointly significant (at least in the statistical sense) as is the model as a whole.

Having shown that misreporting and non-response mechanisms are predictable in the sense that they arise from a model whose parameters (and therefore derivatives with respect to x) are non-zero, Table 4 presents corrected frequency estimates for same-sex behaviour in the USA. According to the model, 7.1% of men have had at least one same-sex sexual partner, significantly more than the self-reported rate of 4.4%. The corrected same-sex-behaviour frequency among women of 4.1% is not much higher than the self-reported rate of 3.9%. Nevertheless, rates of misreporting among non-heterosexuals men and women are significantly above zero, and their rates of non-response are considerably higher than heterosexuals. This contrasts sharply with naive frequency estimates based on face-value interpretations of self-reported data. The bias correction technique does not build in or presuppose positive amounts of misreporting and non-response or any differences between heterosexuals and homosexuals. The model is symmetric in that it allows for heterosexuals who self-report as non-heterosexuals (measured by the frequency M). Because there are no restrictions on the slope parameters requiring that non-response among heterosexuals be less than among non-heterosexuals, the model could in theory produce a lower same-sex frequency than is estimated naively from self reports. Therefore, the higher estimated number of individuals engaging in same-sex behaviour reflects information present in the data rather than priors built into the model.

Table 3. Likelihood ratio test statistics for misreporting/missing at random hypotheses

		Men		Women		
Hypothesis		Lik. ratio	p-value	Lik. ratio	p-value	df
Misreporting is random w.r.t. x	$\beta = 0$	27.3	0.0003	23.5	0.0014	7
Missing at random	$\gamma = \rho$	29.9	0.0001	21.6	0.0030	7
Missing completely at random	$\gamma = \rho = 0$	40.8	0.0002	30.3	0.0070	14
Misrep. and nonresp. independent of x	$\beta = \gamma = \rho = 0$	70.0	0.0000	127.1	0.0000	21
Sexual orientation independent of x	$\phi = 0$	144.6	0.0000	71.4	0.0000	7
All slopes zero	$\beta = \gamma = \phi = \rho = 0$	240.2	0.0000	197.2	0.0000	28

Table 4. Estimated homosexual, misreporting and nonresponse frequencies

Bias-corrected frequency:	Men			Women			
estimates	$\Sigma_i\, p(n'x_i)/n$	se	Naive	$\Sigma_i\, p(n'x_i)/n$	se	Naive	Naive formula
Homosexual (g)	0.071	0.014	0.043	0.041	0.009	0.039	$N_g/(N_s + N_g)$
Misreport|Homosexual (m)	0.41	0.002	0	0.39	0.001	0	–
Nonrespond|Homosexual (n)	0.33	0.233	0.043	0.21	0.174	0.05	N_n/N
Nonrespond|Heterosexual (r)	0.03	0.051	0.043	0.04	0.061	0.05	N_n/N
Misreport|Heterosexual (M)	0.013	0.008	0	0.016	0.012	0	–

V. Conclusion

Previous attempts at measuring the incidence of non-heterosexual behaviour have produced highly variable results. The range has been between 1% and 10% according to Badgett (1995), although larger estimates are not unheard of (Kinsey *et al.*, 1948; Sorenson, 1973; Dover, 1978). None of the previous attempts deals effectively with the problems of misreporting and non-response. Against this background of inconsistent empirical findings and questions of methodological reliability, the model in this paper attempts to simultaneously deal with misreporting and non-response within a unified maximum-likelihood framework.

According to the misreporting- and non-response-corrected model, 7.1% of men and 4.1% of women have had at least one same-sex sexual partner. That translates into 15.8 million non-heterosexual Americans: 5.9 ($\pm$1.3) million women and 10.0 ($\pm$1.9) million men based on total US Census Bureau figures from July 2001. More than four million non-heterosexuals go uncounted using the naive model.

The estimates reported above demonstrate interesting patterns that condition probabilities of misreporting and non-response. Among non-heterosexual males, misreporting is most likely for those who have high income, are young, and live in small cities. Among non-heterosexual females, non-response is most likely among those who have low income, are older, and hold more academic degrees. Older heterosexuals are disproportionately represented among non-responders. These relationships suggest that face-value interpretations of survey data concerning sexual behaviour distort the true joint distribution of behavioural variables and their covariates. The data reject formal tests of both strong and weak forms of the missing-at-random hypothesis.

Four assumptions were required to estimate the model. The assumptions effectively require the model to revert to face-value frequencies whenever the independent variables fail to be informative. Near the point in parameter space where the hypothesized covariates x do not influence misreporting and non-response probabilities, the assumptions force the model to produce estimated same-sex frequencies equal to empirical face-value same-sex frequencies. Similarly, the assumptions dictate that predicted non-response probabilities coincide with observed non-response rates whenever the coefficients on x are zero. This symmetry allows the data to decide whether conditioning on x changes the likelihood that an individual will misreport or non-respond. Whether or not to adjust away from face-value interpretations based on self-reported frequencies and, if so, in which direction, depends on the data. Adjustment can occur in either direction.

There are no parameter restrictions preventing the model from estimating lower same-sex frequencies than those derived from naive, or face-value, computations. Thus, the paper's main finding that, after accounting for misreporting and non-response, there are four million more behaviourally non-heterosexual individuals in the USA – fully one-third more than would be estimated otherwise – reflects information derived solely from the data rather than priors built into the model. By simultaneously dealing with misreporting and non-response within an otherwise conventional maximum likelihood framework, the estimators enjoy consistency and asymptotic efficiency, conditional on correct model specification. A similar modelling approach could be applied in other settings where misreporting and non-response are suspected to be significant problems.

Acknowledgements

The authors thank Brian Bucks, Ted Juhl, Joseph Sicilian, and Wim Vijverberg for helpful comments.

References

Allegretto, S. and Arthur, M. (2001) An empirical analysis of homosexual/heterosexual male earnings differentials: unmarried and unequal?, *Industrial and Labor Relations Review*, **54**, 631–46.

Alm, J., Badgett, M. V. L. and Whittington, L. A. (2000) Wedding bell blues: the income tax consequences of legalizing same-sex marriage, *National Tax Journal*, **53**, 201–14.

Arabsheibani, G. R., Marin, A. and Wadsworth, J. (2001) Gays' pay in the UK, Working Paper, University of Wales, Aberystwyth.

Badgett, L. M. V. (1995) The wage effects of sexual orientation discrimination, *Industrial Labor Relations Review*, **48**, 726–39.

Badgett, L. M. V. (1997) Beyond biased samples: Challenging the myths on the economic status of lesbians and gay men, in *Homo Economics: Capitalism, Community and Lesbian and Gay Life* (Eds) A. Gluckman and B. Reed, Routledge, London, pp. 65–72.

Badgett, L. M. V. (2001) *Money, Myths and Change: The Economic Lives of Lesbians and Gay Men*, University of Chicago Press, Chicago.

Baker, S. G. (1995) Marginal regression for repeated binary data with outcome subject to non-ignorable non-response, *Biometrics*, **51**, 1042–52.

Berg, N. (2005) Non-response bias, in *Encyclopedia of Social Measurement*, vol. 2 (Ed.) K. Kempf-Leonard, Academic Press, London, pp. 865–73.

Berg, N. and Lien, D. (2002) Measuring the effect of sexual orientation on income: evidence of discrimination?, *Contemporary Economic Policy*, **20**, 394–414.

Black, D. A., Berger, M. C. and Scott, F. A. (2000a) Bounding parameter estimates with nonclassical measurement error, *Journal of the American Statistical Association*, **95**, 739–48.

Black, D., Gates, C., Sanders, S. and Taylor, L. (2000b) Demographics of the gay and lesbian population in the United States: evidence from available systematic data sources, *Demography*, **37**, 139–54.

Black, D., Gates, C., Sanders, S. and Taylor, L. (2002) Why do gay men live in San Francisco?, *Journal of Urban Economics*, **51**, 54–76.

3 Black, D. A., Makar, H. R., Sanders, S. G. and Taylor, L. (forthcoming) The effects of sexual orientation on earnings, *Industrial Labor Relations Review*.

Bloom, D. E. and Glied, S. (1992) Projecting the number of new AIDS cases in the United States, *International Journal of Forecasting*, **8**, 339–65.

Chauncey, G. (1994) *Gay New York: Gender, Urban Culture and the Making of the Gay Male World, 1890–1940*, Basic, New York.

Collum, C. S. (1993) Co-parent adoptions: lesbian and gay parenting, *Trial*, **29**, 28.

Conaway, M. R. (1992) The analysis of repeated categorical measurements subject to nonignorable non-response, *Journal of the American Statistical Association*, **87**, 817–24.

4 D'Emilio, J. (1989) Gay politics and community in San Francisco since World War II, in *Hidden From History: Reclaiming the Gay and Lesbian Past* (Eds) M. B. Duberman *et al.*, New American Library, New York, pp. 456–73.

Dover, K. J. (1978) *Greek Homosexuality*, Harvard University Press, Cambridge.

Ekholm, A. and Skinner, C. (1998) The muscatine children's obesity data reanalyzed using pattern mixture models, *Applied Statistics*, **47**, 251–63.

Embree, B. G. and Whitehead, P. C. (1991) Validity and reliability of self-reported drinking behaviour: dealing with the problem of response bias, *Journal of Studies on Alcohol*, **54**, 334–44.

Florida, R. (2002) *The Rise of the Creative Class*, Basic Books, New York.

Forster, J. J. and Smith, P. W. F. (1998) Model-based inference for categorical survey data subject to non-ignorable non-response, *Journal of the Royal Statistical Society Ser. B*, **60**, 57–70.

Hausman, J. A., Abrevaya, J. and Scott-Morton, F. M. (1998) Misclassification of the dependent variable in a discrete-response setting, *Journal of Econometrics*, **87**, 239–69.

Heckman, J. J. (1979) Sample selection bias as a specification error, *Econometrica*, **47**, 153–61.

Kinsey, A. C., Pomeroy W. B. and Martin, C. E. (1948) *Sexual Behaviour in the Human Male*, Saunders, Philadelphia.

Kupek, E. (1998) Determinants of item non-response in a large national sex survey, *Archives of Sexual Behaviour*, **27**, 581–89.

Laumann, E. O., Gagnon, J. H., Michael, R. T. and Michaels, S. (1994) *The Social Organization of Sexuality: Sexual Practices in the United States*, University of Chicago Press, Chicago.

Lee, B. and Marsh, L. C. (2000) Sample selection bias correction for missing response observations, *Oxford Bulletin of Economics and Statistics*, **62**(2), 305–22.

Lien, D. and Rearden, D. (1990) Missing measurements in discrete response models, *Economics Letters*, **32**, 231–35.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1994) Weighted least squares analysis of repeated categorical measurement with outcomes subject to non-response, *Biometrics*, **50**, 11–24.

Little, R. J. (1993) Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, **88**, 125–34.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis With Missing Data*, Wiley, New York.

Marquie, J. C. and Baracat, B. (2000) Effects of age, education and sex on response bias in a recognition task, *Journal of Gerontology*, **55B**(5), 266–72.

Michael, R. T., Gagnon, J. H., Laumann, E. O. and Kolata, G. (1994) *Sex in America: A Definitive Study*, University of Chicago Press, Chicago.

Murray, S. O. (1999) *Homosexualities*, University of Chicago Press, Chicago.

Nandram, B. and Choi, J. W. (2002) Hierarchical Bayesian non-response models for binary data from small areas with uncertainty about ignorability, *Journal of the American Statistical Association*, **97**, 381–388.

Newport, F. (2001) American attitudes toward homosexuality continue to become more tolerant, *The Gallup Organization*, <http://www.gallup.com/subscription/?m=f&c_id=10680>, June, 4.

Pearl, D. K. and Fairley, D. (1985) Testing for the potential for non-response bias in sample surveys, *Public Opinion Quarterly*, **49**(4), 553–60.

Plug, E. and Berkhout, P. (2004) Effects of sexual preferences on earnings in the Netherlands, *Journal of Population Economics*, **17**, 117–31.

Rubin, D. B. (1987) *Multiple Imputation for Non-response in Surveys*, Wiley, New York.

Roy, J. and Lin, X. (2002) Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices, *Journal of the American Statistical Association*, **97**, 49–52.

Sorenson, R. C. (1973) *Adolescent Sexuality in Contemporary America*, World Publishing, New York.

Stasny, E. A. (1991) Hierarchical models for the probabilities of a survey classification and non-response: an example from the national crime survey, *Journal of the American Statistical Association*, **86**, 296–303.

Thomas, R. (2001) Estimated population mixing by country and risk cohort for the HIV/AIDS epidemic in Western Europe, *Journal of Geographical Systems*, **3**, 283–301.

Turner, H. A. (1999) Participation bias in AIDS-related telephone surveys: results from the national AIDS behavioural survey (NABS) non-response study, *The Journal of Sex Research*, **36**, 52–66.

Whitehead, J. C., Groothuis, P. A. and Blomquist, G. C. (1993) Testing for non-response and sample selection bias in contingent valuation: analysis of a combination phone/mail survey, *Economics Letters*, **41**, 215–20.

Wu, L. (2002) A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies, *Journal of the American Statistical Association*, **97**, 955–64.

Appendix

The empirical results reported in the body of the paper are derived using modified logistic functional forms as follows:

$$g(\phi' x) = (1 + a_g \mathrm{e}^{-} \phi' x)^{-1} \quad \text{(A1)}$$

$$m(\beta' x) = 0.5(1 + a_m \mathrm{e}^{-} \beta' x)^{-1} \quad \text{(A2)}$$

$$n(\gamma' x) = 0.5(1 + a_n \mathrm{e}^{-} \gamma' x)^{-1} \quad \text{(A3)}$$

$$r(\rho' x) = 0.5(1 + a_r \mathrm{e}^{-} \rho' x)^{-1}, \quad \text{(A4)}$$

where the constants a_g, a_m, a_n and a_r are given by the formulas:

$$a_g = \frac{1 - 2M - N_n/N}{N_g/(N_s + N_g) - M} - 1 \quad \text{(A5)}$$

$$a_m = 1/M - 1 \quad \text{(A6)}$$

$$a_n = N/N_n - 1 \quad \text{(A7)}$$

$$a_r = N/N_n - 1. \quad \text{(A8)}$$

The presence of the term '0.5' arises from requiring the probabilities to lie in the unit interval, explained as follows. The expression $1 - m(\beta' x) - n(\gamma' x)$ is the probability of self-reporting non-heterosexual conditional on non-heterosexual behaviour. Because it is a probability, it must be bounded in the unit interval. The idea that either misreporting or non-response would exceed 50% so wildly contradicts the priors, that is applied a parameterization that bounds m and n from above by 0.50. The upper bound of 0.5 is also imposed on the heterosexual refusal probability r out of symmetry considerations, so that the non-response parameters ρ and γ can be more easily compared.

Alternative specifications that do away with the upper bound restrictions are possible, such as the following:

$$g(\phi' x) = (1 + a_g \mathrm{e}^{-} \phi' x)^{-1} \quad \text{(A9)}$$

$$m(\beta' x, \gamma' x) = a_m \mathrm{e}^{\beta} x/(1 + a_m \mathrm{e}^{\beta} x + a_n \mathrm{e}^{\gamma} x)^{-1} \quad \text{(A10)}$$

$$n(\beta' x, \gamma' x) = a_n \mathrm{e}^{\gamma} x/(1 + a_m \mathrm{e}^{\beta} x + a_n \mathrm{e}^{\phi} x) \quad \text{(A11)}$$

$$r(\rho' x) = (1 + a_r \mathrm{e}^{-} \rho' x)^{-1}. \quad \text{(A12)}$$

This specification ensures that the probability non-heterosexuals correctly report their sexual behaviour $1 - m(\cdot) - n(\cdot)$ is contained in the unit interval for any value of M (in the unit interval) without limiting the range of $m(\cdot)$ or $n(\cdot)$. In practice, this model encounters problems with numerical stability in computing maximum likelihood estimates. Therefore, the analysis proceeds under the specification given by (A1)–(A3).

A second detail worth clarifying is the role of the constants a_g, a_m, a_n and a_r. They serve to centre the respective probabilities at their corresponding empirical face-value frequencies when the slope parameters are zero, in accordance with assumptions 1–4. Those assumptions result in the equations:

$$g(0) = \left(\frac{N_g}{N_s + N_g} - M\right) \Big/ \left(1 - \frac{N_n}{N} - 2M\right) \quad \text{(A13)}$$

$$m(0) = M \quad \text{(A14)}$$

$$n(0) = \frac{N_n}{N} \quad \text{(A15)}$$

$$r(0) = \frac{N_n}{N} \quad \text{(A16)}$$

In finite samples the numerator and denominator in the formula for $g(0)$, Equation A30, may be either negative or positive. To guarantee that $g(0) > 0$, M must be chosen such that one of two inequalities holds:

$$0 < M < \min\left\{\frac{N_g}{N_s + N_g}, \frac{1}{2}\left(1 - \frac{N_n}{N}\right)\right\} \text{ or}$$

$$\max\left\{\frac{N_g}{N_s + N_g}, \frac{1}{2}\left(1 - \frac{N_n}{N}\right)\right\} < M < 1$$

Asymptotically, $g(0)$ approaches the population frequency $\mathrm{plim}[g(\phi' x)] \equiv g$ with respect to sample size N, as the slope parameters approach zero. Therefore, $g(0)$ must be positive in the limit. To see this, denote the plims of $g(0)$, $m(0)$, $n(0)$ and $r(0)$ as g_l, m_l, n_l and r_l, respectively, and use the calculations

$$\mathrm{plim}\left[\frac{N_g}{N_s + N_g}\right] = g_l(1 - m_l - n_l) + (1 - g_l)M,$$

and

$$\mathrm{plim}\left[\frac{N_n}{N}\right] = g_l n_l + (1 - g_l) r_l \quad \text{(A17)}$$

to substitute into the right hand side of Equation A20. Thus,

$$\mathrm{plim}[g(0)] = \frac{g_l(1 - m_l - n_l) + (1 - g_l)M - M}{1 - 2M - g_l n_l - (1 - g_l) r_l} \quad \text{(A18)}$$

The zero slope restriction together with assumptions 1–4 allows the substitution of m_l for M and n_l for r_l, implying that

$$\mathrm{plim}[g(0)] = g_l|_{\phi=\beta=\gamma=\rho=0} \quad \text{(A19)}$$

Together, the four assumptions in equations ([redacted])–([redacted]) imply a functional relationship between 5

$g(0)$, $m(0)$, $n(0)$, and $r(0)$ and M, conditional on the observed number of self-reported heterosexuals and non-heterosexuals, N_s and N_g:

$$g(0) = \left(\frac{N_g}{N_s + N_g} - M\right) \bigg/ \left(1 - \frac{N_n}{N} - 2M\right) \quad \text{(A20)}$$

$$m(0) = M \quad \text{(A21)}$$

$$n(0) = \frac{N_n}{N} \quad \text{(A22)}$$

$$r(0) = \frac{N_n}{N} \quad \text{(A23)}$$

The likelihood function for the parameters $[\phi' \; \beta' \; \gamma' \; \rho']'$ and M, Equation 1 in the body of the paper, has first order conditions as follows:

$$\frac{\partial L}{\partial \phi} = \sum_{i=1}^{N} \left\{ \frac{z_{0i}}{A_{0i}} [-(1 - M - r_i) + m_i] + \frac{z_{1i}}{A_{1i}} [-M + (1 - m_i - n_i)] + \frac{z_{2i}}{A_{2i}} [-r_i + n_i] \right\} \frac{\partial g_i}{\partial \phi} = 0 \quad \text{(A24)}$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{N} \left\{ \frac{z_{0i}}{A_{0i}} - \frac{z_{1i}}{A_{1i}} \right\} g_i \frac{\partial m_i}{\partial \beta} = 0 \quad \text{(A25)}$$

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{N} \left\{ -\frac{z_{1i}}{A_{1i}} + \frac{z_{2i}}{A_{2i}} \right\} g_i \frac{\partial n_i}{\partial \gamma} = 0 \quad \text{(A26)}$$

$$\frac{\partial L}{\partial \rho} = \sum_{i=1}^{N} \left\{ -\frac{z_{0i}}{A_{0i}} + \frac{z_{2i}}{A_{2i}} \right\} (1 - g_i) \frac{\partial r_i}{\partial \rho} = 0 \quad \text{(A27)}$$

where

$$\frac{\partial g_i}{\partial \phi} = \frac{a_g \mathrm{e}^{-\phi' x_i}}{(1 + a_g \mathrm{e}^{-\phi' x_i})^2} x_i \quad \text{(A28)}$$

$$\frac{\partial m_i}{\partial \beta} = \frac{0.5 a_m \mathrm{e}^{-\beta' x_i}}{(1 + a_m \mathrm{e}^{-\beta' x_i})^2} x_i \quad \text{(A29)}$$

$$\frac{\partial n_i}{\partial \phi} = \frac{0.5 a_n \mathrm{e}^{-\gamma' x_i}}{(1 + a_n \mathrm{e}^{-\gamma' x_i})^2} x_i \quad \text{(A30)}$$

$$\frac{\partial r_i}{\partial \rho} = \frac{0.5 a_r \mathrm{e}^{-\rho' x_i}}{(1 + a_r \mathrm{e}^{-\rho' x_i})^2} x_i, \quad \text{(A31)}$$

and

$$A_{0i} = (1 - g(\phi' x_i))(1 - M - r(\rho' x_i)) + g(\phi' x_i) m(\beta' x_i) \quad \text{(A32)}$$

$$A_{1i} = (1 - g(\phi' x_i))M + g(\phi' x_i) \times (1 - m(\beta' x_i) - n(\gamma' x_i)) \quad \text{(A33)}$$

$$A_{2i} = (1 - g(\phi' x_i)) r(\rho' x_i) + g(\phi' x_i) n(\gamma' x_i). \quad \text{(A34)}$$

The estimates presented in the paper are solutions in $[\phi' \; \beta' \; \gamma' \; \rho']'$ and M to the system of equations above. Expressions for the second derivative matrix from which variance estimators were computed is available from the authors upon request.