methods, data, analyses | Vol. 13(1), 2019, pp. 111-137 DOI: 10.12758/mda.2018.09

The Longitudinal Item Count Technique: A New Technique for Asking Sensitive Questions in Surveys

Alessandra Gaia [1,2] *& Tarek Al Baghal* [2]

[1] *University of Milan-Bicocca*

[2] *Institute for Social and Economic Research, University of Essex*

Abstract

Asking respondents sensitive questions directly may lead to socially desirable responding. As alternative, some have proposed using the Item Count Technique (ICT). The problem with ICT methods is that these can have low statistical efficiency, but also do not provide an indicator of the behavior at the respondent level. We propose a new variant of the ICT to overcome these issues: the Longitudinal Item Count Technique (LICT). Instead of administering different lists (one including the sensitive item and one without) to two random groups in a single survey, the LICT administers both lists to each respondent, but at different survey waves. The sensitive attribute can be estimated as the difference within individuals across waves. Like the ICT, the LICT can be extended to a two-list version. In this paper we discuss the assumptions, implementation, limitations, and ethical implications of this novel technique, and present application of the method in the *Understanding Society* Innovation Panel, estimating the prevalence of the gay, lesbian, and bisexual population in the United Kingdom. In this first application, the LICT in some ways appeared to provide better estimates than the traditional ICT, but also provided some inconsistency in estimates. We discuss the implications of these results and point to routes for further research.

Keywords: Item Count Technique; sensitive questions; social desirability; longitudinal data; LGBT research

The Item Count Technique (ICT) – also called “Unmatched Count Technique” or “List Experiments” – is used to improve the measurement of sensitive topics, reducing social desirability bias. This promising technique protects respondents’ privacy when it works as planned, with no “ceiling” and “floor” effects (i.e. every or no item in the list applying). The ICT, introduced by Smith, Federer and Raghavarao (1974), is an indirect questioning technique to ask sensitive questions in surveys. Instead of inferring the population prevalence of a sensitive behavior by asking respondents directly whether they engaged in that behavior, using the ICT the researcher can extrapolate this information experimentally.

Specifically, in the ICT sample members are randomly divided into two groups; respondents in each group are presented with a list of items and asked to count how many items apply to them. Each group’s list is identical but for the sensitive item appearing only in one of them. Items should be selected such that it is reasonable for respondents to select some but not all items. While ICT methods produce estimates that can be useful in estimating prevalence of sensitive behaviors within subgroups and for regression analyses (see Corstange 2009; Holbrook & Krosnick 2010; Blair & Imai 2010; Imai 2011; Glynn 2013), such results from the ICT are typically imprecise due to low statistical efficiency. The lack of indicators at the respondent-level is also problematic as ICT methods do not allow analyses at the individual-level, but rather at the aggregate-level only.

To overcome these issues, we propose a variation to the ICT: the Longitudinal Item Count Technique (LICT). Instead of splitting the sample in two groups, all respondents are presented with the list which includes the sensitive item in one survey wave and the list that does not include the sensitive item in another survey wave. Since the entire sample is used, there are less concerns of statistical efficiency, as with standard ICT. Importantly, LICT methods also provide an individual-level indicator of the behavior of interest under certain circumstances, since both lists (with and without the sensitive item) are administered to each respondent. In these cases, analyses can be made directly at the individual-level, including multivari-

Acknowledgments

We are grateful to Noah Uhrig for contributing to the study design and to the development of the Longitudinal Item Count Technique. We are also thankful for comments from Peter Lynn and Annette Jäckle on the preliminary draft of this paper. Finally, we thank two anonymous reviewers and the guest editors of the special issue for their helpful comments on this manuscript.

The second author is funded by a research award from the UK Economic and Social Research Council (award no. ES/K005146/1) for “Understanding Society: the UK Household Longitudinal Study, phase 3.”

Direct correspondence to

Alessandra Gaia, Department of Sociology and Social Research,
University of Milan-Bicocca, Milan, Italy
E-mail: alessandra.gaia@unimib.it

ate methods such as regression models without the need for multiple steps such as those proposed by Imai (2011).

The circumstances where these meaningful individual-level indicators are met mostly likely occur when the items are time invariant, e.g. items that refer to past events, like where the respondents grew up ("I have grown-up in the country-side"), dates in the past which are significant to the respondents, like birthdays of significant others ("My father's birthday is in October"), *etc*. If the selected items are not time invariant (e.g. "I have travelled to Spain"), the event may occur between data collection waves. If that is the case, respondents answering the survey question accurately would report a higher number of items in the second wave compared to the first survey wave.

Time invariance in LICT methods is not always necessary, except that the LICT also rests on the assumption that there is no trend in the list items (upward or downward, across waves). If there is a trend in the list items (including the sensitive behavior) measuring differences will be confounded with change over time. As long as there is not a trend, individual respondent time variability is acceptable, although it will increase the variance of estimates. In some ways, individual time variability may be desirous in the LICT. Generally, the LICT allows researchers to identify whether the trait of interest applies to the respondent, although it provides less privacy than the ICT. In particular, if a respondent remembers the lists across waves, time invariant items may lead respondents to realize they will be reporting on the sensitive behavior by reporting higher or lower counts (depending on which list is presented at which wave).

Conversely, time variant items may allow respondents to maintain a sense of anonymity intended by ICT methods. For example, travelling to Spain may occur between waves, or a tattoo may be removed – in either case, changes to the counts are therefore not directly related to the indication of the sensitive behavior. Further, in the LICT design proposed here, respondents are divided where half of the sample receives the list with the sensitive item in an earlier wave and the list without in a later wave, while the reverse ordering occurs for the other group. To the extent that time variant behaviors do not trend and are distributed equivalently across groups over time, the averaging of estimates will tend to eliminate any bias introduced by time invariant items.

Further, for both the ICT and LICT to work properly, items should be selected to avoid "ceiling" and "floor" effects. If lists contain the non-sensitive items where all items are likely to be selected among respondents ("ceiling" effect), those with the sensitive item list would self-identify by counting all the items. There is also some concern that respondents may view themselves as self-identifying in the case where the list has items where the respondent is likely to select none ("floor" effect). However, this "floor" effect is less problematic, as it requires the assumption that the interviewer can infer that respondents with the sensitive-item list are indicating the sensitive behavior applies to them when reporting a count of one (Kuha & Jackson 2014).

While LICT methods have not yet been explored in research previously, ICT has been used to estimate sensitive behaviors across a number of disciplines. For example, disciplines like development economics or political studies often adopt the ICT (at times referring to it as "list experiments") to elicit very sensitive behaviors – e.g. vote buying in Turkey (Çarkoğlu & Aytaç 2015), voter intimidation in Russia (Frye et al. 2018) attitudes toward Female Genital Mutilation in Ethiopia (De Cao & Luz 2015) the presence of drug trafficking organizations in Mexico (Magaloni et al. 2012); and in conflict settings such as contemporary Afghanistan (Blair et al. 2014). These studies can be extended to explore these phenomena across time using LICT. The implementation of the technique in fields such as development economics is facilitated by the fact that often researchers implement small scale experiments which require observation before and after treatment, where the measures across time allow for implementation of the LICT. Given the frequent need for indicators of sensitive behaviors in many disciplines, LICT may be of particular use.

We motivate the usage of the technique in the next section through the description of a sensitive topic asked in surveys: sexual orientation. Then we describe the features of this innovative technique and the underlying assumptions, provide guidance on its implementation, discuss its limitations, and the ethical implications associated with it. We then present an empirical application of the method on the sensitive topic of sexuality. The implementation of the method is conducted using experimental data from a large scale nationally representative survey of the UK population, the Innovation Panel of *Understanding Society*: the UK Household Longitudinal Study. Three sets of estimates are compared using this experimental data: first, standard direct questions frequently asked in surveys to measure sexual identity; second, we explore ICT and LICT indicators measured at two consecutive waves of the longitudinal study; third, we examine extensions of the ICT and LICT using two lists to generate estimates. We conclude with a discussion of our findings, and implications for further research.

Measuring Sensitive Questions in Surveys: Sexual Orientation

This substantive topic of analysis for the current research, sexual orientation, is chosen for both the importance and the complexity of obtaining reliable estimates in this area. Indeed, providing sound statistical information on the gay, lesbian, or bisexual populations (also called "sexual minorities") is needed to inform policy makers on disadvantage and discrimination. However, obtaining good quality data is methodologically challenging, as sexuality is one of the most sensitive topics when asked about directly in social surveys.

An additional complication is that classification of people's sexuality is complex as "sexual orientation" is a multidimensional construct involving three different dimensions: sexual attraction, sexual behavior, and self-identification (Laumann et al. 1994). "Heterosexual/homosexual/bisexual attraction" indicates whether a person is sexually attracted by someone of the same sex, of the opposite sex, or of both sexes, whereas "heterosexual/homosexual/bisexual behavior" indicates whether someone has had sexual experiences with someone of the same sex, opposite sex, or of both sexes. And sexual identity indicates self-identification into "heterosexual", "homosexual", "bisexual", or "other" sexual identities. Classification of the population could occur along any of these three dimensions (sexual attraction, behavior, and identity) or amongst any combination of them, and it is not clear which are most relevant for population estimation much less monitoring of equality (Aspinal 2009). Until now, large scale multi-purpose UK studies have measured sexual identity as self-identification into "heterosexual", "homosexual", "bisexual", or "other" sexual identities, rather than these various dimensions.

In addition to being a sensitive behavior "non-heterosexual" sexual identity, homosexual attraction and homoerotic behavior are also rare in the general population. Indeed, nationally representative surveys suggest a low prevalence of "non-heterosexual" sexual identity, homosexual attraction and homoerotic behavior in the UK. Results from the UK National Survey of Sexual Attitudes and Lifestyles III show that 3.3% of respondents identified as gay, lesbian, bisexual or other, 3.2% in the UKHLS and 1.9% self-identify as gay, lesbian or bisexual (the option "other" was not provided) in the 2013 British Social Attitudes Survey. In terms of same-sex sexual attraction and homoerotic behavior, data from the National Survey of Sexual Attitudes and Lifestyles III (2010) show that 10.6% of respondents declare being attracted by a person of the same sex and 10.5% declare having had sexual experiences with a person of the same sex.

Overall, there appears to be a low true prevalence of the behaviors of interest, which may have consequences for using methods such as the ICT and LICT. Although Ahlquist (2017) finds that the ICT does not perform well with rare behaviors, Kiewiet de Jonge and Nickerson (2013) find empirical evidence that the ICT is more effective in estimating low prevalence behaviors than high prevalence. In particular, they find that while low prevalence items do not show evidence of artificial inflation (more reports than expected), high prevalence items show a tendency toward artificial deflation (less reports than expected). Given the possibility that the ICT (and by extension LICT) may bolster the measurement of low prevalence behaviors, the sensitive nature of sexual behaviors and the complexity of measuring sexual orientation, we consider the estimation of the all three dimensions of sexual orientation (attraction, behavior, and identity), as an interesting case study for the first implementation of the LICT.

Methodology of the ICT and LICT

In the ICT, survey sample members are divided randomly into two groups, with each being provided a list for which to provide a count of items that apply to them. One list has an additional item, the sensitive behavior of interest. The mean difference in list counts across the two groups theoretically should range between 0 and 1. The result is the estimated prevalence of the sensitive behavior in the population. Formally, the estimated prevalence of the sensitive item using ICT is calculated as following:

$$\hat{p}_{ICT} = \overline{x}_{a+s} - \overline{x}_a \qquad (1)$$

where:

$\overline{x}_{a+s}$ is the average number of items counted in list *a* plus the sensitive item;

$\overline{x}_a$ is the average number of items counted in list *a*.

As long as the two samples are independent the variance is the sum of the variances of each of these means, that is $Var\left(\overline{x}_{\mathrm{a+s}}\right) + Var\left(\overline{x}_{\mathrm{a}}\right)$. Since the ICT only uses half of the sample for each mean estimate, there is a loss in precision in the estimate, and the variance is larger than if the entire sample was used for each mean.

As outlined above, one alternative to solve this problem of precision, as well as provide individual-level estimates, is the LICT. Each respondent is given both lists, one with the sensitive item and one without, and asked for counts of relevant items. These lists are given in different waves, although which lists goes in the earlier wave and which list goes in the later can vary. In particular, it is recommended that the sample is divided randomly such that half gets the list without the sensitive item and half the list with the sensitive item in the earlier wave, with each group getting the other list in the later wave. This balancing allows for effects from time invariant items to potentially average out, assuming events are equally likely to occur for groups over time. The LICT then takes the differences in lists within individuals, opposed to the mean group differences of the ICT. The prevalence of the sensitive behavior is estimated as the mean of the within individuals differences for the entire sample, formally:

$$\hat{p}_{LICT} = \left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}\left(x_{i,\ a+s,\ w(s)} - x_{i,\ a,\ w}\right)\right] \qquad (2)$$

where:

n is the total number of respondents

$x_{i,\ a+s,\ w(s)}$ is the number of items counted in list *a* plus the sensitive item for respondent *i* at the wave with the sensitive item in the list;

$x_{i,\ a,\ w}$ is the number of items counted in list *a* for respondent *i* at the wave without the sensitive item in the list.

The variance of this estimate is based on the difference of dependent observations, hence can be expressed as

$$\begin{aligned} Var\left(\hat{p}_{LICT}\right) &= Var\left(x_{i,\ a+s,\ w(s)}\right) + Var\left(x_{i,\ a,\ w}\right) \\ &\quad -2Cov\left(x_{i,\ a+s,\ w(s)}, x_{i,\ a,\ w}\right) \end{aligned} \tag{3}$$

Where the covariance term accounts for the dependency in measures. This expression can be simplified as $Var\left(\bar{d}\right)$ where $d_i = \left(x_{i,\ a+s,\ w(s)} - x_{i,\ a,\ w}\right)$, and there is no need to compute the separate variances and the covariance. It is then possible to take the difference at the individual level and apply the standard variance estimator to the mean of these individual differences.

Both the ICT and the LICT can also be extended using two lists. The Two-List ICT has been proposed to take advantage of the full sample in a cross-sectional setting, to overcome efficiency problems (Droitcour et al. 1991, Biemer & Brown 2005). Each subsample receives one list with the extra item of interest and one short list without the item of interest (list sets *a* and *b*). As such the estimated prevalence of the sensitive item in the Two List ICT can be formalized as:

$$\hat{p}_{2ICT} = \left(\hat{p}_{s1} + \hat{p}_{s2}\right)/2 \tag{4}$$

where:

$$\hat{p}_{s1} = \bar{x}_{a+s} - \bar{x}_a$$

$$\hat{p}_{s2} = \bar{x}_{b+s} - \bar{x}_b$$

Each list sets *a* and *b* lead to an ICT estimate in the same way as in (1), but then these are averaged to take the overall sample mean. The estimated variance for the Two Lists ICT is as follows:

$$Var\left(\hat{p}_{2ICT}\right) = \left(\frac{1}{4}\right)\left(Var\left(\hat{p}_{s1}\right) + Var\left(\hat{p}_{s2}\right) + 2\rho_{s1s2}\sqrt{Var\left(\hat{p}_{s1}\right)Var\left(\hat{p}_{s2}\right)}\right) \tag{5}$$

Where ρ_{s1s2} is the correlation between the estimators of $\hat{p}_{s1}$ and $\hat{p}_{s2}$, with the expectation that this correlation is negative (Biemer & Brown 2005). The variance can also be estimated (as it is done here) using just the first two terms, i.e. $\left(\frac{1}{4}\right)\left(Var\left(\hat{p}_{s2}\right) + Var\left(\hat{p}_{s2}\right)\right)$, given the complications in estimated ρ_{s1s2} (see Biemer & Brown 2005). However, using this form of the variance will likely overestimate the true variance, as the last term in (5) is likely negative. This overestimate means a reduction in precision and wider confidence intervals, but conversely means there will be greater conservativism in significance testing.

While Two-List ICT methods improve efficiency in estimates, there is still a lack of individual indicators. Since the LICT already uses the full sample, the benefit of having Two-List LICT is that there are multiple indicators of the sensitive behavior, rather than one, which may solidify conclusions by relying on multiple rather than single data points. Like the LICT, the Two-List LICT is estimated within individuals, as all respondents receive both lists with and without the sensitive item. In one wave, respondents receive list *a* with the addition of the sensitive item, and list *b* without the additional sensitive item, and in the other wave (again the order of wave can vary), the other version of each list *a* and *b* is given. Like the Two-List ICT, the Two-List LICT prevalence can be estimated via averaging the estimated prevalence of each of the two list sets *a* and *b,*

$$\hat{p}_{2LICT} = \left(\hat{p}_{LICT(a)} + \hat{p}_{LICT(b)}\right)/2 \tag{6}$$

Where $\hat{p}_{LICT(a)}$ and $\hat{p}_{LICT(b)}$ are estimated separately via (2). The variance of the Two-List LICT then takes the form of the Two-List ICT reported in Biemer and Brown (2005)

$$\begin{aligned} Var\left(\hat{p}_{2LICT}\right) = \left(\frac{1}{4}\right)\Bigg(&Var\left(\hat{p}_{LICT(a)}\right) + Var\left(\hat{p}_{LICT(b)}\right) \\ &+2\rho_{LICT(a),LICT(b)}\sqrt{Var\left(\hat{p}_{LICT(a)}\right)Var\left(\hat{p}_{(LICT(b))}\right)}\Bigg) \end{aligned} \tag{7}$$

Where $\rho_{LICT(a),LICT(b)}$ is the correlation between the estimators of $\hat{p}_{(LICT(a))}$ and $\hat{p}_{(LICT(b))}$. Given both list sets are used for each individual, the correlation estimate is more direct, and this is the variance estimator used in the following empirical example.

Data and Methods

Data come from an experiment implemented in the *Understanding Society* Innovation Panel waves 8 and 9 (IP8 and IP9) (University of Essex 2018). *Understanding Society: the UK Household Longitudinal Study* (UKHLS) is a multidisciplinary study that focuses on a wide range of topics such as living arrangements, fertility, housing, economic activity, income, health, and political attitudes. *Understanding Society* includes an Innovation Panel (IP), a separate sample used to test methodological innovations in longitudinal surveys, in general, and *Understanding Society*, in particular. The Innovation Panel target population is adults (aged 16+) living in Great Britain. The study aim is to interview each adult member of the household and individuals are followed when they move to other parts of Great Britain. Sample members are interviewed every 12 months. The Innovation Panel mirrors *Understanding Society* in its design and it is a stratified, clustered, probability sample. Prior to the fifth wave (IP5), all interviews were conducted by interviewers, but moved to sequential mixed-mode web and CAPI design at IP5. Two-thirds of households were allocated to the mixed-mode design, while the other third were administered the standard single-mode CAPI design. In the mixed-mode treatment, if any household member did not respond to the web survey within three weeks, an interviewer was sent to attempt a face-to-face interview. A mop-up period allows respondents to complete in either web or telephone interviews, although no respondents in the sample completed via telephone. All experimental allocations used in the current study are made independent of the mixed-mode experiment (described in detail in Jäckle et al. 2017).

To ensure that results of the various measures explored are comparable, and because the analysis of interest is across lists across waves, the analytic sample is defined as those who answered all lists given across both waves. Respondents who did not answer all of the lists, including those not responding to any list within a wave or those only responding at one wave are not included in this analysis. Overall, refusal to list questions across both waves was low, ranging from 3.4% of respondents in IP8 on a question on sexual behavior to 0.5% of respondents in IP9 on a question on sexual identity. Also "don't know" answers were rare, to levels lower than 0.7% in all items and waves. Further, due to the possibility that respondents could change waves in the mixed-mode allocation, the data are further restricted to respondents answering in the same mode across wave. This restriction removes any effect that the change of mode could have on responses across waves within respondents. This analytic sample has 1370 respondents.

Experimental Design

Experimental design

The LICT in the IP was designed to measure all three dimensions of sexual orientation (attraction, behavior, and identity), using two lists for each dimension, six in total. The lists are then repeated at the subsequent survey wave to derive the longitudinal element of the ICT. Respondents were randomly allocated at IP8 to one of two conditions. Each of the two conditions received three lists without a key and three with a key item, with the two groups differing on which set of lists were received. At IP9, each group received the reverse set of lists; i.e. if the respondent received a list with the key item at IP8, that list with the same non-key items was presented at IP9 minus the key item or *vice versa*. Given two lists were used for each dimension, Two-List ICT and LICT estimates can also be made. Table 1 shows the experimental design of the LICT.

Table 1 LICT implemented at IP8 and IP9

	IP8	IP9
Group 1	List A	List A + S1
	List B + S1	List B
	List C	List C + S2
	List D + S2	List D
	List E	List E + S3
	List F + S3	List F
Group 2	List A + S1	List A
	List B	List B + S1
	List C + S2	List C
	List D	List D + S2
	List E + S3	List E
	List F	List F + S3

Note: S1 refers to being sexual attracted from someone of the same sex, S2 refers to having had homoerotic sexual experiences (sexual experiences with someone of the same sex), and S3 refers to self-identifying as gay, lesbian, or bisexual.

A basic check of whether the randomization worked tested differences across groups on age (in 7 categories), sex (male, female), marital status (single, formerly married, married), education (university/professional degree, A-level/GSCE, less education) and urbanicity (urban, rural). Generally, the randomization appears to have worked, with all comparisons across conditions not significantly different at p<0.05.

Before the sexual identity ICT questions, the respondent was presented with a brief preamble which explained what was needed for each question; that is, only the counts of behaviors relevant to them. The wording of the introduction (as well as the full question wording for each ICT question) is presented in Appendix 1. As examples, three item lists are presented below, one on sexual attraction, one on sexual behavior and one on sexual identity, each including the sensitive item of interest. After each list on the same screen, respondents were presented with the question: “How many statements are true for you?” with the options “None are true”, “One statement”, “Two statements” “Three statements”, “Four statements”, “Five statements”. Questions without the key item did not have the “Five statements” response option.

Example of item count on sexual attraction:

I have at least once been sexually **attracted** to someone who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

Example of item count on sexual experience:

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

Example of item count on sexual identity:

I would describe myself as **being** …

- gay, lesbian or bisexual
- stylish and fashionable
- disabled
- patient
- British

At each wave, the ordering of item counts (i.e. the different lists) was randomized across respondents, and the statements within lists were also randomized.

The wording of the ICT questions was designed with the aim of mixing non-sensitive items that were expected to be high prevalence with non-sensitive items that were expected to be low prevalence; this is consistent with the indication of the literature (see Glynn 2013). Indeed, if all items in the list are of a high prevalence, gay, lesbian, and bisexual respondents may count all items in the list, and thus self-identify themselves as gay, lesbians, and bisexuals, i.e. a "ceiling effect"; conversely, if all "non-sensitive" items are very rare (and perceived by respondents as being more rare than belonging to the gay, lesbian, and bisexual population), a "floor" effect may occur.

Therefore we combined items that we expected to be low prevalence (e.g. "I would describe myself as being disabled"), with items that we expected to be high prevalence (e.g. "I would describe myself as being British"). When items were designed, in early 2014, items: "I consider myself as being British" (list E) and "I consider myself as being European" (list F) were considered non-sensitive high prevalence items. However, the debate on the United Kingdom European Union membership (which developed in conjunction with the referendum, held on 26th June 2016) pervaded public opinion during the fieldwork for IP9 (summer 2016). This parallel timing may have increased the sensitive nature of these two items, and altered the estimating prevalence of the two items at IP9. Finally, the questions were designed so that the list of items would fit together and make sense to respondents – as suggested by Droitcour et al. (1991).

To explore the possibility of "ceiling" and "floor" effects, Figures 1 and 2 present the distribution of the items reported as true for each list which does not include the sensitive item. We focus on the extremes of the distribution (i.e. 0 and 4 true statements). In the dimensions of attraction (lists A and B) and behavior (lists C and D), the large majority (29.2% - 44.0%) of respondents, in both waves, reported that none of the items presented applied to them; conversely, in the identity questions (lists E and F) the "floor" effect was not problematic, as "none of the statements are true" was selected by only a small percentage of respondents (2.2% - 3.8%).

The evidence for "ceiling" effects is mixed. While lists A (attraction), list C (behavior) and E (identity) resulted with only a small proportion of respondents selecting that all "four statements are true", ranging between 1.1% and 5.1%, lists B (attraction) and F (identity) respondents reporting that all four behaviors range between 16% and 20%. Similarly, while not quite as high, list D had 7.4% of respondents (in IP8) and 10.4% (in IP9) selecting four statements are true. The more limited evidence for "ceiling" effects is reassuring, as "ceiling" effects are more problematic to ICT than "floor" effects (Kuha & Jackson 2014).

In addition to the ICT, respondents were also asked a direct question on sexual identity; sample members were randomly allocated to two different protocols, which vary in question wording and in mode of administration. These two protocols are currently adopted in two large scale studies in the United Kingdom, i.e. *Understanding Society*: the UK Household Longitudinal Study (UKHLS) and the Integrated Household Survey (IHS). The protocols for the two studies are as follows:

Protocol 1 – UKHLS:

The question is asked in self-completion either by Computer Assisted Self-Interview (CASI) or by Web.

Protocol 2 – IHS:

The question is asked Face-to-Face (in Computer Assisted Personal Interview, CAPI) with the aid of a showcard

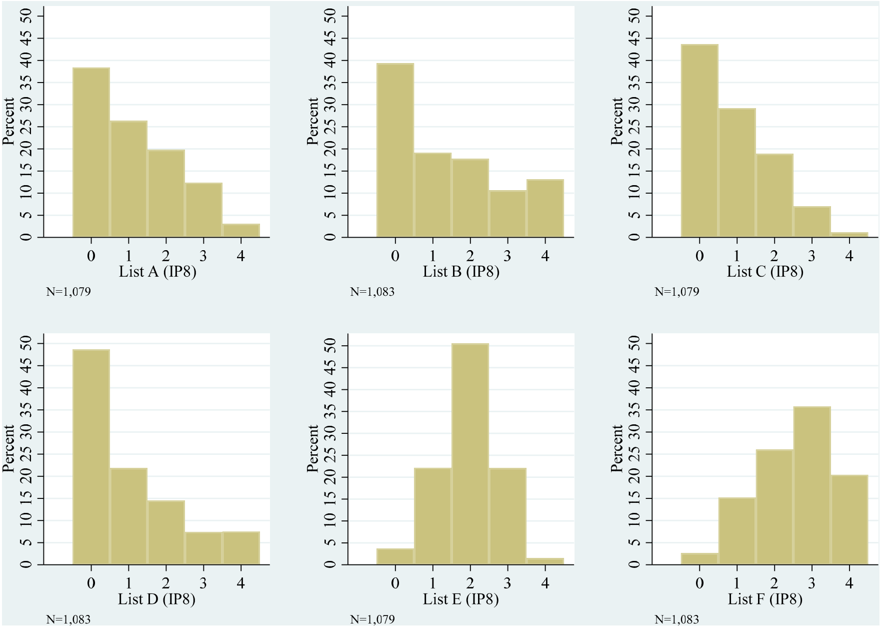

Figure 1 Distribution of Reported Items Excluding Sensitive Items, IP8

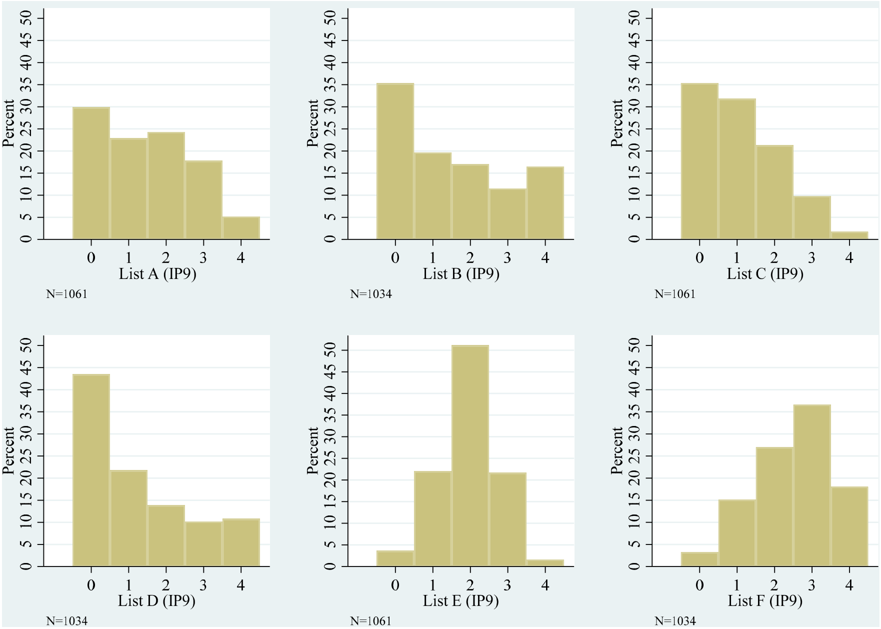

Figure 2 Distribution of Reported Items Excluding Sensitive Items, IP9

The visual design was identical in the Web and CASI versions of the UKHLS question. The question wording for the two protocols, the showcard, and the interviewer instructions are presented in Appendix 1. The ICT questions were separated from the direct sexual identity question in the questionnaire in order to avoid carry-over effects between these survey tasks.

Sample members were randomly allocated to receive either the UKHLS or IHS protocol. The experimental allocation was fully crossed with the allocation to the two lists ICT groups. Respondents were given the same protocol/question in both waves. Deviations to the experimental allocations were implemented to accommodate the mixed-mode nature of the survey design (Jäckle et al. 2017). Specifically, respondents completing the survey by Web answered the question according to the self-completion UKHLS protocol, regardless of their original allocation.

Results

Most surveys have attempted to directly measure sexual identity in questionnaires using a single question (or a small set of questions). These standard forms of questions are the basis of comparison for the Two-List LICT proposed here. Table 2 below presents the self-reported sexual identity using the three different protocols: the UKHLS Web protocol; the UKHLS face-to-face protocol using CASI; and the IHS protocol directly asked by an interviewer using a showcard. While most respondents provided a response at both waves, the UKHLS protocols, which offers an explicit "Prefer Not to Say" option and are self-administered, has more respondents refusing to respond than the IHS protocol.

In all instances, the large majority of responses indicated a heterosexual identification, with more than ninety-percent identifying so in all cases. Slightly more respondents identified as heterosexual in the IHS protocol in both waves, which was asked directly by an interviewer with a showcard. The small cell sizes for non-heterosexual responses make significance testing of the entire response distributions unreliable. However, tests of heterosexual/non-heterosexual responses (binary) show that the UKHLS CASI protocol elicited significantly less (at $p<0.05$) heterosexual responses than either the UKHLS Web ($t(1362)=-2.76$, $p<0.01$) or IHS protocol at IP8 ($t(1362)=-2.04$, $p<0.05$). At IP9, the UKHLS CASI protocol received significantly less heterosexual responses than the IHS protocol ($t(1356)=-2.22$, $p<0.05$), but is not significantly different from the UKHLS Web protocol. While not conclusive, these results are suggestive that, as expected, interviewer-administered questions may lead to more responses seen as socially desirable.

Although the above results suggest mode may reduce socially desirable reporting, it is unlikely to have entirely removed these pressures. As such, item count

Table 2 Self-reported sexual identity using direct questioning

	IP8			IP9		
	UKHLS-Web	UKHLS-CASI	IHS	UKHLS-Web	UKHLS-CASI	IHS
Heterosexual	94.9% (n=590)	91.6% (n=348)	95.1% (n=350)	93.4% (n=581)	91.6% (n=348)	94.6% (n=348)
Gay or Lesbian	1.6% (n=10)	1.6% (n=6)	0.8% (n=3)	1.9% (n=12)	1.8% (n=7)	0.8% (n=3)
Bisexual	1.9% (n=12)	1.1% (n=4)	2.2% (n=8)	1.1% (n=7)	2.4% (n=9)	1.6% (n=6)
Other	NA	1.3% (n=5)	1.1% (n=4)	1.0% (n=6)	1.3% (n=5)	0.3% (n=1)
Prefer Not to Say/ Refused	1.6% (n=10)	4.5% (n=17)	0.5% (n=2)	2.6% (n=16)	2.9% (n=11)	1.4% (n=5)
Don't Know	NA	NA	0.3% (n=1)	NA	NA	1.4% (n=5)
n	622	380	368	622	380	368

techniques may improve reporting and estimates. Table 3 presents the estimates from the IP8 and IP9 ICT, as well as the LICT using data from both waves. Standard errors for each estimate are also presented. These standard errors show that as expected, given the use of the full sample in the LICT versus half in each ICT estimate, the LICT improves efficiency over the ICT estimators. In every comparison between LICT and ICT estimates, LICT estimates have smaller standard errors.

Beyond that result, it is difficult to make other substantive conclusions. This difficulty is largely due to negative values that occur throughout the estimates. If ICT and LICT methods work, negative values should not occur, as respondents with longer lists (i.e. with the sensitive item) are expected on average to provide higher counts. This negative value indicates a negative prevalence of a sensitive behavior, and so is not interpretable. There is some evidence presented in Table 3 to suggest how this may occur.

For example, the IP8 ICT estimate for List B is negative, while at IP9 the List B estimate is positive. This result may occur if those assigned to the List B without the sensitive item at IP8 truly had more non-sensitive items to report on average than those assigned to List B + S (with the sensitive item) at IP8, particularly given the expected low prevalence of the behavior. Respondents with the higher true average without the sensitive behavior in the list could report a higher mean at one wave

Table 3 ICT and LICT estimates

Dimension	IP8 ICT	IP9 ICT	LICT
Attraction			
List A	0.12	-0.05	0.04
(S.E.)	(0.06)	(0.07)	(0.03)
List B	-0.08	0.21	0.07
(S.E.)	(0.08)	(0.08)	(0.03)
Experience			
List C	0.15	0.05	0.09
(S.E.)	(0.06)	(0.06)	(0.03)
List D	0.07	0.09	0.09
(S.E.)	(0.07)	(0.08)	(0.03)
Identity			
List E	-0.01	-0.04	-0.03
(S.E.)	(0.04)	(0.04)	(0.02)
List F	-0.20	0.02	-0.09
(S.E.)	(0.06)	(0.06)	(0.03)

(in this case List B at IP8) than those given the list with the sensitive behavior. Since these same respondents with the higher average are asked the same list with the sensitive item and the group with the lower average asked the list with only non-sensitive items, the expected difference would now be positive. Also, since the higher average respondents would also add in reports of the sensitive behavior, this average could be even larger than the negative value identified. This pattern is what occurs for List B in IP8 and List A in IP9.

This explanation may not actually be what is occurring, and does not clearly explain all of the negative values in Table 3. There are negative values for List E and List F estimates at IP8. At IP9, while the List F ICT estimate is now positive, which could fit with the above explanation, the List E estimate is still negative, and somewhat larger in absolute value. Other explanations may also explain these negative values in ICT estimates, for example various forms of measurement error, such as counting and reporting error of relevant items.

The LICT also leads to negative estimates for List E and List F, and group differences cannot explain these values in the same way, given estimates are within individuals for the entire sample. One explanation is that the items used in these lists are not necessarily time invariant as these can change within respondents. For example, a respondent could count they were healthy (in List F) in one wave, but could be feel unhealthy in the other wave. However, to the extent that changes occur equally over groups assigned to different lists at each wave, these changes should balance out and negative estimates avoided.

While these time invariant items are very much a possible explanation for these negative values in the LICT, as well as other measurement errors (e.g. counting), it should be pointed out that List E contains the item being "British" and List F has the item being "European". As noted above, the lead-up and vote for the UK to leave the European Union occurred during the IP9 fielding period, which may have affected respondents' counts of these items in a differential way than from IP8. If this was the case, which seems possible, the need to avoid a trend (i.e. an event affecting one wave differentially) in the LICT is violated. If this explanation is the case, it underscores the need to avoid items that may trend (although in this case, the possible trend was unforeseen at the design stage).

This trend explanation does not obviously explain the ICT estimates seen for Lists E and F at IP8 and IP9, as these are both cross-sectional estimates. To the extent that the trend explanation holds, at least LICT results are understandable. The LICT also appears to provide better estimates elsewhere, as there are no other negative estimates, unlike for the ICT. Further, the estimates across lists within a dimension (which are estimating the same sensitive item) vary less for LICT estimates than for ICT estimates. The similarity in LICT estimates across lists within dimension suggests the possibility (although not certainly) that the LICT estimates do not depend on list, whereas with ICT the larger variation across lists does not suggest this possibility.

Although the direct questions asked only about identity, which can be a very different construct to attraction and experience, it is also potentially useful to compare ICT and LICT estimates to these direct questions. Using the results originally presented in Table 2 as a baseline is also suggestive about the usefulness of estimates of list methods. For example, while the standard of assessing methods to improve reporting of sensitive behaviors is "more is better" (e.g. Tourangeau & Yan 2007), ICT estimates in Table 3 are at times very much more than those of the direct questions. For example, the UKHLS and IHS protocols provide estimates ranging from 2.7% to 3.5% identifying as being homosexual or bisexual. Comparatively, based on List A at IP8, the ICT estimates 12% of respondents have homosexual attraction and using List B the ICT provides an estimate 21% for the same (these may be due to the differences in non-sensitive items across groups, explained above). Conversely, for the LICT estimates for homosexual attraction is 4% based on List A and 7% on List B, so more than the direct questions, but not as drastically as the ICT estimates. The ICT estimate for homosexual experience based on List A is also 15%; however, the remainder of ICT estimates is relatively smaller or negative.

A suggested improvement to the ICT which may improve estimates is the Two-List ICT (Biemer & Brown 2005). In this case, Two-List ICT averages esti-

Table 4 Two-List ICT and Two-List LICT Estimates

Dimension	IP8 Two-List ICT	IP9 Two-List ICT	Two-List LICT
Attraction	0.02 (0.05)	0.08 (0.05)	0.06 (0.04)
Experience	0.11 (0.05)	0.07 (0.05)	0.09 (0.03)
Identity	-0.11 (0.04)	-0.01 (0.04)	-0.06 (0.03)

mates from the two lists within each dimension presented in Table 3, within waves. The LICT can also be extended to the Two-List LICT using the same averaging of estimates from lists within dimension. The estimates of Two-List ICT and Two-List LICT and the standard errors for these are presented in Table 4.

Both methods lead to negative estimates for Identity (Lists E and F), continuing to suggest problems with the method, noting the potential issues with these specific lists. However, there are no other negative values identified for any other estimate, which is an improvement over single-list ICT estimates, but consistently the same for LICT estimated. The Two-List ICT estimates are relatively smaller due to the averaging effect, and the drastically larger values are generally gone. The Two-List ICT estimate standard errors are also smaller than the single-list ICT estimates, demonstrating the benefit of Two-List ICT over the single-list version (even with the possibly conservative estimate of variance). Comparatively, the Two-List LICT estimates and standard errors are largely the same, given the small variation in individual list estimates. This consistency is reassuring in that lack of consistency (as in the ICT) is suggestive of possible problems. While there is still problematic evidence, and it does not prove the success of the LICT, lack of consistency is not a problem in the current application.

Discussion and Conclusions

This paper describes a new technique for collecting data on sensitive topics in surveys, extending on Item Count Technique methods: the Longitudinal Item Count Technique. Unlike the traditional ICT, this method uses the full sample and provides individual-level data. While results suggest some problems, the LICT results also provide evidence of the method's potential usefulness. The main problem identified is negative LICT estimates in two instances. Certainly negative estimates are problematic in any item count method; a negative prevalence is obviously not a true outcome. However, it is suggested that in this instance, the failure of the LICT to produce realistic estimates are due to the violation of the assumption that there is no trend in the data over time. The two lists that led to negative LICT estimates contained non-sensitive items regarding being British and European; the second administration of these lists occurred during the lead up-to and aftermath of the UK referendum to leave the European Union. Although problematic, if these negative values are due to items that trended, then future implementations of LICT may be able to avoid this problem with careful selection of items. Still, this explanation is not the only one which may explain the problems identified. In particular, the LICT lists used time variant items, which may have caused instability in responses; however, the balancing of lists across waves with a two-group design hopefully countered much of this impact.

Evidence suggesting the potential usefulness of the LICT exists in that it outperformed traditional ICT methods in a number of ways: it had lower standard errors, varied less on lists measuring the same dimension, and provided estimates that were greater, but not drastically so, than differing direct questions on sexual identity, the sensitive behavior of interest. While these results do not prove that the LICT is reliable or accurate, it is suggestive and at least does not prove that the method definitively does not work.

To ensure that the LICT method is useful, further research is needed. In particular, more applications of the LICT are needed using differing sensitive behaviors, especially where true values are known (if possible). The LICT methods here were all completed using self-completion data collection (CASI and Web). Research using face-to-face interviewing is also needed, as self-completion may have a differential impact on response and respondents, as some respondents may not be able to self-complete the questions.

From a design perspective, the downside of the LICT is that it requires multiple waves of data collection, which increases costs, while ICT or direct questions can be handled in a cross-sectional study. It should also be noted that other guidelines for the design of the traditional ICT are relevant also for the LICT (see Glynn (2013) for a recent summary of guidelines). Among the important design issues, in the application of the LICT, researchers need to consider whether an ethical approval is needed for data collection. Indeed, the LICT poses more challenges than the ICT from an ethical point of view, as respondents are revealing their sensitive behaviors by answering both, and they may not be aware of revealing them.

Furthermore, if respondents do realize they are being asked to reveal their sensitive behavior without being asked explicitly may lead to survey drop-out, or, in the context of a longitudinal study, panel attrition. The impact of asking the sensitive behavior to all respondents in the LICT may vary on which list (with or without the sensitive item) is presented at the earlier and later waves. For example, respondents may remember having answered already the short list in an earlier wave, the additional item in the later wave may make the realization of revealing the sensitive behavior more likely. Additionally, the length between waves may impact the method; longer lags between waves may increase the chance respondents do not remember whether they answered a similar question before. Shorter lengths could have the opposite effect.

References

Ahlquist, J. (2018). List experiment design, non-strategic respondent error, and item count technique estimators. *Political Analysis, 26*(1), 34–53. doi:10.1017/pan.2017.31

Aspinal, P. J. (2009). *Estimating the size and composition of the lesbian, gay, and bisexual population in Britain* (Equalities and Human Rights Commission Research Report 37). Retrieved November 12, 2018, from the Equalities and Human Rights Commission website: https://www.equalityhumanrights.com/sites/default/files/research-report-37-estimating-lesbian-gay-and-bisexual-population-in-britain.pdf

Biemer, P., & Brown, G. (2005). Model-based estimation of drug use prevalence using item count data. *Journal of Official Statistics*, *21*(2), 287–308.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis*, *20*(1), 47–77. doi: 10.1093/pan/mpr048

Blair, G., Imai, K., & Lyall, J. (2014). Comparing and combining list and endorsement experiments: Evidence from Afghanistan. *American Journal of Political Science*, *58*(4), 1043–1063. doi: 10.1111/ajps.12086

Corstange, D. (2009). Sensitive questions, truthful answers? Modeling the list experiment with LISTIT. *Political Analysis, 17*(1), 45–63. doi: 10.1093/pan/mpn013

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research, 40*(1), 169–93. doi: 10.1177/0049124110390768

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base-rates for sensitive behavior. *Personnel Psychology, 47*(4), 817–828. doi: 10.1111/j.1744-6570.1994.tb01578.x

De Cao, E., & Lutz, C. (2015). *Measuring attitudes regarding female genital mutilation through a list experiment* (CSAE Working Paper Series No. 2015-20), Retrieved November 12, 2018, from the Centre for the Study of African Economies, University of Oxford website: https://www.csae.ox.ac.uk/materials/papers/csae-wps-2015-20.pdf

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The Item Count Technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, S. Sudman, (eds) *Measurement Errors in Surveys*, pp. 185–210. Hoboken, New Jersey: John Wiley & Sons.

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly, 77*(S1), 159–172. doi: 10.1093/poq/nfs070

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly, 74*(1), 37–67. doi:10.1093/poq/nfp065

Jäckle, A., Gaia, A., Al Baghal, T., Burton, J. & Lynn, P. (eds) (2017). *Understanding Society the UK household longitudinal study Innovation Panel, waves 1-9, user manual*. Colchester: University of Essex.

Jensen, N. M., Mukherjee, B. & Bernhard, W. T. (2014). Introduction: survey and experimental research in international political economy. *International Interactions*, *40*(3), 287–304. doi: 10.1080/03050629.2014.899222

Johnson, A., London School of Hygiene and Tropical Medicine. Centre for Sexual and Reproductive Health Research, NatCen Social Research, & Mercer, C. (2017). *National Survey of Sexual Attitudes and Lifestyles, 2010-2012*. [data collection]. *2nd Edition*. UK Data Service. SN: 7799, doi:10.5255/UKDA-SN-7799-2

Kiewiet de Jonge, C. P. and Nickerson, D. W. (2014). Artificial inflation or deflation? assessing the Item Count Technique in comparative surveys. *Political Behavior 36*(3), 659–682. doi:10.1007/s11109-013-9249-x

Kuha, J. & Jackson, J. (2014). The item count method for sensitive survey questions: modelling criminal behavior. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 63*(2), 321–341. doi: 10.1111/rssc.12018

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

Magaloni, B., Diaz-Cayeros, A., Romero, V. & Matanock, A. M. (2012). The enemy at home: exploring the social roots of criminal organizations in Mexico. Available at: https://ssrn.com/abstract=2122950

NatCen Social Research. (2014). *British Social Attitudes Survey, 2013*. [data collection]. UK Data Service. SN: 7500. doi: 10.5255/UKDA-SN-7500-1

Smith, L. L., Federer, W. T. & Raghavarao, D. (1974). A comparison of three techniques for eliciting answers to sensitive questions. *American statistical association. Proceedings of the social statistics section* pp. 447–452. Washington D.C.: American Statistical Association.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. doi: 10.1037/0033-2909.133.5.859

University of Essex. Institute for Social and Economic Research, NatCen Social Research, & Kantar Public. (2017). *Understanding Society: Waves 1-7, 2009-2016 and Harmonised BHPS: Waves 1-18, 1991-2009*. [data collection]. *9th Edition*. UK Data Service. SN: 6614.

University of Essex. Institute for Social and Economic Research. (2018). Understanding Society: Innovation Panel, Waves 1-10, 2008-2017. [data collection]. 9th Edition. UK Data Service. SN: 6849. doi: 10.5255/UKDA-SN-6849-10

Appendix 1: Question wording

Item Count Technique (CASI & WEB)

Introduction

“The next set of questions will ask you to count the number of statements that are true for you. Please only count the number of statements. We are not interested in knowing which statements are relevant for you.”

Group 1

Item count list A

I have at least once been sexually **attracted** to someone who …

- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements

Item count list B + sensitive item

I have at least once been sexually **attracted** to someone who …

- is the same sex as me
- wears the latest trends and fashions
- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list C

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements

Item count list D + sensitive item

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- is the same sex as me
- wears the latest trends and fashions
- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list E

I would describe myself as **being** …

- stylish and fashionable
- disabled
- patient
- British

How many statements are true for you?

None are true
One statement

Two statements
Three statements
Four statements

Sexuality item count list F + sensitive item

I would describe myself as **being** …

- gay, lesbian or bisexual
- healthy
- tolerant
- European
- working class

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Group 2

Sexuality item count list A + sensitive item

I have at least once been sexually **attracted** to someone who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list B

I have at least once been sexually **attracted** to someone who …

- wears the latest trends and fashions

- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?
None are true
One statement
Two statements
Three statements
Four statements

Sexuality item count list C + sensitive item

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?
None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list D

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- wears the latest trends and fashions
- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?
None are true
One statement
Two statements
Three statements
Four statements

Sexuality item count list E + sensitive item

I would describe myself as **being** …

- gay, lesbian or bisexual
- stylish and fashionable
- disabled
- patient
- British

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list F

I would describe myself as **being** …

- healthy
- tolerant
- European
- working class

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements

Direct questions:

Protocol 1 – IHS

Mode: Face-to-Face with showcard

Question wording: “Which of the options on this card best describes how you think of yourself? Please just read out the number next to the description.”

SHOWCARD
27. Heterosexual / Straight
21. Gay / Lesbian
24. Bisexual
29. Other

Note: “Don’t Know” and “Refuse” were not displayed in the showcard. Interviewers recorded “Don’t Know” and “Refuse” if those where spontaneous answers of the respondent.

Mode: Telephone

Question wording: “I will now read out a list of terms people sometimes use to describe how they think of themselves: “Heterosexual or Straight”, “Gay or Lesbian”, “Bisexual”, or “Other”. As I read the List Again please say ‘yes’ when you hear the option that best describes how you think of yourself.

Heterosexual or Straight
Gay or lesbian
Bisexual
Other”

Interviewer Instruction: on first reading, read list to end without pausing. Note that “heterosexual or straight” is one option “gay or lesbian” is one option. On second reading, please pause briefly after each option.

Protocol 2 – UKHLS

Mode: WEB or CASI

“Which of the following options best describes how you think of yourself?

Heterosexual or Straight
Gay or Lesbian
Bisexual
Other
Prefer not to say”