Rev Econ Household
DOI 10.1007/s11150-008-9038-1

Sexual orientation and self-reported lying

Nathan Berg · Donald Lien

Received: 1 March 2007 / Accepted: 4 June 2008

Abstract This paper examines empirical links between sexual orientation and self-reported lying using data collected in several waves of Georgia Institute of Technology's World Wide Web Users Survey. The data include questions about sexual orientation, lying in cyberspace, and a broad range of demographic information. According to the theoretical framework of Gneezy (Am Econ Rev 95: 384–395, 2005) on the economics of deception, individuals conceal or falsify information when the expected benefit of lying exceeds its costs in terms of psychic disutility. If non-heterosexuals expect to benefit more by falsifying information, then this theory predicts higher rates of lying among non-heterosexuals. The data show that gays and lesbians do indeed report lying more often than heterosexuals, both unconditionally in bivariate correlations and after controlling for demographic and geographic differences. These empirical results are consistent with the conclusion that non-heterosexuals expect higher benefits from concealing personal information because of anti-homosexual discrimination.

Keywords Deception · Sexual orientation · Gay · Misreporting · Non-response

N. Berg (✉)
School of Economic, Political, and Policy Sciences, University of Texas-Dallas,
800 W. Campbell Rd., GR31, Richardson, TX 75083-3021, USA
e-mail: nberg@utdallas.edu; prof.berg@gmail.com

D. Lien
Department of Economics, University of Texas-San Antonio, 6900 North Loop 1604 West,
San Antonio, TX 78249-0633, USA
e-mail: don.lien@utsa.edu

1 Introduction

This paper undertakes an empirical investigation of statistical associations between self-reported sexual orientation and self-reported lying, bringing together two distinct literatures concerning deception (Gneezy 2005; Hurkens and Kartik 2006; Miettinen 2006; Sanchez-Pages and Vorsatz 2006; Demichelis and Weibull 2006; Wang et al. 2006; Fischbacher 2007; Dreber and Johannesson 2008) and the economics of non-heterosexuality (Badgett 2001). A growing literature within economics connects sexual orientation to economic variables such as personal income, household income, geographical location and health outcomes (Badgett 1995; Allegretto and Arthur 2001; Black et al. 2003; Plug and Berkhout 2004; Black et al. 2002; Bloom and Glied 1992; Turner 1999). Much of this literature necessarily relies on survey data, for example, self-reported same-sex sexual contacts reported in the GSS, self-reported number and gender of roommates in U.S. Census data, or self-reported sexual orientation in other surveys.

Relying on self-reports about sexual orientation raises several potential problems. One problem is that definitions of categories such as "gay" are somewhat ambiguous (Murray 1999; Chauncey 1994), with different non-heterosexuals choosing different labels and ascribing distinct meanings to those labels. Another potential problem is that truthful self-reporting about one's sexual orientation may, in some environments, subject survey respondents to risks or other elevated levels of expected costs in the form of discriminatory treatment by employers or others outside the workplace (Berg and Lien 2002). On the other hand, a number of researchers who study dishonesty have put forward the hypothesis that people generally receive positive utility from telling the truth (Gneezy 2005; Mazar and Ariley 2006; Wang et al. 2006; Fischbacher 2007). These two ideas—of deceit responding to expected benefits, and deceit incurring costs in the form of psychic disutility—can be combined and applied to the case of misreporting sexual orientation.[1]

The main prediction that this paper seeks to test is motivated by a straightforward cost–benefit theory of lying (Gneezy 2005) applied to survey respondents' decisions about whether to reveal non-heterosexual sexual orientation. Holding other factors that influence individual propensities to lie constant, the cost–benefit theory predicts that non-heterosexuals are more likely to lie or falsify information whenever the expected benefit of concealing non-heterosexuality is greater than the disutility of lying. In some special environments with high degrees of acceptance of homosexuality, the expected benefit of concealing homosexual behavior might be negligible. In such environments, one would expect those small benefits of concealment to be greatly offset by the psychic benefits of being "out" or openly transmitting information in a way that takes no pains to conceal homosexuality. The expected benefit of concealing homosexuality would also be small in cases where data collectors can persuade survey respondents that their responses will remain confidential and protected with adequate measures to insure data security.

[1] The literature on the psychology of heuristics (Gigerenzer et al. 1999) is rich with alternative models in which decision makers do not weigh costs and benefits at all, but rather employ simple strategies to deal with commonly encountered tasks such as taking a survey or, for homosexuals, choosing whether to conceal sexual orientation depending on contextual cues in the environment.

There would seem to be plenty of evidence, however, to motivate nontrivial expected benefits of concealment and falsification for non-heterosexuals in many parts of the U.S. and beyond, given highly visible condemnation of homosexuality by some politicians and religious leaders in the U.S., lethal violence suffered by homosexual victims of hate crime, and indirect evidence of workplace discrimination (Badgett 1995; Berg and Lien 2002).

Based on the theory that non-heterosexuals, on average, have larger expected benefits from lying about sexual orientation and therefore lie more often than heterosexuals, we investigate the empirical link between lying and sexual orientation by means of an Internet survey collected by a team of researchers at the Georgia Institute of Technology. The advantage of these data is that they contain information about both sexual orientation and self-reported lying. Using a rich set of socio-economic controls collected as part of this survey, we estimate empirical probability models measuring the expected change in the probability of lying as a function of non-heterosexuality. We estimate separate models by gender and allow age to enter the probability-of-lying function with distinct intercepts, slopes and second-order curvature for heterosexuals and non-heterosexuals. This reveals new information about the gay-straight differential in rates of lying and how it changes as a function of age. Given a randomly selected information transmission about any topic, this gay-straight differential in the probability of lying, in turn, enables us to estimate a lower bound and ultimately quantify the much more difficult-to-observe rate at which non-heterosexuals lie specifically about sexual orientation, which is important because it can lead to significant undercounts of non-heterosexual populations.

The next section specifies a simple probability model that provides a lower bound and one-parameter estimator for the probability that non-heterosexuals conceal their sexual orientation, expressed as a function of variables that can be more directly observed. Section 3 describes the empirical model and the theory's predictions about its parameters. Section 4 describes the data, and Sect. 5 presents the main empirical results of gay-straight differentials in the probability of lying. Finally, Sect. 6 concludes with a discussion of the results, their possible interpretations, and implications for future research.

2 Simple model for recovering the rate at which gays[2] lie about sexual orientation

Measuring who is gay, or non-heterosexual, can be difficult. Social stigma may make it awkward for some non-heterosexuals to self-identify as such. And anti-homosexual sentiment among employers and others who are in a position to impose real economic costs on individuals identified as non-heterosexual can rationalize the decision to lie about sexual orientation following straightforward cost–benefit calculus.

[2] As a linguistic shortcut, non-heterosexual sexual orientation is sometimes referred to as "gay" in this paper, while recognizing the subtle nuances of different labels and the important distinction between behavioral labels and identity that are emphasized in other literatures on sexual preference and identity. According to this abbreviated formulation, the category "gay" includes lesbians.

Gneezy (2005) provides a simple model in which decisions about whether to lie follow the cost–benefit calculus under the added assumption that agents experience psychic disutility from lying and therefore choose to incur this cost only when the benefits of lying are sufficiently large. If the psychic disutility from lying is relatively stable across groups, then widely varying rates of lying would imply varying benefits from lying.

The experiments of Gneezy (2005) and Dreber and Johannesson (2008) demonstrate strong empirical links between the experimentally controlled benefits-of-lying variable and individual propensities to lie. Dreber and Johannesson's (2008) experimental results additionally show strong gender effects in propensities to lie, with males demonstrating greater willingness than women to lie to obtain a given benefit, which suggests that men may have systematically lower levels of psychic disutility from lying. With these results, one would predict, in a somewhat-to-severely homophobic society, that the benefit of lying is systematically linked to sexual orientation. But the gay-straight differential in rates of lying should differ by gender if the costs of lying also vary by gender.

If non-heterosexuals face higher expected benefits from lying about sexual orientation, there is the possibility that this would affect propensities to lie in other contexts depending on the shape of the disutility-of-lying curve as a function of quantity of lies. If the marginal cost of a lie is constant and much less variable than the benefit of lying, then observed gay-straight differences in lying decisions would be directly related to the unobserved benefit of lying. If on the other hand the disutility of lying curve is concave or convex in the quantity of lies, the observed gay-straight difference in rates of lying about all topics would consist of two components, one reflecting differential benefits of lying and the second reflecting different quantities of mendacity and consequently different marginal costs of lying.

A number of researchers who study non-heterosexual populations use self-reported sexual orientation without any correction for the possibility of higher-than-average rates of misreporting and non-response among gays and lesbians, raising concern about undercounting the gay and lesbian population. Berg and Lien (2006) use a statistical model that does correct for misreporting and non-response to show that population estimates of the gay and lesbian population in the U.S. are indeed dramatically underestimated if one relies solely on face-value interpretation of self-reported sexual behavior. In the context of this literature and the important sensitivity of this most basic demographic statistic—the frequency of non-heterosexuals in the population—to systematic misreporting among gays, a primary quantity of interest to researchers of non-heterosexual populations is the rate at which non-heterosexuals falsify their sexual orientation. Thus, a key objective of this paper is to provide information relevant for estimating the probability of lying about sexual orientation among non-heterosexuals.

This section develops a simple probability model that relates the observed gay-straight difference in probabilities of lying to websites (about all topics) to the probability of lying about sexual orientation in particular, providing a lower bound on the probability of lying about sexual orientation conditional on being gay. The data introduced later in this paper do not provide a direct measure of lying about sexual orientation. Instead, the survey item measuring self-reported falsification refers to "personal information" and therefore covers a range of topics and information transmissions, only a fraction of which has any chance of revealing sexual orientation.

Denote the unobserved probability of lying about sexual orientation (SO) conditional on non-heterosexual sexual orientation (abbreviated as "gay") as:

$$\Pr(\text{lie about SO}|\text{gay}). \tag{1}$$

The unobserved probability in (1) can be expressed as a function of two probabilities that are directly observable from the data presented in later sections and an unobserved parameter s measuring the fraction of all information transmissions that reveal a writer's sexual orientation, regardless of whether he or she is homosexual. Its complement, $1 - s$, measures the chance that a randomly selected electronic communication does not reveal any information correlated with sexual orientation. Expressing the probability of lying about sexual orientation among gays as a function of s allows us to consider a range of possible values depending on context and assumptions about the frequency of information transmissions that have any chance of revealing information about sexual orientation.

We assume that gays lie more often about sexual orientation than straights, and that straights' rate of lying about their sexual orientation is close to zero:

$$\Pr(\text{lie about SO}|\text{gay}) > \Pr(\text{lie about SO}|\text{straight}) \approx 0. \tag{2}$$

As an agnostic uniform prior, we also assume that the rate of lying about topics not connected to sexual orientation is the same for gays and straights, represented by the parameter L:

$$\Pr(\text{lie about non-SO}|\text{gay}) = \Pr(\text{lie about non-SO}|\text{straight}) \equiv L. \tag{3}$$

Then the probability of lying about any topic in general can be expressed as a weighted average of the probabilities of lying conditional on the topic, either non-SO or SO, with weights equal to s and $1 - s$. Unconditional on whether an information transmission is correlated with sexual orientation or not, the Law of Total Probability yields the following two expressions for the probability of lying among gays and straights:

$$\Pr(\text{lie}|\text{gay}) = \Pr(\text{lie about SO}|\text{gay})s + L(1 - s), \tag{4}$$

and

$$\Pr(\text{lie}|\text{straight}) = \Pr(\text{lie about SO}|\text{straight})s + L(1 - s) \approx L(1 - s). \tag{5}$$

Finally, putting Eqs. (4) and (5) together, the rate of lying about sexual orientation among gays can be computed as:

$$\Pr(\text{lie about SO}|\text{gay}) = [\Pr(\text{lie}|\text{gay}) - \Pr(\text{lie}|\text{straight})]/s. \tag{6}$$

In Eq. 6, the two probabilities on the right-hand side are directly observable from our data. Equation 6 also provides a useful inequality in the form of a lower bound on the probability that gays lie about their sexual orientation, which is equal to the observed gay-straight difference in lying probabilities concerning all topics:

$$\Pr(\text{lie about SO}|\text{gay}) \geq \Pr(\text{lie}|\text{gay}) - \Pr(\text{lie}|\text{straight}). \tag{7}$$

The empirical model introduced in the next section provides a measure of the right-hand side of inequality (7), which is the estimated gay-straight differential in rates of lying (conditional on a vector of other observable information), and can be interpreted as an estimated lower bound on the unobserved probability Pr(lie|gay). By adding one auxiliary assumption about s, this observed gay-straight differential in rates of lying can be easily transformed to an estimate of Pr(lie about SO|gay), by scaling the observed differential up by the factor $1/s$, as given in Eq. 6.

3 Empirical model and predictions

The basic prediction to be tested using the empirical model introduced in this section is whether non-heterosexuals lie at higher rates than non-heterosexuals. This follows from the theory that some fraction of information transmissions does in fact concern sexual orientation and that the expected benefit of lying about sexual orientation is greater for non-heterosexuals and in some cases exceeds the psychic disutility of lying.

The dependent variable is a binary measure of self-reported falsification of information represented by $y_i = 1$ if individual i has ever lied and 0 otherwise. Let x_i represent a vector containing all observable information relevant for predicting whether individual i has ever lied. One key component of x_i based on the theory that gays face higher benefits from lying is the non-heterosexual, or gay, indicator variable G_i. Another important component of x_i is the individual's age A_i, and possible nonlinear functions of age, based on the theory that younger individuals with longer expected durations of remaining life have higher benefits from lying (possibly nonlinear in age, with a decreasing marginal effect of age on lying) and consequently higher rates of lying. Generational effects caused by shifting norms in culture could also play a role but are not necessary for age to predict lying. In the most parsimonious parameterization of the empirical model, nonlinearity is captured by allowing the probability of lying to depend on A_i^2. In keeping with the theory that gays and straights face different benefits from lying about sexual orientation over the life course, both age terms are interacted with sexual orientation. Finally, letting z_i represent all remaining regressors aside from G_i , A_i, A_i^2, and their interactions, the empirical model can be written as:

$$\Pr(y_i = 1|G_i,\ A_i, z_i) = f(\delta_0 G_i + \alpha_1 A_i + \alpha_2 A_i^2 + \delta_1 G_i A_i + \delta_2 G_i A_i^2 + \beta z_i), \tag{8}$$

where all parameters are estimated separately for men and women.

Other control variables in the vector z_i include experience using the Internet, with more Internet-savvy respondents predicted to be more sensitive to the potential for data security problems and therefore more motivated to conceal information. Education controls are needed so that gay-straight differences in levels of education do not lead to spurious gay-straight differences in lying that should be attributed to education and experiences that take place as a result of education, such as socialization in the use of computers and survey taking. Occupational controls are similarly useful because of anecdotal evidence that gays select disproportionately into particular industries, and because working in a job that requires a high degree of familiarity with computer and Internet-related technology is likely to sensitize workers to potential security problems that leads to concealment motives that are distinct from workers in other sectors of the economy.

Because respondents' operating systems are recorded in the survey data and the choice of a non-Windows operating system such as Macintosh OS or Linux might correlate with differential attitudes about data sharing, vulnerabilities in data security and other factors that could influence the expected benefit of lying, this variable is included in the empirical model and predicted to correlate positively with lying. Disability is a demographic variable that, surprisingly, correlates positively with non-heterosexual sexual orientation in these survey data and is therefore included as a control. Finally, linguistic, ethnic and geographic controls are included to capture different cultural attitudes toward non-heterosexuality and lying, which are likely to vary by ethnicity and place. When the empirical model is estimated using pooled data collected at various points in time over several waves of the survey, point-in-time or survey-wave fixed effects are added to control for rapid cultural change, current events and news reporting about anti-homosexual violence, legislative initiatives of special interest to the homosexual population, and innovations in Internet security that might influence lying behavior.

In the empirical work reported in the next section, the functional form of the probability function $f(X)$ (where X is a linear index based on all the variables in the model) is specified as the linear probability model, $f(X) = X$, despite its well-known drawbacks. Among these drawbacks are the possibility of predicted probabilities outside the unit interval and heteroskedasticity. The linear probability model has advantages, however, because estimated coefficients can be directly interpreted as differences in the probability of lying, which are easier to interpret than odds ratios when several interaction terms are included in the model (Ai and Norton 2003). There are numerous empirical papers that adopt the linear probability model because of its advantages of straightforward interpretation of interaction terms and estimation of fixed effects models,[3] both of which apply to the model we use. A key advantage of the linear specification is that marginal effects, even with several interaction terms, are not dependent on the mean values of all regressors as is true in the logit and probit models. Heteroskedasticity is taken care of with a more general error structure (using STATA's "robust" option) that produces more conservative standard errors controlling for predictable differences in variance and model mis-specification. Regarding illogical predicted probabilities outside the unit interval, the linear probability model is least likely to produce expected probabilities outside the unit interval for proportions that are not too close to 0 or 1, which is the case with our data in which the mean rate of lying is 40%. And Cox (1970) reports that, in the range 0.2–0.8, linear, probit, and logit models all give similar predictions, as is the case with our data. The qualitative findings

[3] Grignon et al. (2006) use the linear probability model to estimate the effect of free health care programs on the probability of utilizing healthcare, arguing that the interpretation of fixed effects is superior to that obtained by logit or probit models, despite the disadvantages of heteroskedasticity and probability estimates that can lie outside the unit interval. Similarly, Drago et al. (2008) argue that measurement of fixed effects (or dummy variables) can be better accomplished with the linear probability model than with logit and probit models.

of our model are replicated in Appendix A with a re-estimated empirical model using a logistic specification of *f*(*X*).

In the specification given in Eq 8, the null hypothesis of no difference between gays' and straights' rates of lying is H_0: $\delta_0 = \delta_1 = \delta_2 = 0$. The Results section reports *P*-values for this test computed separately for men and women.

4 Data

The data analyzed in this paper come from three waves of the Georgia Tech WWW User Survey collected in April 1997, October 1997, and April 1998. A total of ten waves were collected between 1994 and 1998[4]. However, only the seventh, eighth and ninth waves provide consistent sample items concerning sexual orientation, self-reported lying and the other regressors in the empirical model. The sample was restricted to adult respondents, aged 19 and up, so that most respondents would have already had the opportunity to finish high school.

Table 1 reports means for the dependent variable and all regressors in the empirical model broken out by gender, sexual orientation, and self-reported lying. The columns in Table 1 labeled "self-reported liar" are based upon the dependent variable, a survey item that asks survey respondents whether they have lied while entering information online that is requested or required by websites.[5] The wording suggests a context of information transmission that excludes chat, email, and other direct person-to-person online tools for dating and coupling.

The first row of Table 1 shows unconditional rates of self-reported lying of 43% among men and 32.5% among women. The non-heterosexual columns indicate rates of lying for non-heterosexual men of just over three percentage points higher than heterosexual men and, for non-heterosexual women, more than 11 percentage points higher than heterosexuals. The second row in Table 1 is the first independent variable in the empirical model, an indicator variable for non-heterosexual status that shows rates of non-heterosexuality of between 8 and 9% for both men and women.[6] Non-heterosexuals appear to be, at least unconditionally, fairly evenly

[4] See http://gvu.cc.gatech.edu/what/websurveys.php for details.

[5] The Wave-7, Wave-8 and Wave-9 sample item asked: "Some websites ask for you to register with the site by providing personal information. When asked for such information, what percent of the time do you falsify the information?" The response choices were: "I've never falsified information," "Under 25% of the time," "26–50% of the time," "51–75% of the time," or "over 75% of the time." From this list of valid responses, a binary variable measuring self-reported lying was constructed. All estimated models reported below use two fixed effect dummies for Wave-8 and Wave-9 survey respondents and are reproducible using ordered categorical probability models that utilize the rest of the measurable variation in self-reported lying but suffer from the disadvantage of more cumbersome marginal effects.

[6] Non-heterosexual sexual orientation is a binary indicator equal to one for individuals who self-report their sexual orientation as something other than heterosexual. The survey item states: "*Note: Although this is a sensitive question, the answer can help Internet developers to understand the needs of current Web users. It is not intended to offend.* How would you classify yourself?" Valid responses are: "None of your business!," "Heterosexual," "Gay Male," "Lesbian," "Bisexual" and "Transgender." Any of the last four of these valid responses maps into the category "non-heterosexual." Around 5% said that sexual orientation was "None of your business!" These non-responders were eliminated from the sample and are not considered outside of Appendix B, which discusses the empirical correlates of item non-response.

Table 1 Mean values broken out by gender, sexual orientation, and self-reported lying

	Men				Women			
	All	Non-Heterosexual	Self-reported liar	Self-reported non-liar	All	Non-Heterosexual	Self reported liar	Self-reported non-liar
Self-reported liar	0.430	0.462	1.000	0.000	0.325	0.437	1.000	0.000
Non-Heterosexual	0.083	1.000	0.090	0.079	0.087	1.000	0.117	0.072
Age	37.0	36.3	32.7	40.2	36.5	34.2	33.2	38.1
Age-Squared	1517.7	1446.5	1173.2	1777.3	1463.5	1278.5	1207.6	1586.7
Age × Non-Heterosexual	3.0	36.3	3.0	3.1	3.0	34.2	3.6	2.7
Age-Squared × Non-Heterosexual	120.6	1446.5	107.5	130.4	111.0	1278.5	120.5	106.4
Years On Internet	2.2	2.2	2.5	2.1	1.8	2.2	2.1	1.7
Completed High School	0.996	0.995	0.996	0.996	0.997	0.994	0.997	0.997
Some College	0.926	0.922	0.934	0.919	0.919	0.943	0.938	0.911
Completed College	0.572	0.568	0.586	0.562	0.535	0.571	0.579	0.514
Completed Graduate Degree	0.235	0.226	0.219	0.248	0.198	0.223	0.212	0.192
Works in Software or Computer Biz	0.323	0.317	0.395	0.269	0.206	0.292	0.255	0.182
Works in Education	0.165	0.165	0.188	0.149	0.218	0.216	0.232	0.211
Works as a Manager	0.133	0.129	0.114	0.148	0.104	0.073	0.094	0.108
Works as Other Professional	0.229	0.230	0.201	0.251	0.237	0.229	0.235	0.238
Non-Windows OS User (Mac/Linux)	0.295	0.293	0.322	0.275	0.175	0.244	0.195	0.165
Disabled	0.064	0.081	0.051	0.074	0.070	0.096	0.065	0.072
Household Income in $1000 units	55.7	50.2	53.8	57.1	50.3	47.4	49.8	50.5
Native Language Non-English	0.086	0.072	0.099	0.077	0.030	0.037	0.038	0.027
Asian	0.032	0.027	0.036	0.029	0.019	0.017	0.024	0.016
Black	0.013	0.014	0.012	0.013	0.024	0.022	0.024	0.023
Hispanic	0.026	0.028	0.023	0.028	0.020	0.021	0.021	0.019

Table 1 continued

	Men				Women			
	All	Non-Heterosexual	Self-reported liar	Self-reported non-liar	All	Non-Heterosexual	Self reported liar	Self-reported non-liar
Latino but not Hispanic	0.009	0.010	0.007	0.010	0.005	0.004	0.007	0.005
Indigenous Person	0.004	0.006	0.005	0.004	0.004	0.005	0.005	0.004
Other Race	0.017	0.017	0.019	0.016	0.018	0.038	0.019	0.018
Africa	0.004	0.007	0.004	0.004	0.001	0.000	0.001	0.002
Asia	0.012	0.016	0.015	0.010	0.005	0.007	0.007	0.005
Canada	0.058	0.048	0.060	0.057	0.046	0.049	0.046	0.045
Central America	0.001	0.001	0.000	0.002	0.001	0.000	0.000	0.001
Europe	0.079	0.052	0.101	0.062	0.029	0.028	0.037	0.025
Middle East	0.003	0.003	0.003	0.002	0.003	0.004	0.004	0.002
Oceania	0.030	0.027	0.029	0.030	0.021	0.030	0.027	0.019
South America	0.008	0.007	0.005	0.010	0.001	0.001	0.000	0.002
West Indies	0.001	0.001	0.001	0.001	0.001	0.000	0.001	0.001
8th-Wave Survey	0.226	0.266	0.216	0.234	0.258	0.243	0.266	0.254
9th-Wave Survey	0.275	0.284	0.300	0.256	0.319	0.330	0.346	0.306
Sample Size	18597	1550	7993	10604	9354	812	3040	6314

distributed across self-reported-liar and non-liar columns, with slightly more non-heterosexuals in the liar category, especially among women.

The next regressor is the row labeled Age, which shows a significant age difference between self-reported liars and non-liars. The average male self-reported liar is 7.5 years younger than the average male self-reported non-liar, and the average self-reported female liar is almost 5 years younger than the average female non-liar. This large bivariate relationship between age and rates of lying will play an important role in our effort to untangle age effects from sexual orientation, because non-heterosexuals in our sample also tend to be younger. It is worth noticing in the row labeled Age that the age differential between liars and non-liars is much larger than between non-heterosexuals and heterosexuals (given approximately by the column under the heading "all," because non-heterosexuals have only a small effect on the overall mean). Following a standard modeling tradition in labor economics of parsimoniously capturing nonlinear effects of age using a quadratic function of age, the variable Age-Squared is included in the regression along with interaction terms between Age and Non-Heterosexual, and Age-Squared and Non-Heterosexual. These interaction terms effectively allow the data to estimate separate nonlinear curves relating age to the probability of lying for gays and straights.

The variable Years On Internet is a categorical variable indicating ranges for how long respondents have been using the Internet. These categorical values map into actual time durations as follows: 0 corresponds to "less than six months;" 1 corresponds to "six to 12 months;" 2 corresponds to "one to three years;" 3 corresponds to "four to six years;" and 4 corresponds to "seven years or more." The average respondent is in the "one to three years" category, and self-reported liars have considerably more online experience than non-liars. Transforming this variable to actual years using the midpoint method reveals grand and within-sample means that are nearly identical.

Table 1 contains four educational attainment variables that are nested such that, for example, a respondent who completed college (i.e., Completed College = 1) will also have values of 1 for all lower-level educational attainment variables (i.e., Completed High School = 1 and Some College = 1). This allows us to see the marginal effects, if any, of degree completion on the probability of lying. The reference category is respondents with no high school degree. Including this reference category (not shown in Table 1), the five education variables are exhaustive and mutually exclusive, with no clear differences between liars and non-liars except that college graduates appear to lie at slightly higher rates. Interestingly, female non-heterosexuals appear to have more education than other women, which is not true of male non-heterosexuals.

Four occupation control variables are included in Table 1 to control for the possibility of systematically different attitudes toward lying and online data collection. Confirming predictions, those who work in the computer or software-related industries are noticeably more likely to lie. Those in the education industry are also slightly more likely to lie. On the other hand, workers in the occupational categories of Managers and Other Professionals are less likely to admit to lying. Interestingly, users of non-Windows operating systems, such as Apple OS or Linux, are slightly more likely to falsify information, and respondents with disability are slightly less likely to lie.

Household Income, measured in units of $1000, is a transformed categorical variable based on a sample item that elicits annual household income in the following seven categories: less than $10,000, $10,000 to $19,000, $20,000 to $29,000, $30,000 to $39,000, $40,000 to $49,000, $50,000 to $74,000, $75,000 to $99,000, or over $100,000. The transformed variable uses the midpoint method for the first six (bounded) income brackets. The seventh category, "over $100,000," is represented by $112,500, the result of adding the size of largest bounded range (i.e., $25,000) to the largest midpoint (i.e., $82,500). The midpoint method for transforming bracketed income data into dollars is known to be problematic, because all estimates in the model are potentially affected by the arbitrary choice of number to represent the largest unbounded bracket. The empirical model estimates (reported below) were re-estimated with the categorical income variable, and with dummy variables for income brackets, which did not substantively change any measured effects of non-heterosexuality on the probability of lying. For ease of interpretation, the dollar-transformed Household Income measure is reported. Interestingly, Table 1 shows that non-heterosexuals have significantly lower levels of household income, on the order of $5,500 for men and $3,000 for women.

Those who list their "primary language" as any language other than English are indicated by the variable Native Language Non-English, and these respondents have slightly higher rates of self-reported falsification. The ethnic distribution among survey participants is clearly nonrepresentative of the U.S. and nonrepresentative of the world in general, with more than 90% of respondents reporting their ethnicity as white. Since white ethnicity is the reference category, its percentage can be computed as one minus percents Asian, Black, Hispanic, Latino but not Hispanic, Indigenous Person, and Other Race. This suggests that, at least in the late 1990s when the survey was conducted, the population of volunteer online survey respondents was mostly white.[7] The geographic indicators take U.S. as the reference category (percent U.S. can be computed from Table 1 as one minus the sum of all the mean values among the geographical indicators listed there). Eighty percent of male respondents and 89% of female respondents live in the U.S. Slightly higher rates of non-heterosexuality are observable among Canadian and European men, and women in the Other Race category. Overall, no large correlations between ethnicity and geography, on the one hand, and lying on the other, are present in the data, with the possible section of the large representation of European men among self-reported liars.

The two rows in Table 1 labeled 8th-Wave Survey and 9th-Wave Survey contain mean values for indicators that capture any fixed effects resulting from differences in the formatting or timing of different survey waves. The empirical model is

[7] Another, potentially more troubling, possibility is that respondents in the Georgia Tech Survey were disproportionately white relative to the broader population of volunteer online survey respondents. We have no data to reliably cross-validate the demographics in our sample against reliable population characteristics of online survey respondents. Broader interpretations of our model's results depend on the maintained assumption of representativeness, defined narrowly with respect to the population of online survey respondents.

estimated using pooled data from three waves, collected during 1997 and 1998, with sample-wave fixed effects to control for possible time-specific or survey-specific influences on rates of lying. Time might influence lying on the Internet because of rapidly changing cultural attitudes and technological change, for example, news reports of employers monitoring workers' email communication, identity theft, and other breaches of data security that could change levels of awareness about privacy on the Internet through time.

Next, we estimate the conditional probability of self-reported lying as a function of sexual orientation while controlling for the variables summarized above.

5 Results

Table 2 reports the empirical model in Eq 8 estimated separately for male and female subsamples under the column heading "Model 1." The same model—but with an additional indicator variable for respondents who were ever married (i.e., currently married, divorced, or widowed) and an interaction of ever married with non-heterosexual status—is reported under the column heading "Add Marital." The first five rows of Table 2 contain coefficients on Non-Heterosexual, Age, Age-Squared, Age × Non-Heterosexual, and Age-Squared × Non-Heterosexual, which jointly determine the probability-of-lying curves as a function of Age.

In Model 1, the gay-straight differential in the probability of lying is determined by the three coefficients, Non-Heterosexual, Age × Non-Heterosexual, and Age-Squared × Non-Heterosexual, which must be jointly tested to decide whether gays and straights have statistically distinguishable probabilities of lying. Although the *t*-statistics on the individual coefficients are small in the male sample, the data reject the null that all three of them are zero, with *p*-values of 0.043 among men and 0.000 among women in Model 1. This implies a statistically significant gay-straight differential in the probability of lying. After the marital variables are added (reported under the column heading "Marital Added"), the coefficients on the four variables which depend on non-heterosexual status are jointly significant among women but not among men. In the Add Marital model, the variable EverMarried reveals a large, negative effect of having ever been married on the probability of lying.

The difficulty in interpreting such coefficients on marital status in models with sexual orientation, however, is the large degree of negative correlation between marital status and sexual orientation. Although this negative correlation is not as perfect as some might speculate, it is large enough to warrant further empirical investigation. Around 19% of gay men in the sample were ever married (computed as 288 gay men who responded that they were married, divorced or widowed, out of a total of 1550), compared with 60% of straight men. Similarly, 30% of lesbian women (245 out of 812) were ever married, compared with 60% of straight women.

Parsing the effect of household structure (e.g., whether coupled in a long-term cohabitating relationship) apart from non-heterosexuality is difficult. With more data on coupling and cohabitation, one would hope to better identify the effect of these and other fundamental variables that make up the structure of the household.

Table 2 Linear probability models of self-reported lying

	Men				Women			
	Model 1*		Add Marital		Model 1		Add Marital	
	Coeff	*t*	Coeff	*t*	Coeff	*t*	Coeff	*t*
Non-Heterosexual	−0.085	−0.7	−0.027	−0.2	0.351	2.1	0.374	2.2
Age	−0.028	−16.5	−0.024	−13.6	−0.017	−6.7	−0.014	−5.4
Age-Squared	0.000	10.3	0.000	8.5	0.000	4.0	0.000	3.1
Age × Non-Heterosexual	0.005	0.8	0.001	0.2	−0.010	−1.2	−0.012	−1.3
Age-Squared × Non-Heterosexual	0.000	−0.6	0.000	−0.1	0.000	0.6	0.000	0.7
EverMarried			−0.051	−5.4			−0.040	−3.3
EverMarried × Non-Heterosexual			0.016	0.5			0.006	0.2
Years On Internet	0.053	14.4	0.052	14.3	0.042	8.7	0.040	8.4
Completed High School	−0.093	−1.7	−0.091	−1.6	−0.129	−1.5	−0.126	−1.4
Some College	0.021	1.5	0.020	1.4	0.018	1.0	0.016	0.9
Completed College	0.023	2.8	0.022	2.6	0.024	2.1	0.022	1.9
Completed Graduate Degree	−0.001	−0.1	0.000	0.0	0.034	2.4	0.032	2.3
Works in Software or Computer Biz	0.059	4.9	0.059	4.9	0.060	4.0	0.058	3.8
Works in Education	−0.010	−0.8	−0.012	−0.9	−0.020	−1.3	−0.023	−1.5
Works as a Manager	0.019	1.5	0.021	1.6	0.010	0.6	0.007	0.4
Works as Other Professional	0.003	0.3	0.004	0.4	0.010	0.7	0.007	0.5
Non-Windows OS User (Mac/Linux)	−0.001	−0.1	−0.002	−0.2	0.002	0.2	0.002	0.1
Disabled	−0.002	−0.2	−0.002	−0.2	0.006	0.3	0.006	0.4
Household Income in $1000 units	0.002	0.9	0.000	1.6	0.002	0.8	0.000	1.2
Native Language Non-English	−0.007	−0.4	−0.009	−0.5	0.017	0.5	0.017	0.5
Asian	−0.049	−2.2	−0.052	−2.4	0.027	0.7	0.025	0.6
Black	−0.027	−0.9	−0.031	−1.0	−0.007	−0.2	−0.012	−0.4

Table 2 continued

	Men				Women			
	Model 1*		Add Marital		Model 1		Add Marital	
	Coeff	*t*	Coeff	*t*	Coeff	*t*	Coeff	*t*
Hispanic	−0.030	−1.1	−0.029	−1.1	−0.020	−0.5	−0.022	−0.6
Latino but not Hispanic	−0.019	−0.4	−0.021	−0.5	0.081	1.1	0.081	1.1
Indigenous Person	0.102	1.9	0.106	2.0	0.105	1.4	0.105	1.4
Other Race	0.013	0.5	0.012	0.4	−0.031	−0.8	−0.033	−0.9
Africa	−0.040	−0.8	−0.037	−0.7	−0.193	−1.9	−0.187	−1.9
Asia	0.063	1.8	0.065	1.8	0.053	0.8	0.056	0.8
Canada	0.010	0.7	0.006	0.4	−0.006	−0.3	−0.008	−0.4
Central America	−0.364	−5.5	−0.365	−5.5	−0.154	−1.1	−0.168	−1.3
Europe	0.052	3.0	0.046	2.7	0.021	0.6	0.015	0.5
Middle East	0.056	0.9	0.063	1.0	0.149	1.5	0.148	1.4
Oceania	−0.034	−1.7	−0.038	−1.9	0.067	2.0	0.063	1.9
South America	−0.160	−3.7	−0.161	−3.8	−0.321	−3.4	−0.327	−3.5
West Indies	−0.019	−0.2	−0.020	−0.2	0.162	0.9	0.159	0.8
8th-Wave Survey	0.021	2.4	0.020	2.3	0.043	3.7	0.042	3.6
9th-Wave Survey	0.062	7.4	0.061	7.3	0.050	4.4	0.049	4.4
Constant	1.047	15.9	0.993	15.0	0.731	7.3	0.697	6.9
Adjusted R^2	0.1254		0.1236		0.0704		0.0716	
Unconditional rate of self-reported lying	0.430				0.325			
Sample Size	18,597				9,354			

* In Model 1, the *p*-value for the joint test that the three coefficients Non-Heterosexual, Age × Non Heterosexual and Age-Squared × Non-Heterosexual are all zero is 0.043 among men and 0.000 among women. In the Add Marital model (with information about marrital status), these three coefficients remain jointly statistically significant among women but not for men

Without such information, data that identify non-heterosexuality as cohabitating same-sex non-kin fail to count single homosexuals. Another problem is that data relying on self-reported sexual orientation are difficult to combine consistently with data on marital status, since the equivalent information about coupling and household structure is missing for homosexuals.

Apart from the very different joint effects of sexual orientation and age across gender evident in Table 2, the remaining variables in the empirical model that are statistically significant all point in the same direction and are of similar magnitudes for men and women. The row labeled Years On Internet shows large effects (of similar magnitude) of experience on the propensity to lie, with an extra 3 years of online experience leading to an extra 4 or 5 percentage points in the probability of lying. Coefficients on the successively nested educational attainment dummies show a statistically significant increase in the probability of lying only among college graduates; lower- and higher-level degree completion have little effect. Those employed in computer and software industries are roughly six percentage points more likely to lie, consistent with the theory that greater familiarity with potential shortcomings in data security—and perhaps the strategic intent of those in the business of data collection—tends to raise levels of suspicion and noncompliance in online requests for information.

Although statistically insignificant, the size of the household income coefficient is large enough to reach levels of economic significance. A household income difference of $10,000 predicts an increase in the probability of lying of 2 percentage points, and a household income difference of $100,000 predicts an increase in the probability of lying of 20 percentage points. The ethnic coefficients are difficult to interpret in isolation because they are highly correlated in some cases with geographical variables. Indigenous persons of both genders admit to lying at a rate of 10 percentage points higher than average. Central and South Americans have much lower-than-average rates of lying. The rate of lying appears to have been increasing over the year and a half during which the three survey waves were collected, as indicated by positive survey-wave coefficients.

To compute the magnitude of the gay-straight differential in expected probabilities of lying, one must plug in particular values for the Age variable, as in Fig. 1, which shows the probability of lying over the entire age range for gay men, straight men, lesbian women and straight women. Figure 1 reveals that the higher rate of lying among non-heterosexuals is critically connected to life-course. Gay men tell lies at similar rates to straight men while in their 20s (because young straight men apparently have lie often for other reasons) but lie at rates 5–8 percentage points higher at ages over 50. In contrast, while in their 20s, lesbian women's rate of lying is 12–15 percentage points higher than straight women's, but hardly distinguishable from other women at ages over 50. All four subgroups—gay men, straight men, lesbian women and straight women—lie at higher rates when young and lower rates with increasing age. In contrast, the gay-straight differential grows from small to large as a function of age in the case of men, but shrinks from large to small among women.

The gay-straight differentials in the probability of lying observable in Fig. 1, which refer to lying about any topic in general on an online survey, imply empirical

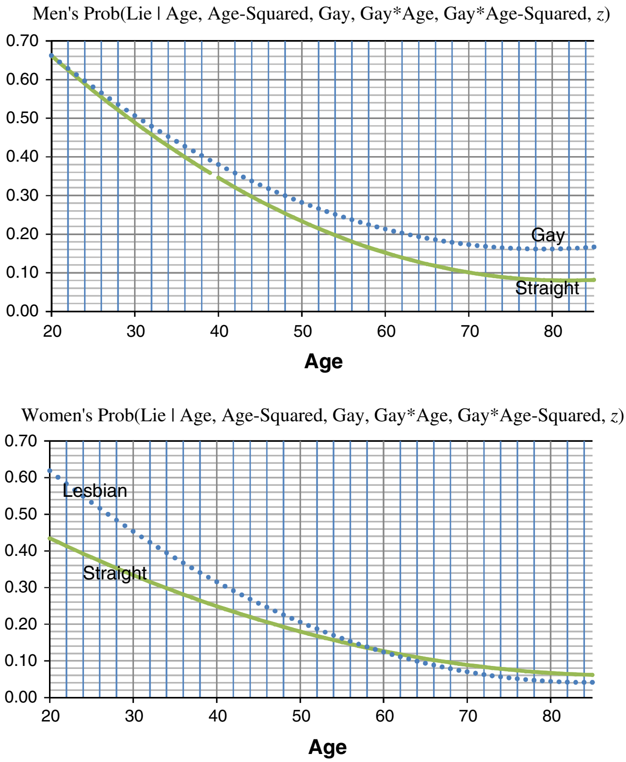

Fig. 1 Probability of lying as a quadratic function[8] of age and sexual orientation, by gender

bounds on the probability that gays lie in particular about sexual orientation. Recalling from Eq. (6) that the probability of lying about sexual orientation among gays can be expressed as $1/s$ times the gay-straight differential in the probability of lying about topics in general, a range of magnitudes can be simulated for the unobserved rate at which gays conceal sexual orientation, as a function of age together with assumptions one wishes to impose on s. (Recall that s measures the fraction of information transmissions that contain information correlated with sexual orientation, which therefore could be used as a predictor to help reveal who is gay.) Somewhat counterintuitively, the most conservative estimate of the rate at which gays lie about their sexual orientation corresponds to the most inclusive assumption about s, namely that all information transmissions can be used to predict sexual orientation and therefore that $s = 1$. In that case, gay men's rate of concealing their sexual orientation would be largest among gay men over 50, in the range of 5–8%. For lesbian women, the rate of concealing sexual orientation would be largest when they are in their 20s, in the range of 12–15%. If instead one assumes

[8] The test statistic for the null hypothesis that gays and straights have identical regressions (of lying on Age and Age-squared while controlling for all other variables in the model presented in Table 2) is distributed as $F(3, 18591)$ for men and $F(3, 9348)$ for women, with observed p-values of 0.044 and 0.000, respectively.

that only half of information transmissions contain information that correlates with sexual orientation, then older gay men's rate of misreporting their sexual orientation would be in the 10–16% range, and young lesbian women's misreporting rates would be 24–30%. If one assumes that $s = 0.20$, then older gay men are predicted to lie about their sexual orientation 25–40% of the time and young lesbian women at a rate of 60–75%.

6 Conclusion

The data show a positive association between self-reported non-heterosexuality and self-reported lying. The magnitude of the gay-straight differential in rates of lying is very large for women, especially young women, and more modest among men. Men and women of both sexual orientations lie less often as they grow older, but the gay-straight differential in the probability of lying grows with age among men and shrinks among women. This positive correlation between non-heterosexuality and the propensity to lie is consistent with a cost–benefit theory of lying in a world where non-heterosexuals can benefit by avoiding costs imposed by others on those with minority sexual orientations. If good proxies for the expected benefit of concealing homosexuality were available (e.g., the inverse of Florida's (2002) index of openness toward homosexuality), then we would expect to see much larger gay-straight differences in rates of lying in high-benefit-of-lying environments (i.e., places with high intolerance of homosexuals). The estimates in this paper should be interpreted as an average across individuals in many different expected-benefit environments.

The simple probability model in Sect. 2 provides a multiplicative transformation of the observed gay-straight differential in the probability of lying about topics in general to the more specific probability of lying about sexual orientation. The gay-straight differential in the probability of lying about topics in general ranges from zero to eight percent among men, and 0 to 15% among women, depending on age. To convert these differentials into estimates of the probability that gays lie about their sexual orientation requires multiplication by the factor $1/s$, where s is the fraction of information transmissions that are possibly correlated with sexual orientation. For example, the question, "Are you married?," should be interpreted as part of the fraction of information transmissions that belong to s. With s as small as 20%, predicted probabilities of non-heterosexuals lying about their sexual orientation can be well over 50%, depending on age and gender.

From a policy point of view, there is a possible disconnect between the average effects based on our regression analyses and the most pressing issues of concern to non-heterosexual individuals and households (e.g., Alm et al. 2000). In particular, non-heterosexuals suffering the most unhappy consequences of anti-homosexual discrimination, where policy is perhaps most needed to help achieve norms of equality and nondiscrimination, would be in the upper tail of the benefit-of-lying distribution and consequently the probability-of-lying distribution as well. If this distribution is highly skewed, with most of the gay population located in relatively open, urban environments, then averages might not provide accurate descriptions of the subsets of non-heterosexuals most likely to conceal their sexual orientation. In particular, mean rates of misreporting non-heterosexual sexual orientation might appear rather low even while self-reporting problems are severe in areas where policy analysts and activists are most interested in advancing nondiscrimination protections. In any case, the contribution of this paper is aimed at more correctly counting the non-heterosexual population by quantifying the rate at which non-heterosexuals conceal sexual orientation.

One unambiguous conclusion is that the possibility of differential rates of misreporting by gender and sexual orientation should probably be considered in future empirical research concerning non-heterosexual populations. The bottom line for frequency estimates of a relatively rare type such as non-heterosexuality is that systematic misreporting can play a very large role. Positive associations between non-heterosexuality and lying in the survey data reported here imply that existing estimates based on face-value interpretations of survey data are likely to underestimate the true frequency of non-heterosexual behavior, especially when expected benefits of concealing non-heterosexuality are large.

Appendix A

Logit model of self-reported lying

	Men		Women	
	Coeff	*t*	Coeff	*t*
Non-Heterosexual	−0.255	−0.5	1.326	1.4
Age	−0.104	−11.2	−0.065	−4.5
Age-Squared	0.001	5.5	0.000	1.8
Age × Non-Heterosexual	0.013	0.5	−0.036	−0.6
Age-Squared × Non-Heterosexual	0.000	−0.2	0.000	0.2
Years On Internet	0.245	14.1	0.205	8.5
Completed High School	−0.443	−1.7	−0.659	−1.6
Some College	0.100	1.5	0.105	1.1
Completed College	0.100	2.6	0.110	2.0
Completed Graduate Degree	−0.006	−0.1	0.165	2.5
Works in Software or Computer Biz	0.267	4.6	0.290	3.9
Works in Education	−0.027	−0.4	−0.073	−1.0
Works as a Manager	0.105	1.6	0.067	0.8
Works as Other Professional	0.031	0.5	0.066	0.9
Non-Windows OS User (Mac/Linux)	−0.005	−0.1	0.010	0.2
Disabled	−0.016	−0.2	0.024	0.3
Household Income in $1000 units	0.001	0.9	0.001	0.7
Native Language Non-English	−0.036	−0.5	0.074	0.5
Asian	−0.222	−2.3	0.115	0.7
Black	−0.124	−0.9	−0.031	−0.2

Appendix continued

	Men		Women	
	Coeff	*t*	Coeff	*t*
Hispanic	−0.135	−1.1	−0.087	−0.5
Latino but not Hispanic	−0.109	−0.5	0.362	1.1
Indigenous Person	0.458	1.9	0.494	1.5
Other Race	0.057	0.5	−0.148	−0.8
Africa	−0.183	−0.8	−1.087	−1.4
Asia	0.285	1.8	0.252	0.8
Canada	0.047	0.7	−0.025	−0.2
Central America	−2.123	−2.9	−0.919	−0.8
Europe	0.226	3.0	0.089	0.6
Middle East	0.249	0.9	0.664	1.4
Oceania	−0.159	−1.7	0.314	2.1
South America	−0.776	−3.5	−2.037	−1.9
West Indies	−0.118	−0.3	0.776	0.9
8th-Wave Survey	0.093	2.3	0.206	3.6
9th-Wave Survey	0.286	7.4	0.242	4.4
Constant	2.084	6.7	0.905	1.9
p-Value for H_0: Non-Hetero coeffs = 0*	0.026		0.000	
Pseudo R^2	0.0957		0.0575	
Unconditional rate of self-reported lying	0.430		0.325	
Sample Size	11,897		9,354	

* The *p*-value is for the joint test that the three coefficients Non-Heterosexual, Age × Non-Heterosexual and Age Squared × Non-Heterosexual are all zero. The test statistic is distributed as $F(3, 18{,}561)$ for men and $F(3, 9318)$ for women

Appendix B

Previous work emphasizes the importance of jointly accounting for misreporting *and* non-response when using data with self-reported sexual orientation (Berg 2005; Berg and Lien 2006). This appendix provides additional detail on how the data were cleaned and how item non-response correlates with important demographic characteristics.

Recall that the data consists of three waves of surveys. These data contain responses from a total of 34,498 individuals aged 19 and older, but only 27,951 provided valid responses to all variables used in the empirical model. The other 6,547 (=34,498 − 27,951) individuals non-responded to at least one sample item, raising the question of systematic inclusion or exclusion from the sample due to correlations between the event of non-response and other variables in the model.

There were two non-response possibilities for the dependent variable regarding the frequency of falsifying information: those who left the item blank, referred to as

"Not Say;" and those who responded that the question was "Not Applicable," perhaps because these responders never faced a website that requested personal information. Eliminating these invalid dependent-variable responses led to 3,913 individuals being dropped. These individuals are not included in any of the reported results here, and the "non-responders" label is defined to refer only to respondents who have a valid dependent variable observation but at least one missing response among the other variables in the empirical model listed in Table 1.

Rates of lying are 0.395 among responders and 0.421 among non-responders, are not very different. Regarding rates of self-reported sexual orientation, only 28.1% of those in the non-responder category non-responded to the sexual orientation sample item. Among the 71.9% of non-responders who did provide a valid response to the sexual orientation sample item (but non-responded to a sample item other than the sexual orientation), $5/71.9 = 7.0\%$ report their status as non-heterosexual, roughly 1 percentage point lower than the responder sample.

The sample item with the highest rate of non-response was Household Income, with 70.1% of non-responders having left this item blank. Average Household Income among non-responders with a valid Household Income response is $42,730, which is significantly lower than mean Household Income among responders. Other interesting correlations are as follows. Those who work in education are more likely to non-respond, as are those whose native language is non-English. Those who report having a disability are more than twice as likely to be non-responders. More than 11% of non-responders refused to identify their ethnicity, although a large majority of both responders and non-responders who did provide ethnicity are white. Few systematic differences in geography between responders and non-responders were apparent, and both groups are roughly 60% male.

As for why non-responders choose to non-respond, there is one survey item that provides some information. Respondents were asked to choose from a list of 15 issues which was the "most important issue facing the Internet." Privacy concerns are noticeably higher among non-responders, with 33% of non-responders versus 26% of responders saying that privacy is the most important issue facing the Internet.

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80*, 123–129. doi:10.1016/S0165-1765(03)00032-6.

Allegretto, S., & Arthur, M. (2001). An empirical analysis of homosexual/heterosexual male earnings differentials: Unmarried and unequal? *Industrial & Labor Relations Review, 54*(3), 631–646. doi:10.2307/2695994.

Alm, J., Badgett, M. V. L., & Whittington, L. A. (2000). Wedding bell blues: The income tax consequences of legalizing same-sex marriage. *National Tax Journal, 53*(2), 201–214.

Badgett, L. M. V. (1995). The wage effects of sexual orientation discrimination. *Industrial & Labor Relations Review, 48*(4), 726–739. doi:10.2307/2524353.

Badgett, L. M. V. (2001). *Money, myths, and change: The economic lives of lesbians and gay men*. Chicago: University of Chicago Press.

Berg, N. (2005). Non-response bias. In Kempf-Leonard, K. (Ed.), *Encyclopedia of social measurement* (vol. 2, pp. 865–873). London: Academic Press.

Berg, N., & Lien, D. (2006). Same-sex sexual behavior: U.S. frequency estimates from survey data with simultaneous misreporting and non-response. *Applied Economics, 38*(7), 757–769. doi: 10.1080/00036840500427114.

Berg, N., & Lien, D. (2002). Measuring the effect of sexual orientation on income: Evidence of discrimination? *Contemporary Economic Policy, 20*, 394–414. doi:10.1093/cep/20.4.394.

Black, D., Gates, C., Sanders, S., & Taylor, L. (2002). Why do gay men live in San Francisco? *Journal of Urban Economics, 51*, 54–76. doi:10.1006/juec.2001.2237.

Black, D. A., Makar, H. R., Sanders, S. G., & Taylor, L. (2003). The effects of sexual orientation on earnings. *Industrial & Labor Relations Review, 56*(3), 449–469. doi:10.2307/3590918.

Bloom, D. E., & Glied, S. (1992). Projecting the number of new AIDS cases in the United States. *International Journal of Forecasting, 8*(3), 339–365. doi:10.1016/0169-2070(92)90052-B.

Chauncey, G. (1994). *Gay New York: Gender, urban culture, and the making of the Gay Male World, 1890–1940*. New York: Basic.

Cox, D. R. (1970). *The analysis of binary data*. London: Methuen.

Demichelis, S., & Weibull, J. W. (2006). Efficiency, communication and honesty, SSE/EFI Working Paper Series in Economics and Finance No. 645.

Drago, F., Galbiati, R., & Vertova, P. (2008). Prison conditions and recidivism, Institute for the Study of Labor (IZA) Working Paper 3395.

Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters, 99*(1), 197–199. doi:10.1016/j.econlet.2007.06.027.

Gigerenzer, G., Todd, P. M., & The ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Grignon, M., Perronnin, M., & Lavis, J. N. (2006). Does free supplementary health insurance help the poor to access health care?: Evidence from France, CHEPA Working Paper, McMaster University.

Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review, 95*(1), 384–394. doi:10.1257/0002828053828662.

Fischbacher, U. (2007). Lies in disguise: An experimental study on cheating, IBZ Working Paper.

Florida, R. (2002). *The rise of the creative class*. New York: Basic Books.

Hurkens, S., & Kartik, N. (2006). (When) would I lie to you? Comment on “Deception: The role of consequences,” Unitat de Fonaments de l’Anàlisi Econòmica (UAB) and Institut d’Anàlisi Econòmica (CSIC) Working Paper.

Mazar, N., Ariely, D. (2006). Dishonesty in everyday life and its policy implications. *Journal of Public Policy and Marketing, 25*(1), 1–25.

Miettinen, T. (2006). Promises and conventions: An approach to pre-play agreements, Working Paper.

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

Plug, E., & Berkhout, P. (2004). Effects of sexual preferences on earnings in the Netherlands. *Journal of Population Economics, 17*, 117–131.

Sanchez-Pages, S., & Vorsatz, M. (2006). Enjoy the silence: An experiment on truth-telling, University of Maastricht Working Paper.

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

Wang, J. T., Spezio, M., & Camerer, C. F. (2006). Pinocchio’s Pupil: Using eyetracking and pupil dilation to understand truth-telling and deception in games, Cal-Tech Working Paper.