A Validation Study of Individual-Level Survey Methodologies for Sensitive Questions

Ahra Wu, Andrew R Wood, and Randy T. Stevenson

July 15, 2019

Summary

Randomized Item Response Theory (RIRT) and the Item Count Response Technique (ICRT) both aim to extract individual-level information using variants of survey techniques originally designed to assess phenomena at the group-level. While both methods build on the insights from Item Response Theory (IRT), they rely on distinct survey techniques and use unique input information. This paper offers a comparison of RIRT and ICRT methods in terms of their ability to accurately recover individual-level information. We examine the properties of each method using simulations and explore how each methods performs under a variety of real-world conditions. Our findings offer insights into the advantages and disadvantages of each methods and can be used to inform the design of surveys of sensitive issues.

Introduction and Motivation

Traditional surveys struggle to obtain reliable responses when assessing sensitive topics such as illicit or taboo behavior. Social desirability bias and low response rates often lead to a biased picture of the phenomenon under study by significantly reducing reported rates of anti-social or "bad" behavior and increasing reported rates of pro-social or "good" behavior. Randomized response designs (RRD) and list experiments (LE) are two potential pathways to overcome these biases and recover reliable information.[1] Both of these techniques offer a form of question-level anonymity wherein survey researchers cannot know for certain whether a respondent is admitting to the anti-social behavior. Instead, these techniques allow researchers to estimate prevalence rates for the targeted behavior at the group level; sacrificing information at the individual level to obtain reliable estimates of group behaviors.

While group-level information can be useful on its own, some research projects are explicitly interested in assessing sensitive issues at the individual level (i.e., estimating the probability with which each respondent engages in the targeted sensitive behavior). To facilitate this goal, Randomized Item Response Theory (RIRT) and the Item Count Response Technique (ICRT) can be used to probabilistically estimate individual-level information while maintaining question-level anonymity.[2] While both methods rely on insights from Item Response Theory (IRT), they approach the problem of extracting individual-level information from partially anonymized data in different ways.

RIRT and ICRT are relatively new statistical methods and, while their effectiveness at recovering individual-level information has been assessed in isolation (Bockenholt and van der Heijden (2007); de Jong and Pieters (2019)), they have not been rigorously evaluated in comparison with one another. As we shall see, despite their similarities, these two methods take unique approaches to understanding sensitive issues and offer different insights into the phenomenon. This paper examines these similarities and differences and provides a guide to the advantages and disadvantages of both methods in particular applications.

Understanding the Models

To understand how RIRT and ICRT methods work, it will be helpful to consider their foundations in randomized response designs (RRD) and list experiments (LE). Both of these methods were initially developed

[1] Endorsement experiments are another possible pathway.

[2] See Fox (2005) and Fox et al. (2018) for details on RIRT and de Jong & Pieters (2019) for details on ICRT.

to overcome social desirability bias and the low response rates frequently seen in surveys of sensitive behaviors.

To make things clear, we adopt a running example using unsafe pre-natal behaviors as the sensitive behavior to be assessed. This example is similar to that used by de Jong & Pieters (2019) in their discussion of the ICRT method.

Randomized response designs use a randomization process with a known probability distribution (e.g. a dice roll) to determine either which question the respondent answers (in the Warner model) or which response they should give (in the crosswise model) (Warner 1965; Yu et al. 2008). Since only the respondent knows the outcome of the randomization process, and thus the meaning of their response, respondents are free to answer the question honestly. However, since researchers know the probability distribution of the randomization process, they can estimate the group-level prevalence of the phenomenon under study.

For example, in seeking to learn about unsafe pre-natal behaviors, a researcher could use a Warner RRD where respondents are asked to roll a die and, for example, if the die roll is a 6, respond to the statement "I have smoked while pregnant." Otherwise, the respondent is asked to respond to the statement "I have not smoked while pregnant." Since the respondent's die roll is not known to the researcher, the meaning of their specific response is not known. However, prevalence rates over the whole sample can still be estimated.

While the Warner RRD uses randomization to determine which question is answered, the more recently proposed crosswise RRD takes a different approach. In a crosswise RRD for unsafe pre-natal behavior, the respondent is presented with two statements: "My die roll was 6." and "I have smoked while pregnant." and is asked to roll their die and indicate whether neither or both of the statements are true or only 1 of the two statements is true. Once again, because the die roll us known only to the respondent, the researcher does know whether the individual has smoked while pregnant but can estimate the prevalence of the behavior at the group level.

While there are several different version of the randomized response designs, in our subsequent discussion and in our simulation study, we will examine the crosswise RRD. We focus on this style of RRD because of its recent rise to prominence and its desirable empirical properties. Indeed, crosswise RRD has been shown to have generally better empirical properties than the other styles of RRD.[3] Specifically, prevalence rates from crosswise RRDs are often higher than in other RRDs, and the structure of crosswise RRDs means that there are no obvious strategies a respondent can use to deceive researchers. As a result, we use a crosswise RRD setup as the basis for our RIRT methods and in all subsequent comparisons with LE and ICRT methods.

Define T_i as a dichotomous variable that takes the value 1 if the sensitive statements is true for respondent i and 0 otherwise. Once again, because of the randomization, T_i is only known to the respondent with researchers only observing Y_i, which is respondent i's answer to the question asked in the crosswise design (i.e., whether neither or both the items are true or only one is). This response is a product of T_i and the response to the random item (e.g., "My die roll was 6."), which has probability q (1/6 for the dice roll). With this, the the prevalence rate of the sensitive behavior (i.e., the probability that $T_i = 1$ for an average respondent) can be calculated as follows:

$$Pr(Y_i = 1) = Pr(T_i = 1) * q + Pr(T_i = 0) * (1 - q)$$

Having N respondents answering the same RRD question, the number of people who agrees with both or neither of the two items, M, is $Pr(Y = 1) * N$. Knowing N and q, the mean value of $Pr(T_i = 1)$ is as follows.

$$\begin{aligned} M &= Pr(Y = 1) * N \\ &= (Pr(T_i = 1) * q + (1 - Pr(T_i = 1)) * (1 - q)) * N \end{aligned}$$

Solving for $Pr(T_i = 1)$ (Warner (1965)),

[3]There is still some debate over whether this assessment is accurate. See Höglinger & Jann (2018) for a discussion of the potential risks of crosswise RRD.

$$Pr(T_i = 1) = \frac{\frac{M}{N} - 1 + q}{2q - 1}$$

In contrast to RRDs, list experiments ask respondents to indicate the number of items from a list to which they answer affirmatively (Miller 1984).[4] Respondents are divided into two groups: one with a list composed only of non-sensitive items and the other with a list that adds one sensitive question. Since respondents in the treatment group are only being asked to indicate the number of items they agree with instead of the specific items, they are thought to have very little reason to misrepresent their behaviors and will answer honestly.

As a consequence, random assignment to the two groups and the assumption that response patterns for non-sensitive items should be the same between groups (i.e., that the inclusion of the sensitive item in the second group does not impact responses to the non-sensitive item) allows one to use differences in the response patterns across groups to estimate the prevalence of the sensitive item across the population.

Again, T_i is a respondent i's agreement to the sensitive item in the list experiment design. U_{ij} and U_{lj} denote the respondent's agreement to the non-sensitive items $j \in \{1, 2, \ldots, J\}$ in the list experiment design, in the control group or in the treatment group, respectively. An individual in the control group's response to the list experiment can be described below.

$$Y_{i,control} = \sum_{j=1}^{J} (U_{ij})$$

On the other hand, the response of an individual in the treatment group to the list experiment is:

$$Y_{l,treatment} = T_l + \sum_{j=1}^{J} (U_{lj})$$

N_1 and N_2 indicate the number of respondents in the control group and in the treatment group, respectively. Aggregating individuals' responses for each group, the difference between the average number of items agreed with provides and estimate of the average response rate of $T_i = 1$.

$$Pr(T_i = 1) = \frac{1}{N_2} \sum_{l=1}^{N_2} T_l = \frac{1}{N_2} \sum_{l=1}^{N_2} Y_{l,treatment} - \frac{1}{N_1} \sum_{i=1}^{N_1} Y_{i,control}$$

In the pre-natal smoking example, one can imagine a LE where respondents are provided with a variety of innocuous statements along with, in the treatment group, a single statement of "I have smoked while pregnant." If, for example, the average number of statements agreed with in the treatment group match those in the control group, then one would conclude that, in this group, the rate of smoking while pregnant is indistinguishable from zero.

List experiments are a useful alternative to RRDs and, in some ways, offer a simpler survey technique for assessing sensitive issues. For example, in an RRD setting, respondents must be introduced to the concept of randomizing their response, properly use a randomization device, and respond in the way that accurately reflects both their true answer and the outcome of the randomization process. In settings where attention is limited, respondents have low reading comprehension, or procuring a randomization device is difficult, a list experiment may be preferable since its requirements on these dimensions are much lower. Indeed, from a respondent's perspective, LEs look like a simple combination of several traditional survey questions into one question - something that might appear simpler and easier to respond to than many other alternative methods of assessing sensitive issues.

[4] List experiments are sometimes referred to as the "item count technique." Indeed, the Item Count Response Technique is so named because of this alternative name. However, in this paper, we consistently refer to this method as a List Experiment (or LE).

Of course, this simplicity does have its drawbacks. Most notably, LEs have considerably less statistical power and require a larger sample size to effectively assess a sensitive issue. This is because an LE must divide the sample in half to create treatment and control groups for comparison.[5] In addition, researchers using LEs must carefully consider the items placed on the list since ceiling effects (where respondents refuse to indicate that all the statements are true) and floor effects (where respondents refuse to indicate that none of the statements are true) are both possible and may significantly bias final results (Blair and Imai 2012; Ahlquist 2017).

Recovering Individual-Level Information

The core idea behind both of these techniques is that, by obfuscating the individual-level information, one can gain more reliable, honest answers to sensitive survey questions. Several validation studies have shown that these techniques can outperform direct questioning and produce reliable insights where direct questions produce considerable bias (e.g., Blair et al. (2015), Rosenfeld et al. (2015)). That said, individual-level information is often of direct importance to a research project so, until recently, these methods have sometimes not been used even when there is obvious concerns about social desirability bias..

To remedy this concern and retrieve useful individual-level information, these techniques can, with minor alterations to the survey structure, be combined with methods from Item Response Theory. Doing so results in Randomized Item Response Theory (from RRDs) and the Item Count Response Technique (from LEs) models.

RIRT

Randomized Item Response Theory is a straightforward extension of randomized response designs where, instead of asking a single sensitive question using an RRD, researchers ask each individual a set of closely related questions. These related questions can then be aggregated within individuals using methods from IRT to produce individual-level estimates of the targeted sensitive behavior.

To understand the intuition behind this estimation approach, imagine a deterministic version of this process wherein a single individual is asked the same RRD question (say, about pre-natal smoking) a large number of times (each time re-rolling the die and re-answering the question honestly). Further, imagine that, instead of aggregating responses across individuals, we aggregate only this individual's responses to these identical questions. Using the same equation for a traditional RRD, we would find that the response patterns combined with the randomization probabilities would (with an large enough number of repeated questions) tell us exactly whether the individual has smoked while pregnant or not.

Of course, asking the same individual the same question many times is not an effective survey strategy. However, by instead asking a set of closely related questions that all map on to an underlying sensitive trait of interest (e.g. unsafe pre-natal behaviors), we can use IRT methods to probibilistically estimate an individual's position on the trait in much the same way that IRT methods are conventionally used to, for example, score students on intellectual ability using exam questions. Here, the underlying trait of unsafe pre-natal behaviors is being estimated using (randomized) responses to questions about, for example, pre-natal smoking, alcohol use, or unhealthy pre-natal dietary habits.

Mathematically, this extension of IRT methods is straightforward and simply involves the modification of the IRT model to incorporate the randomization probability used in the questions. We first examine the standard IRT model to understand this modification.

In a standard IRT model T_{ij} is a respondent i's answer to an item j related to an underlying trait or dimension being measured. In the context of tests, where IRT methods were first developed, this would be student i's answer (one if correct; zero otherwise) to a test question j. The probability of the student answering the question correctly is a function of her underlying intellectual ability θ_i. In our unsafe pre-natal behavior example, T_i indicates the true value for the sensitive item. For purposes of understanding a simple IRT,

[5] As we shall see, this concern no longer strictly applies when using a list experiment to extract individual-level information using an ICRT model. In these cases, a control group is no longer necessary. <!– if one is only interested in obtaining individual-level information and is willing to sacrifice the group-level data. –>

suppose that we actually observed this value (as we would in a survey using direct questioning if respondents answered honestly). θ_i denotes the latent trait underlying this attitude. Note that θ_i is a respondent-specific parameter; all of the questions answered by respondent i share the same trait. Also, θ_is are relative ability measures, so scholars often convert them into a standard normal distribution.

In addition to individual-specific ability, the probability of answering a particular question j affirmatively also depends on the attributes of the question. There are many ways to derive question-specific characteristics, but one of the most common approaches is to use a 2-parameter logistic (2PL) model, deriving two parameters for each question, α_j and β_j. This model can be represented in the equation:

$$Pr(T_{ij} = 1) = \Phi(\alpha_j(\theta_i - \beta_j))$$

Here, α_j is often called the "discrimination" parameter of the question j, because it defines how effectively the question differentiates individual-level trait θ_i. As α_j approaches zero, θ_is and T_{ij}s become irrelevant and we cannot successfully sort respondents by their latent traits using the observed answers. Also, if α_j is negative, one's latent trait θ_i is inversely related to T_{ij}, which is a sign that the question is not useful for inferences about the underlying trait. The other parameter in the 2PL model, β_j, is often called the "difficulty" parameter, because $\Phi(-\alpha_j\beta_j)$ is the constant term inside the probit function for each question j. Therefore, $\Phi(-\alpha_j\beta_j)$ denotes the average rate of answering the question j correctly, assuming that the mean of θ_i is zero.

Combining this IRT model to RRD crosswise design described earlier (whose randomization probability is q) we have:

$$\begin{aligned} Pr(Y_{ij} = 1) &= Pr(T_{ij} = 1) * q + Pr(T_{ij} = 0) * (1 - q) \\ &\text{where} \\ Pr(T_{ij} = 1) &= \Phi(\alpha_j(\theta_i - \beta_j)) \end{aligned}$$

Y_{ij} indicates a respondent i's observed response to the RRD question j. Similarly, T_{ij} indicates the respondent's unobserved response to the sensitive item in the same question. However, this simple mathematical approach has significant theoretical implications. Most notably, one must move from assessing a single behavior of interest to understanding an underlying attitude or pattern of behaviors. For example, while our RRD example examined pre-natal smoking specifically, the adaptation to an RIRT setting involved moving to assessing unsafe pre-natal behaviors in general. While this theoretical shift is often acceptable, since general attitudes or patterns of behavior are of interest as well, this approach may be unsuitable for studies seeking to evaluate very specific individual-level behavior. In these settings, ICRT methods may be more appropriate, a topic we will return to after presenting the method in detail.

ICRT

Whereas RIRT uses IRT methods to estimate a single sensitive trait using multiple RRD questions, ICRT approaches sensitive topics in a quite different way and uses IRT methods for a different purpose. Instead of asking a number of sensitive questions, under the ICRT approach, researchers ask only a single sensitive item (in the LE format). Further, they ask this to all respondents without splitting the sample. In addition, however, they also ask a set of direct (traditionally formatted) baseline questions about some set of non-sensitive traits. For example, the researcher might include a well-validated battery of (non-sensitive) questions that tap the respondent's level of compulsiveness or consciousness.

The researcher then uses IRT methods to aggregate the answers to these question batteries (and note that there can be several such batteries) into scores on the corresponding attitude, personality, or behavioral dimension.[6] Thus each individual has a set of scores on these "auxiliary dimensions."

[6]In some cases these traits may be related to the sensitive attitude or behavior under study. However, as we shall see, no such relationship is needed for the basic ICRT model to work.

The next step is the key to this method. The researcher now includes, as non-sensitive items in the list format, a set of items that tap the chosen auxiliary dimensions. Of course, since the researcher knows the respondent's propensity to answer such items in certain ways (based on the respondent's scores on the auxiliary dimensions), this information can be used to determine if the observed item count in the list experiment is greater than the expected sum of non-sensitive items the respondent will answer affirmatively. If so, it is likely that the respondent agrees with the sensitive item.

Again, understanding the deterministic version of this process can be insightful. In this case, imagine that, before the list experiment is given, a respondent answers questions about their age and nationality, for example, indicating that they are a 30 year old German. Next, they are presented with a list experiment that includes statements about their age and nationality - using their previous answers to create the statements. For example, they may see statements like, "I am German." and, "I am 32 years old." In addition, the list of statements also includes the sensitive item, "I have smoked while pregnant." Clearly, despite the list experiment setting, if an individual responds that two of the statements are true, then the researcher can be sure that the respondent has smoked while pregnant since they know, based on the respondent's previous answers, that only one of the two innocuous statements is true.

Obviously, the idea of re-using respondents answers to previous questions in a list experiment is unlikely to work in practice since respondents would likely realize that the answers to the previous questions are known and that they are "giving themselves away" by answering honestly.[7] However, by asking a set of related questions, using IRT methods to score individuals on underlying dimensions, and estimating their response patterns to the list experiment using similar (but not identical) questions, a researcher can accomplish much the same thing and extract individual-level information about the sensitive behavior or attitude of interest.

Indeed, this is exactly the approach taken by de Jong & Pieters (2019) in their model of pre-natal smoking. By first estimating each individuals' underlying score on the dimensions of impulsivity and self-discipline, they compare expected and actual list experiment responses to determine whether a specific individual has smoked while pregnant. For example, when a respondent ranks very low on the estimated dimensions, yet indicates that they agree with a single item, a researcher can infer that the respondent has engaged in the sensitive behavior.

In this model, the researcher asks H number of items for each auxiliary dimension l. Let W_{ijl} denote respondent i's answer to an item j related to an underlying auxiliary dimension l. W_{ijl} is a discrete response: it can be binary (one if correct; zero otherwise) or ordinal on an M-point scale ($m \in \{1, 2, \ldots, M\}$). The probability of the respondent marking the item as correct (or choosing high scale values) is a function of her underlying "ability" on the auxiliary dimension l (ζ_{il}). For example, for a compulsiveness trait, all of the the responses $W_{ij,\text{compulsiveness}}$ share the common trait $\zeta_{i,\text{compulsiveness}}$. Of course, the responses W_{ijl} are also determined by the item parameters κ_j (item j's discrimination parameter) and λ_j (item j's difficulty parameter) as we have described earlier. If W_{ijl} is binary, the following equation describes the 2PL IRT model.

$$Pr(W_{ijl} = 1) = \Phi(\kappa_j(\zeta_{il} - \lambda_j)),\ j = \{1, 2, \ldots, J\}$$

On the other hand, if W_{ijl} is ordinal, a standard graded-IRT model described below is applicable.

$$Pr(W_{ijl} = m) \quad = \quad \begin{cases} \Phi(\delta_{jm}(\zeta_{il} - \eta_{jm})) & \text{if } m = 1 \\ \Phi(\delta_{j(m+1)}(\zeta_{il} - \eta_{j(m+1)})) - \Phi(\delta_{jm}(\zeta_{il} - \eta_{jm})) & \text{if } m \in \{2, 3, ..., M-1\} \\ 1 - \Phi(\delta_{jm}(\zeta_{il} - \eta_{jm})) & \text{if } m = M \end{cases}$$

Next, the researcher includes, as non-sensitive items in the list format, a set of *additional* items that tap the chosen auxiliary dimensions. Now, let U_{ikl} denote respondent i's unobserved binary response (one if correct;

[7] The possibility of using less obviously manipulative questions and inserting a distracting task between the questions and the list experiment does open up the possibility to effectively use this technique.

zero otherwise) to the auxiliary item k that maps on to the auxiliary dimension l in the LE question. The researcher assumes that these auxiliary items in the LE also share the common auxiliary trait ζ_{il}. Of course, since the researcher has a good guess about the respondent's propensity to answer such items in certain ways (based on W_{ijl}), the researcher can also produce estimates of expected number of these auxiliary items the respondent will answer affirmatively. In the actual estimation, utilizing the estimated probabilities of marking each auxiliary item as correct ($Pr(U_{ikl} = 1)$), the researcher can simulate each value of U_{ikl} and sum the values for each respondent i to have the expected sum. This process is done by a data augmentation step in the ICRT Bayesian estimation.

$$\begin{aligned} Pr(U_{ikl} = 1) &= \Phi(\kappa_k(\zeta_{il} - \lambda_k)),\ k \in \{1, 2, \ldots, K\} \\ E(\sum_{k=1}^{K}\sum_{l=1}^{L} U_{ikl}) &= \sum_{k=1}^{K}\sum_{l=1}^{L} U_{ikl} * (Pr(U_{ikl} = 1)) \\ &= \sum_{k=1}^{K}\sum_{l=1}^{L} U_{ikl} * (\Phi(\kappa_k(\zeta_{il} - \lambda_k))) \end{aligned}$$

In addition, the researcher includes one sensitive item in the LE. A respondent i's binary true value of the sensitive item is denoted as T_i. Finally, a respondent i's observed item count in the list experiment, denoted as Y_i, is the sum of $\sum_{k=1}^{K}\sum_{l=1}^{L} U_{ikl}$ and T_i.

$$Y_i = T_i + \sum_{k=1}^{K}\sum_{l=1}^{L} U_{ikl}$$

Having estimates of each respondent's auxiliary traits ($\hat{\zeta_{il}}$), the researcher can determine if the observed item count in the list experiment (Y_i) deviates from the expected sum of auxiliary items - and so indicates that the respondent agrees with the sensitive item.[8] The differences between the two is the respondent's estimated probability of agreeing with the sensitive item.

$$\begin{aligned} Pr(T_i = 1) &= Y_i - E(\sum_{k=1}^{K}\sum_{l=1}^{L} U_{ikl}) \\ &= p_i \end{aligned}$$

Comparison of the Methods

Overall, RIRT and ICRT share some features and have some notable differences. Notably, both require changes to the overall survey design in order to obtain additional information that will be used in the IRT components of each model. Further, while both models share foundations in IRT methods, they utilize IRT in unique ways that present different challenges.

For example, in a RIRT setting, IRT is used to estimate an underlying behavioral trait from a set of related specific behaviors. In this case, it is critical that each individual question coherently maps to the underlying trait. Ensuring that this is the case can be challenging when addressing a novel sensitive issue or one that involves a small set of specific behaviors.

On the other hand, ICRT only uses IRT methods to predict an individual's response to the non-sensitive items on a list experiment where the related underlying dimensions need not be related at all to the phenomenon of interest. In practice, this means that researchers can, for example, always use the same well-validated

[8] Of course, in practice this is all estimated simultaneously rather than sequentially as we have described it.

psychometric measures for all ICRT applications, regardless of how the measures relate to the issue under study. All that matters in these cases is that the underlying IRT models are well founded and produce reliable responses when asked directly.[9]

In addition, as noted before, ICRT models generally focus on determining whether an individual has engaged in a specific behavior or holds a specific attitude. This makes sense since respondents only answer a single question about the sensitive issue. Thus, it is often better to use simple, concrete statements (e.g. "I have smoked while pregnant.") than more complex, conceptual questions that are subject to misinterpretation (e.g. "I have engaged in unsafe behaviors while pregnant.").[10]

In contrast, RIRT methods are clearly better suited for examining underlying traits or patterns of behavior since they focus on aggregating numerous individual responses about specific attitudes and behaviors. While group-level estimates of individual behaviors are still possible using tradition RRD methods, individual responses to specific questions remain unknown even after adapting the model. In some ways, this is good since it implies that the researcher did not, in actuality, deceive the respondent when they indicated that specific answers cannot be determined. However, the continued lack of clarity on individual behaviors an prove problematic when researchers are interested in assessing and perhaps preventing a specific action.[11]

Research Design: Simulation Study

We conduct Monte Carlo simulations to examine RIRT and ICRT methods' ability to recover individual-level information under various conditions. Specifically, we investigate both methods' robustness by the number of respondents, the number of IRT items asked in each design, the overall prevalence rate of the sensitive attitude or behavior, the magnitude of ceiling effects (only applicable to ICRT), and the existence of an opt-out option (e.g., "I prefer not to answer."). Our criteria for examining their performance are the correlation between estimates and the simulated "true" values, the bias of the estimates, and their variance, measured by root mean square error. Overall, our simulations indicate that both designs work well under most of the examined conditions.

Simulation of RIRT and ICRT data sets

Common Individual attributes

Because of their reliance on Item Response Theory, both RIRT and ICRT methods share substantial common ground in their estimation strategy. Utilizing this commonality, we simulate an individual-level data set that contains information about hypothetical respondents' sensitive attitudes or behaviors, their non-sensitive traits, and other personal attributes that affect the sensitive trait. Note that we use the same sensitive trait to simulate both RIRT and ICRT responses: we assume that the underlying trait affects responses in the same way despite the differences in designs. To put it differently, we draw each individual's multiple responses for simulating RIRT responses and draw only one response for simulating ICRT data set from the same latent tendency. Although we understand the differences in both methods' approaches to IRT, starting with a common individual parameters makes our study more structured.

Except for the case where we vary the number of respondents, we simulate $N = 1,000$ respondents. The individual-level data set contains three sets of variables. First, we simulate three personal attributes that might affect one's sensitive attitudes or behaviors. In real-world example, this can be demographic characteristics such as age, gender, socioeconomic status, etc. For the unhealthy pre-natal behavior example, variables such as age, the number of weeks pregnant, etc. can shape one's probability of engaging in unsafe pre-natal

[9] This last point is perhaps more important than one might initially think. While there is some evidence that responses to most psychometric batteries are honestly given, there is always the possibility that social desirability or interviewer effects can still bias responses.

[10] One can, of course, present these more complex statement, but their validity will be subject to considerable dispute.

[11] For example, in an RIRT assessing workplace ethics, responses to the statement "I have stolen from this company." are probably of immediate and direct interest but are, in this approach, simply aggregated into estimates of an individual's "ethical disposition." Of course, if stealing from the company is a rare and highly unethical event, then an individual who has stolen would likely receive a very bad "ethical disposition" score and be flagged for further intervention anyway. For more details on how this process works and the role of IRT methods in this scenario, see Fox (2005).

behavior. Second, we also simulate two non-sensitive traits that are tied to the baseline items used in ICRT. Both personal attributes and scores on the auxiliary dimensions from the non-sensitive traits are randomly drawn from a standard normal distribution. We assume that the personal attributes can be obtained by asking additional, direct questions in a survey. The precise values for the latent traits relating to the non-sensitive attitudes are not observed. We only observe their realized responses that tap these non-sensitive auxiliary dimensions.

Finally, we generate the latent sensitive trait using both personal attributes and non-sensitive traits as predictors. Going back to de Jong and Pieters (2019)'s example, one's impulsiveness may increase the chance of engaging in unsafe pre-natal behaviors. For the magnitude of the effect of these variables on the sensitive trait, we fix their values to cover the range $[-2, 2]$ equally, which we think are reasonable values for the coefficients. By multiplying the personal attributes and the simulated coefficients and summing them, we obtain each individual's sensitive trait θ_i^*. This latent sensitive trait corresponds to the "ability" parameters θ_i in IRT models and is the target of our simulation study where we compare correlations ($\rho(\theta_i, \hat{\theta}_i)$), bias ($E(|\theta_i - \hat{\theta}_i|)$), and RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$). Because these ability parameters are relative to each other, we convert them to a standard normal distribution. Standardization also helps in the simulation because we can set the prevalence rate to 0.5 by default ($\Phi(0) = 0.5$), and we can get a well-balanced distribution of corresponding probabilities.

For a clear illustration of the DGP of θ_i, we provide a formula below. This is the formula used to simulate θ_i under our basic simulation setting ($N = 1000$, $J = 2$, $p = 1/2$, $\pi = 1/2$, and no opt-out option). θ_i^* denotes a respondent i's non-standardized sensitive trait. Assuming that the respondent's sensitive trait is shaped by non-sensitive traits (ζ_{i1} and ζ_{i2}) as well as other personal characteristics (X_{i1}, X_{i2}, and X_{i3}), we generate the sensitive trait scores. The coefficients are fixed throughout the simulations, except for two scenarios mentioned below (prevalence rate and opt-out rate).

$$\theta_i^* = -2 - 1.2\zeta_{i1} - 0.4\zeta_{i2} + 0.4X_{i1} + 1.2X_{i2} + 2X_{i3}$$

Then, we convert θ_i^* to a standard normal distribution, because these parameters are relative to the others'. Now $\theta_i \sim N(0, 1)$ and

$$\theta_i = \frac{\theta_i^* - E(\theta_i^*)}{SD(\theta_i^*)}$$

To summarize, we simulate six individual-level variables in total: two for non-sensitive traits, three for personal characteristics, and the latent sensitive trait. Only the three personal characteristics variables (X) are observed directly in the RIRT and the ICRT design. The two non-sensitive traits (ζ_l) are observed indirectly in the ICRT design by having the responses to the non-sensitive baseline items, but are not observed in RIRT design. Finally, the latent sensitive trait (θ) is not observed in either designs and is the target quantity that we want to recover. This individual-level data set forms the basis of our RIRT and ICRT simulations. However, from this point on, the two simulation models diverge in their specification and must be examined in turn.

RIRT responses

For the RIRT simulations, we simulate multiple (≥ 5) responses to the crosswise RRD design with a randomization probability of 1/6. We do not let the randomization probability vary in this simulation, but it is worth noting that setting the probability sufficiently away from 0.5 is important for the method. Considering the case where randomization is provided by a dice roll, we believe 1/6 both provides the respondents a sense of privacy protection and ensures reliable estimates.[12] For the IRT part of the model, the discrimination parameters α_j are set to equally divide the range $[1, 3]$, and the difficulty parameters β_j

[12] Additional research, conducted for another project, validates this idea and demonstrates a clear trade off between perceptions of privacy (and resulting honest responses) and the model's ability to recover reliable individual-level estimates. For example, in the obvious case when the randomization probability is especially low, e.g., 1/99, the model can effectively recover the parameters of interest but respondents will not answer honestly - resulting in useless estimates.

are set to equally divide the range $[-2,\ 2]$. When the prevalence rate p is not .5, the range of the difficulty parameters are adjusted accordingly. For example, if there are three discrimination parameters, they are set to 1, 2, and 3. If there are five discrimination parameters, they are set to 1, 1.5, 2, 2.5, and 3. In this way, the overall distribution of these parameters remain the same across different numbers of items, and we can compare the correlations, biases, RMSEs across these conditions. Recall that θ_i, the target of our simulations and the key parameter recovered in RIRT applications, is drawn using the process outlined above.

An individual i's binary response to each sensitive item is drawn using these quantities, and we add random noises to the data using the randomization probability (q) of 1/6 and arrive at the following equations.

$$\begin{aligned} Pr(Y_{ij} = 1) &= Pr(T_{ij} = 1) * q + Pr(T_{ij} = 0) * (1 - q) \\ Pr(T_{ij} = 1) &= \Phi(\alpha_j(\theta_i - \beta_j)) \end{aligned}$$

$$\alpha_j \in [1,\ 3],\ \beta_j \in [-2,\ 2],\ p = \frac{1}{2},\ q = \frac{1}{6},\ j \in \{1, 2, \ldots, J\},\ J \in \{5, 9, 13\}$$

Again, T_{ij} denotes a respondent i's agreement with the sensitive item in question j. Furthermore, α_j and β_j are the discrimination and difficulty parameters of the question j. The total number of items are denoted as J. Finally, Y_{ij} describes a respondent i's response to the crosswise RRD question j, choosing "both statements are true or neither of the two is true," where the randomization probability is q. p indicates the overall prevalence rate of sensitive attitudes/behaviors, set to .5 by default.

In sum, the RIRT simulation data set consists of two sets of variables. First, we obtain respondents' three personal attributes described earlier (X) by asking for the information directly during the survey experiment. Second, we obtain respondents' binary responses (Y) (coded one if "both or neither is true", zero otherwise) that are masked by the randomization probability $q = 1/6$. We use this information to predict θ_i, α_j, and β_j.

ICRT responses

Simulating ICRT responses is more complicated. First, we simulate multiple baseline items j for each non-sensitive trait, when the total number of questions per trait is J. As described earlier, we infer two non-sensitive auxiliary dimensions, indexed by l, from the ordinal responses to these questions. The default number of questions per trait (j) is two, and we increase the number to four and six when we examine the effects of the number of baseline items on the ICRT model's performance.[13] Each of these questions are asked on a $5(= M)$-point scale (e.g., from "strongly disagree" to "strongly agree"). Again, for the IRT component of the model, the difficulty (η_{jm}) and discrimination parameters (δ_{jm}) of each response value m from question j are set to equally divide the ranges $[-2,\ 2]$ and $[1, 3]$. p is the overall prevalence rate, set to .5 by default.

The equation below describes the probability of $W_{ijl} = m$ i.e., respondent i choosing a particular point m on the 5-point scale for the baseline question j relating to auxiliary dimension l. Critically, ζ_{il} is respondent i's score (i.e. ability parameter) on the latent non-sensitive dimension l. This process results in a data set with four responses per respondent, two for each non-sensitive trait with a scale from 1 to 5.

$$Pr(W_{ijl} = m) = \begin{cases} \Phi(\delta_{jm}(\zeta_{il} - \eta_{jm})) & \text{if } m = 1 \\ \Phi(\delta_{j(m+1)}(\zeta_{il} - \eta_{j(m+1)})) - \Phi(\delta_{jm}(\zeta_{il} - \eta_{jm})) & \text{if } m \in \{2, 3, 4\} \\ 1 - \Phi(\delta_{jm}(\zeta_{il} - \eta_{jm})) & \text{if } m = 5 \end{cases}$$

$$\delta_{jm} \in [1,\ 3],\ \eta_{jm} \in [-2,\ 2],\ p = \frac{1}{2},\ j \in \{1, 2, \ldots, J\},\ J \in \{2,\ 4,\ 6\},\ l \in \{1,\ 2\}$$

[13] The default number of two items per trait is to faithfully follow up de Jong & Pieters (2019).

Having simulated responses (W_{ijl}) to the baseline questions and scores (ζ_{il}) on the underlying auxiliary dimensions, we now turn to the responses to each item in the LE. Along with the sensitive statement, we add one new non-sensitive statement per auxiliary dimension, so we have three items in the LE. For the two non-sensitive statements, we use scores on the auxiliary dimensions as well as another set of difficulty and discrimination parameters to simulate binary responses (one if agree, zero otherwise) to each item. We use κ_{j1} and λ_{j1} to denote the discrimination and difficulty parameters of these non-sensitive items. Again, κ_{j1} and λ_{j1} are also fixed to equally divide $[1, 3]$ and $[-2, 2]$[14] when we vary the number of items. In actual ICRT estimation, the determination of these difficulty and discrimination parameters is done simultaneously using IRT and the responses to the LE. U_j denotes the binary response to each auxiliary item in the ICRT LE question. T_j denotes the binary response to the sensitive item.

For the sensitive item, we use θ_i, the ability parameter and the target of our simulations, to draw one binary response (T_i) from a Bernoulli distribution with probability ($\Phi(\theta_i)$). This formulation means that the higher one's latent sensitive trait is, the more likely one is to mark the sensitive item as correct. Note that we did not introduce θ_i in our ICRT discussion above. Introducing the ability parameter with a single question does not make much sense in the IRT setting. At the same time, we use the notation to build connections between our RIRT simulations and ICRT simulations.[15] In other words, we assume the case in which the same respondents take both ICRT and RIRT, thus the same underlying sensitive trait shapes both sets of responses. Of course, in actual ICRT estimation, a researcher does not "observe" these values but instead estimates them using the count from the list experiment. Additionally, note that, since, in our simulations, θ_i is the linear combination of the individual attribute of respondents (X_i) and scores on the latent auxiliary dimensions (ζ_{il}), we can rewrite the probability of T_i as such.

$$\begin{aligned} Pr(U_{i11} = 1) &= \Phi(\kappa_1(\zeta_{i1} - \lambda_1)) \\ Pr(U_{i22} = 1) &= \Phi(\kappa_2(\zeta_{i2} - \lambda_2)) \\ Pr(T_i = 1) &= \Phi(\theta_i) \\ &= \Phi(\sum_{l=1}^{2} \zeta_{il}\phi + X_i\gamma) \end{aligned}$$

$$\kappa_1 = 1,\ \kappa_2 = 3,\ \lambda_1 = (-2),\ \lambda_2 = (2)$$

Finally, recalling our discussion of the analysis of list experiments above, as a final step, after simulating individual responses to each item, we sum the number of items marked as correct. The final product of the ICRT simulation yields a data set containing information about respondents' three personal attributes (X), four 5-scaled responses to the preliminary questions related to the auxiliary traits (W), and one count response to the LE (Y).

Modeling the Opt-out Option

As we will discuss in the next section, the addition of an opt-out option allowing respondents to refuse to answer the RRD or LE questions about a sensitive trait can have significant effects on the reliability of an RIRT or ICRT model. In order to properly assess these effects, we build extensions of the two methods above. These modifications are based on mixture of the above probabilities and the probability of opting out, as developed in Bockenholt and van der Heijden (2007) and Fox et al. (2012).[16] We also model the probability of opting out (π_i) as a function of other individual characteristics, denoted as V_i.

In the RIRT setting, our modifications result in the following three equations.

[14] Again, the difficulty parameters are adjusted accordingly when the prevalence rate is not .5.

[15] There are no difficulty and discrimination parameters for the sensitive item (T_i) since only one item is presented in the ICRT setting and, as result, IRT methods would not be able to determine these parameters. In practice, what this means is that the items difficulty and discrimination parameters are set to 0 and 1 respectively.

[16] Bararesi et al. (2014) also provides numerical derivations and proofs for the idea.

$$
\begin{aligned}
Pr(Y_{ij} = 1) &= (1 - \pi_i) * (Pr(T_{ij} = 1) * q + Pr(T_{ij} = 0) * (1 - q)) \\
Pr(Y_{ij} = 2) &= (1 - \pi_i) * (Pr(T_{ij} = 1) * q + Pr(T_{ij} = 0) * (1 - q)) \\
Pr(Y_{ij} = 3) &= \pi_i = \Phi(V_i\varphi)
\end{aligned}
$$

Similarly, the probabilities of agreeing with each item in ICRT also change as follows. Y_i indicates the count number of correct items reported by the respondent i.

$$
\begin{aligned}
Pr(Y_i &= 0) = (1 - \pi_i) * (Pr(U_{i11} = 0) * Pr(U_{i22} = 0) * Pr(T_i = 0)) \\
Pr(Y_i &= 1) = (1 - \pi_i) * (Pr(U_{i11} = 1) * Pr(U_{i22} = 0) * Pr(T_i = 0) + \\
&\quad Pr(U_{i11} = 0) * Pr(U_{i22} = 1) * Pr(T_i = 0) + \\
&\quad Pr(U_{i11} = 0) * Pr(U_{i22} = 0) * Pr(T_i = 1)) \\
&\vdots \\
Pr(Y_i &= 3) = (1 - \pi_i) * (Pr(U_{i11} = 1) * Pr(U_{i22} = 1) * Pr(T_i = 1)) \\
Pr(Y_i &= 4) = \pi_i = \Phi(V_i\varphi)
\end{aligned}
$$

Parameters Under Study

For our simulation study, we vary numerous model parameters to observe how both the RIRT and ICRT models perform under different scenarios. Here we briefly describe the motivation behind our interest in each parameter and detail the values at which we set each.

- *Number of Respondents:* $N = \{100, 300, 500, 700, 1000, 1500, 2000\}$

  The number of respondents included in the sample represents and obvious metric by which to evaluate these models. Our general expectation is that more respondents will be better in both cases. However, we are interested in seeing how each method performs in small samples and determining whether general cutoffs exist where each method is no longer practically viable due to sample limitations.

- *Number of IRT Items*: $N_{item} = \begin{cases} j = \{5,\ 9,\ 13\} & \text{for RIRT} \\ = l * k + 1 = \{5(= 2 * 2 + 1),\ 9(= 4 * 2 + 1),\ 13(= 6 * 2 + 1)\} & \text{for ICRT} \end{cases}$

  As with the number of respondents, we also expect that the number of IRT items asked will be positively correlated with the success of each method. We use the same number of questions in both designs. For RIRT, all of the items are RIRT items: we ask 5, 9, and 13 RIRT items for each scenario. For ICRT, the number of baseline items change: when $N_{item} = 5$, this means that there are two baseline items for each auxiliary trait and we ask one LE ($N_{item} = k * l + 1 = 2 * 2 + 1 = 5$). Similarly, we ask four baseline items per non-sensitive trait ($N_{item} = k * l + 1 = 4 * 2 + 1 = 9$) when $j = 9$, and ask six items per non-sensitive trait when $N_{item} = 13$.

- *Group-level Prevalence Rate*: $\{0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9\}$

  The overall prevalence rate may also have an impact on the model's ability to recover the respondents' sensitive traits. In general, we expect that it will be more challenging to recover latent traits when the incidence of the trait is low in the population. In such a setting, admitting to such a trait would occur rarely, so when combined with the noise from the randomization, accurate estimation becomes difficult.

- *Opt-out Rate:* $\{0.2,\ 0.3,\ 0.4,\ 0.5\}$

  Researchers often debate whether to include an opt-out option in surveys. Since some people may still be reluctant to answer sensitive questions - even in the RRD or LE setting - providing an opt-out option can be used to add another layer of privacy protection. Unfortunately, it may be the case that certain people are more likely to opt out than others. For example, in addition to respondents who do

not want to be singled out for their "bad" behaviors, other respondents may opt out because they do not understand the question format well and do not want to cheat, or because they simply do not like being asked about private matters.

In general, including an opt-out option results in a loss of efficiency since opt-out responses are often excluded from estimations. Even worse, if the probability of opting out is related to confounding factors, a sample selection problem with missing data arises. Our simulations do not address these, more complicated, scenarios, but we do offer insights into how rates of opting out impact the reliability of the two models.

- *Degree of Ceiling Effect (ICRT only)*: {0.1, 0.2, 0.3, 0.4, 0.5 0.6, 0.7, 0.8, 0.9}

  In LE designs, when respondents find that all of the items, including the sensitive item, are true, they have incentives to deviate to the other option. This is because choosing the option "all of the items are true" inherently involves admitting to the sensitive attitude or behavior. This removes the privacy protection provided by the design, induces deception by respondents, and, as a result, violates the LE's assumptions. This phenomenon is called a "ceiling effect" in the literature. The rate of those who think all items are true but deviate to other options is important for researchers to better understand the individuals' sensitive attitude or behavior.[17]

  In de Jong and Pieters' (2019) ICRT model, ceiling effects are modeled directly. Specifically, an additional parameter τ is included in the model that indicates the average probability of deviating to the option "all but one are true" from "all of the items are true." The assumption that the respondents will always deviate to the particular option may not always be met. Another possibility is to choose any other choice, including "none is true." However, it is reasonable to assume that the respondents will deviate only so far as to preserve their privacy and are still willing to be honest about the other, non-sensitive items. Therefore, while this assumption may not be perfect, it is reasonable and is the approach we use in our simulations.

Model Estimation

For all of the simulations except examining the effect of an opt-out option, we use the models based on the above data generating process. Our RIRT estimation is based on Fox (2005), using three individual-level covariates to predict one's sensitive attitude or behavior. Note that the effect of non-sensitive traits is not estimated in the RIRT setting. Their effects will be subsumed into the constant term in predicting the sensitive trait. On the other hand, our ICRT estimation utilizes all of the individual-level attributes, including the non-sensitive traits measured through the baseline items. This is one of the reasons why direct comparison of these two methods can be difficult: In general, RIRT is only interested in measuring the sensitive attributes, even though one can collect other demographic information.

As discussed above, we utilize de Jong and Pieters (2019)' sample WinBUGs code to estimate the quantities of interest. Both are Markov Chain Monte Carlo methods, which is especially useful for IRT models with high dimensionality. For more details of each model, we recommend consulting the two reference articles. For examining the effect of an opt-out option, we modify the two methods accordingly so that they reflect the data generating process described above. We use uninformative normal priors for all of the parameters except the discrimination parameters. We use a uniform prior of $[0, 5]$ for those in RIRT. Their hyperparameters are drawn from a standard normal distribution and passed to an exponential function in ICRT estimation. We run a single chain for each simulation for 5,000 iterations after the initial burn-in of 5,000 using JAGS.

[17] Similarly, there is also a "floor effect," where subjects deviate from "none is true" to "one item is true." However, assuming that the sensitive item is susceptible to social desirability bias and is regarded negatively in general, respondents would not feel much pressure to reveal that they do not agree with the sensitive item. Therefore, we do not conduct simulations for the floor effect in this paper.

Results

RIRT

Number of respondents

Figure 1 shows the simulation results for different numbers of respondents. The x-axis indicates the number of respondents ranging from 100 to 2,000. Here the red boxplots show the distribution of five mean biases ($E(|\theta_i - \hat{\theta}_i|)$) of the estimated individuals' tendency to have sensitive traits for each scenario. Furthermore, the blue boxplots show the distribution of five RMSEs ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$) of the target quantity for a given number of respondents. The smaller both quantities are, the better the model is at recovering the true latent sensitive trait. The red line connecting the boxplots is the mean of the five mean biases for each number of respondents, and the blue line connects the mean of the five RMSEs for each number of respondents. The decreasing relationship between the number of respondents and bias/RMSE becomes more apparent.

Table 1 shows the correlations between θ_i and $\hat{\theta}_i$ as well as the average values used in Figure 1. In general, it is clear that studies using RIRT should aim to maintain a sample that include over 100 respondents. While our simulations indicate near correlation of 0.9 at 2,000 respondents, in most applications correlations around .85, such as those seen with 500 respondents, are likely to be acceptable. For more accurate measures, having more than 1,000 respondents improves the statistics more clearly.

Figure 1: RIRT Results for Varying Number of Respondents

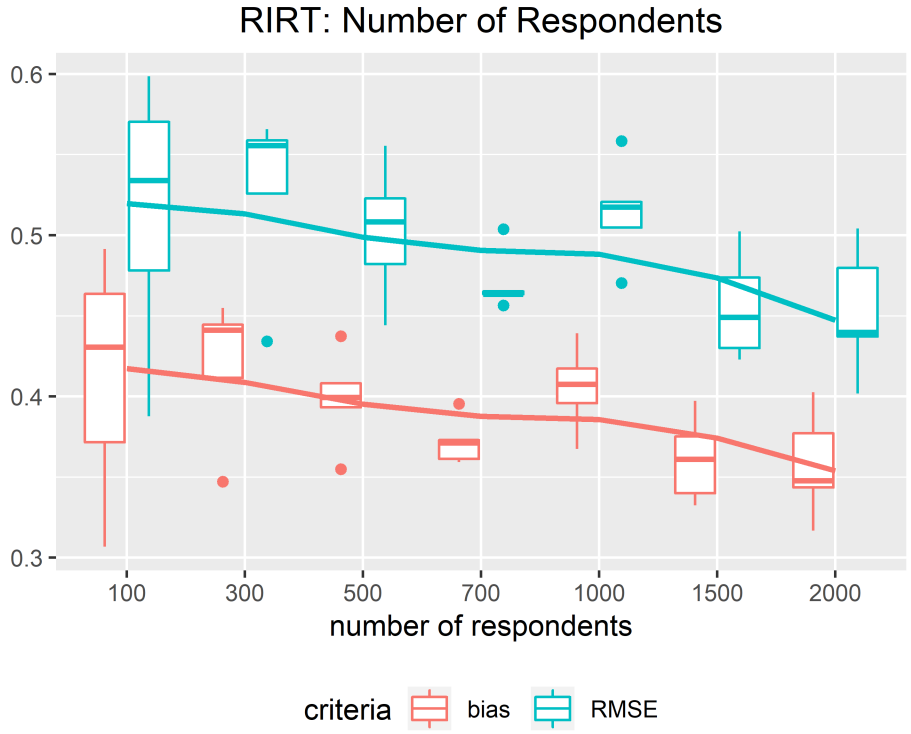

Table 1: RIRT Results for Varying Number of Respondents

	correlation ($\rho(\theta_i,\ \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
$N = 100$	0.853	0.413	0.514
$N = 300$	0.847	0.420	0.528
$N = 500$	0.864	0.399	0.503
$N = 7000$	0.882	0.372	0.470
$N = 1000$	0.857	0.405	0.514
$N = 1500$	0.890	0.361	0.456
$N = 2000$	0.891	0.358	0.453

Number of RIRT Items

Again, to examine whether this result can be attributed to the stochastic process of the simulation, we fix the coefficients to certain ranges and re-run the simulation five times for each condition. Figure 2 and Table 2 summarizes these results. As expected, we can see a clear pattern of decreasing biases/RMSEs as we add more RIRT items. In particular, both biases and RMSEs decrease considerably from asking five questions to asking nine questions. At the same time, the marginal benefit of adding more questions decreases as the number of RIRT items increases. Therefore, we generally recommend that a moderate number of items (e.g. around 10) be used in applied settings.

Figure 2: RIRT Results for Varying Number of Items

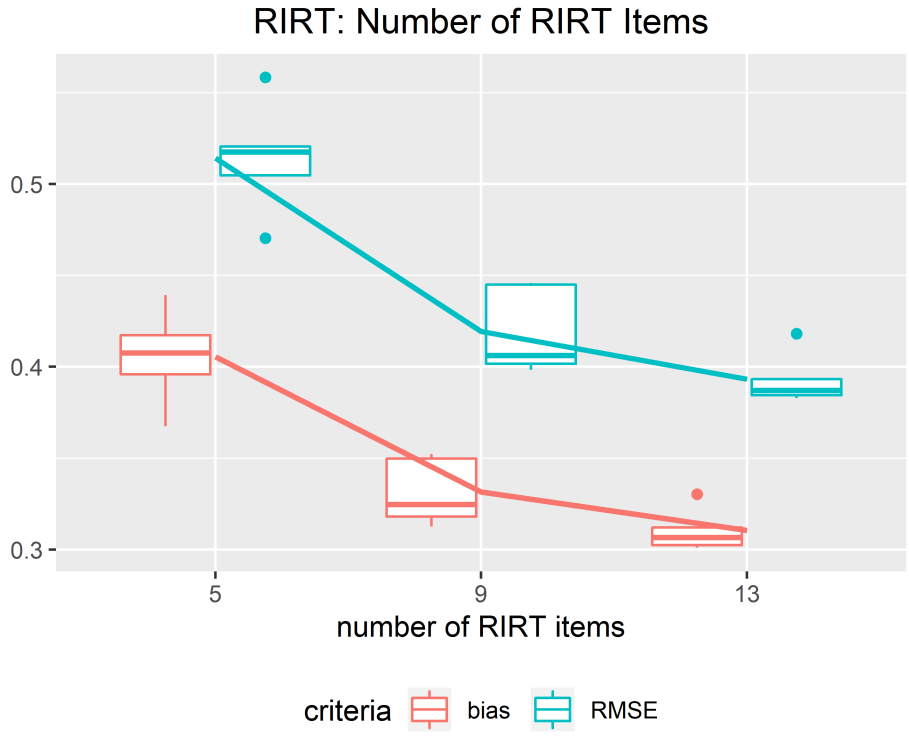

Table 2: RIRT Results for Varying Number of Items

	correlation ($\rho(\theta_i,\ \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
$m = 5$	0.857	0.405	0.514
$m = 9$	0.908	0.331	0.419
$m = 13$	0.919	0.310	0.393

Group-level Prevalence Rate

To examine how prevalence rates affect RIRT estimates, we vary the group-level prevalence rate from 0.1 to 0.9 by increments of 0.1. The results are shown in Figure 3 and Table 3. We find that the bias and RMSE only slightly decrease as the overall prevalence rate increases. In general, the estimates are not strongly affected by the prevalence rate. As shown in Table 3, all of the estimates have correlations around .85. This result suggests that RIRT can be relied upon to assess both rare and commonplace sensitive issues.

Figure 3: RIRT Results for Varying the Overall Prevalence Rate

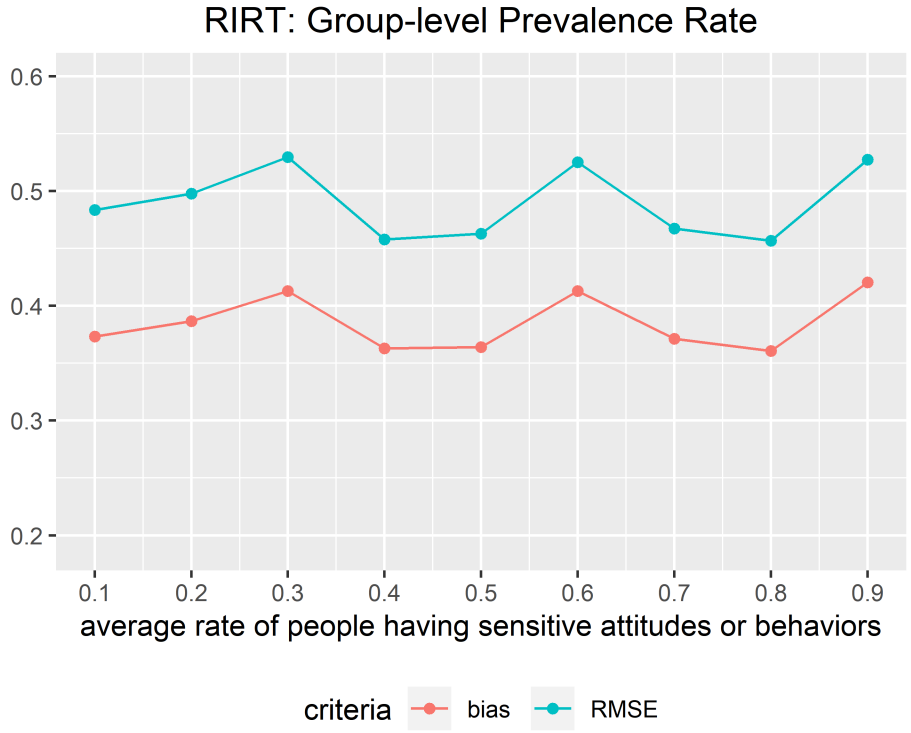

Table 3: RIRT Results for Varying the Overall Prevalence Rate

	correlation ($\rho(\theta_i, \hat{\theta_i})$)	bias ($E(\lvert\theta_i - \hat{\theta_i}\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta_i})^2))^{1/2}$)
$E(\Phi(\theta_i)) = 0.1$	0.876	0.373	0.484
$E(\Phi(\theta_i)) = 0.2$	0.868	0.387	0.498
$E(\Phi(\theta_i)) = 0.3$	0.848	0.413	0.530
$E(\Phi(\theta_i)) = 0.4$	0.889	0.363	0.458
$E(\Phi(\theta_i)) = 0.5$	0.887	0.364	0.463
$E(\Phi(\theta_i)) = 0.6$	0.851	0.413	0.525
$E(\Phi(\theta_i)) = 0.7$	0.885	0.371	0.467
$E(\Phi(\theta_i)) = 0.8$	0.891	0.361	0.457
$E(\Phi(\theta_i)) = 0.9$	0.850	0.420	0.527

Opt-out Option and the Rate of Opting out

Next, we examine the relative performance of both an advanced model that addresses the opt-out option and a simple RIRT model. Furthermore, we also investigate how the performance of these models is affected by the overall opt-out rate.

Figure 4 and Table 4 describe the simulation results. As can be seen in Figure 4, the bias and RMSE do not vary across different values of opt-out rate in RIRT. At the same time, when a researcher provides the opt-out option in the RIRT design and runs the simple version of RIRT by simply dropping the opt-out responses, the expected bias and RMSE are much higher than when the researcher uses the advanced model to directly model the opt-out responses.

We suspect that this is because when estimating each individual's target sensitive trait, the researcher has to rely on the respondent's answers that the respondent did not opt out. In other words, if a respondent opted out three RIRT items out of five, the estimation only relies on two of her responses. Therefore, estimating the respondent's latent sensitive trait becomes more difficult. Even worse, if the answered RIRT items' discrimination parameters are relatively small or even negative, the estimation becomes more challenging for the respondent. Note that we assume that all of the RIRT items map on to the sensitive dimension well in our simulations: we set the discrimination parameters to equally divide the range $[1, 3]$. Still, having missing responses to some questions and dropping them in analyses is susceptible to considerable amount of bias and losing efficiency. At the same time, the steady pattern of bias and RMSE across different opt-out rates when using the advanced model is encouraging. Even when respondents choose to opt out at the rate of 50%, the bias and RMSE do not increase rapidly.

This result clearly demonstrates that the decision to offer opt-out choices in the survey should not occur in isolation from the form of the statistical model ultimately used. If an opt-out option is provided, researchers should use RIRT methods that explicitly account for opt-outs. However, it is unclear whether offering an opt-out option in the first place is a good idea. Specifically, since no obvious strategy for deception exists in crosswise RRDs, it is unclear whether offering an opt-out (and allowing for efficiency loss) actually prevents any other non-cooperative behavior. Ultimately, this is an empirical question about the deceptive behaviors of respondents when faced with a crosswise RRD that should be addressed in future work. For now, we can only emphasize that, when they are implemented in the survey, opt-out options should be directly modeled in the subsequent analysis.

Figure 4: RIRT Results for Varying the Overall Opt-out Rate

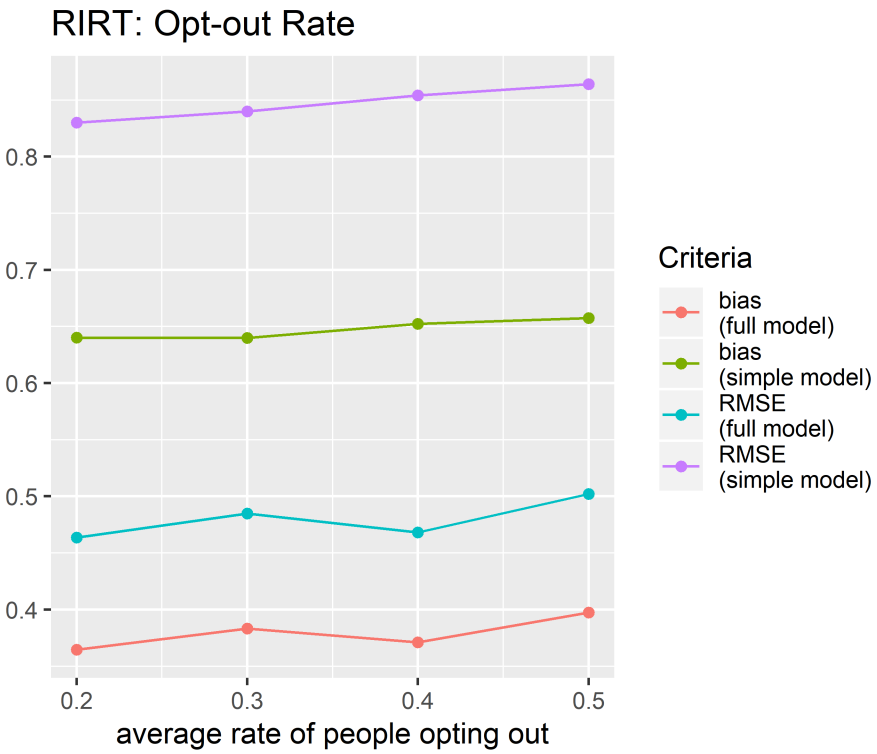

Table 4: RIRT Results for Varying the Overall Opt-out Rate

		correlation ($\rho(\theta_i, \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
advanced model	$E(\pi_i) = 0.2$	0.887	0.365	0.463
	$E(\pi_i) = 0.3$	0.876	0.383	0.484
	$E(\pi_i) = 0.4$	0.884	0.371	0.468
	$E(\pi_i) = 0.5$	0.866	0.397	0.502
simple model	$E(\pi_i) = 0.2$	0.567	0.640	0.830
	$E(\pi_i) = 0.3$	0.550	0.640	0.840
	$E(\pi_i) = 0.4$	0.529	0.652	0.854
	$E(\pi_i) = 0.5$	0.511	0.658	0.864

ICRT

In this section, we turn our attention from the RIRT simulations to the ICRT simulations. We examine the same set of factors as before (i.e. the number of respondents, number of items, prevalence rate, and opt-out rate). In addition, we also examine how ICRT performs under varying degrees of ceiling effects.

Number of respondents

Figure 5 shows the ICRT simulation results for different numbers of respondents. The figure reads similar to Figure 1. Again, the x-axis indicates the number of respondents ranging from 100 to 2,000. The red line and boxplots show the pattern of five mean biases for each category, while the blue line and boxplots show the pattern of RMSEs. As expected, we see a clear pattern of decreasing bias and RMSE in general. In particular, the amount of bias significantly decreases as we bring more respondents into the model. Similar to what we found in Figure 1, adding more respondents to the sample in excess of 1,000 does not seem to result in large

improvements to the estimates. Although the indicators may look reasonable in Figure 5 when the number of respondents is relatively low (≤ 100), we have experienced some model convergence issues.[18] We have not experienced such issues for RIRT, so we recommend using RIRT when the number of respondents is low.

This can be clearly seen in the correlations and bias measures from Table 5. While our simulations suggest that 100 respondents is likely to be acceptable, it is up to individual researchers to determine how small of a sample they and how much bias they are willing to tolerate. Note that the numbers in RIRT and ICRT cannot be compared directly, because the effects of non-sensitive personal attributes (ζ_{il}) are not addressed in RIRT estimation. In other words, the current RIRT simulation setting has a model misspecification issue. We plan to run another set of simulation removing such effects.

Figure 5: ICRT Results for Varying Number of Respondents

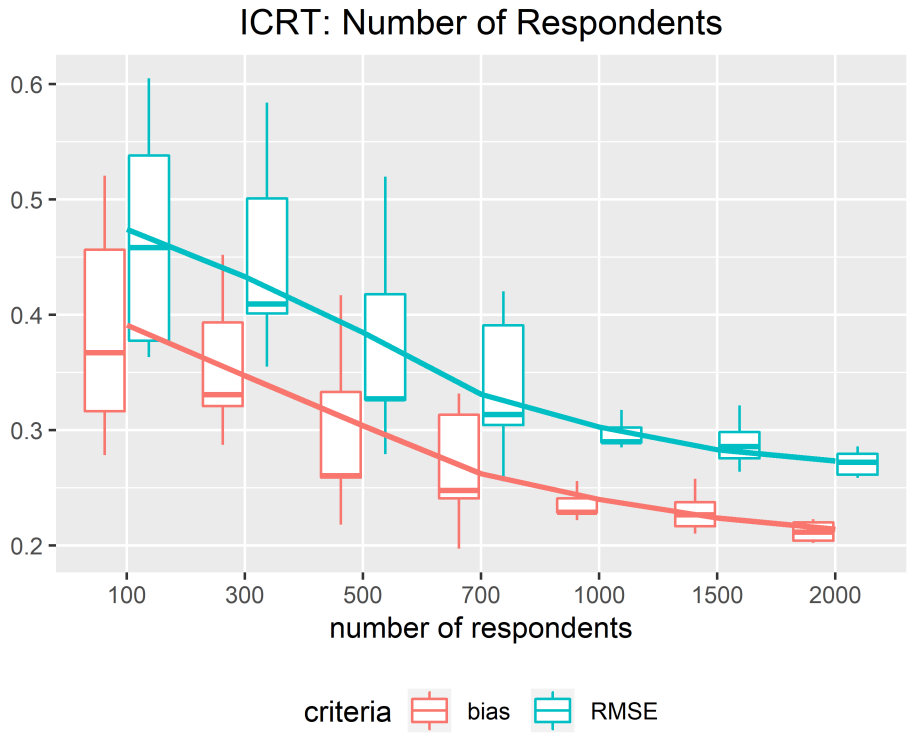

Table 5: ICRT Results for Varying Number of Respondents

	correlation ($\rho(\theta_i, \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
$N = 100$	0.885	0.388	0.469
$N = 300$	0.895	0.357	0.450
$N = 500$	0.926	0.298	0.374
$N = 7000$	0.941	0.266	0.337
$N = 1000$	0.956	0.235	0.297
$N = 1500$	0.958	0.230	0.289
$N = 2000$	0.963	0.212	0.272

[18] ICRT estimates more parameters and requires more computational power than RIRT. Estimating multiple latent traits and using the estimates to predict the LE outcome may make the model more difficult to converge.

Number of IRT Items

Figure 6 shows the simulation results for different numbers of baseline items used to estimate the auxiliary dimensions using IRT. In general, we expect both bias and RMSE to decrease as we ask more baseline items. Having more precise estimates for respondents' non-sensitive attributes increases the accuracy of our predictions for their responses to the list experiment question. This, in turn, helps us estimating the respondent's probability of agreeing with the sensitive item.

As expected, we find a decreasing relationship between the number of baseline items asked and the biases/RMSEs of the estimated target sensitive trait scores. $x = 5$ corresponds to a case asking two baseline items for each of the two auxiliary dimensions ($= 2 * 2 + 1$ LE), while $x = 9$ corresponds to a case asking four baseline items for each of the two auxiliary dimensions ($= 4 * 2 + 1$ LE). Interestingly, adding two additional baseline items per trait to the sample in excess of four does not seem to result in large improvements to the estimates. This can be due to the near-perfect correlation at $x = 9$ as found in Table 5. The model already performs well with four baseline items per auxiliary trait, so adding more baseline items has little room to enhance the estimates.

Figure 6: ICRT Results for Varying Number of Items

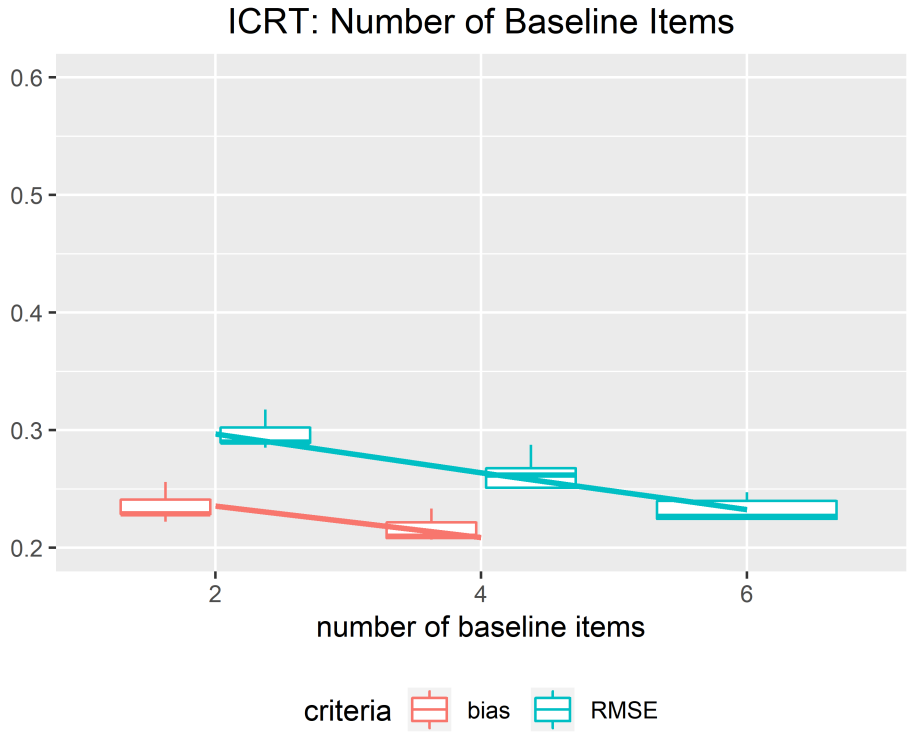

Table 6: ICRT Results for Varying Number of Items

	correlation ($\rho(\theta_i, \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
$m = 5$	0.956	0.235	0.297
$m = 9$	0.965	0.209	0.264
$m = 13$	0.973	0.184	0.233

Group-level Prevalence Rate

Figure 7 and Table 7 display the results of our simulations with respect to the group-level prevalence rates. Similar to our results in the RIRT setting, the overall group-level prevalence rate does not seem to have any impact on ICRT's ability to recover the underlying sensitive attitude or behavior. In fact, while the improvement in RIRT estimates was only very small, there does not appear to be any consistent effect in the ICRT setting. Thus, it appears that both methods are quite robust to variation in overall prevalence rates. Again, similar to the RIRT estimates, all of the estimates are correlated with the true sensitive traits by approximately 0.95.

Figure 7: ICRT Results for Varying the Overall Prevalence Rate

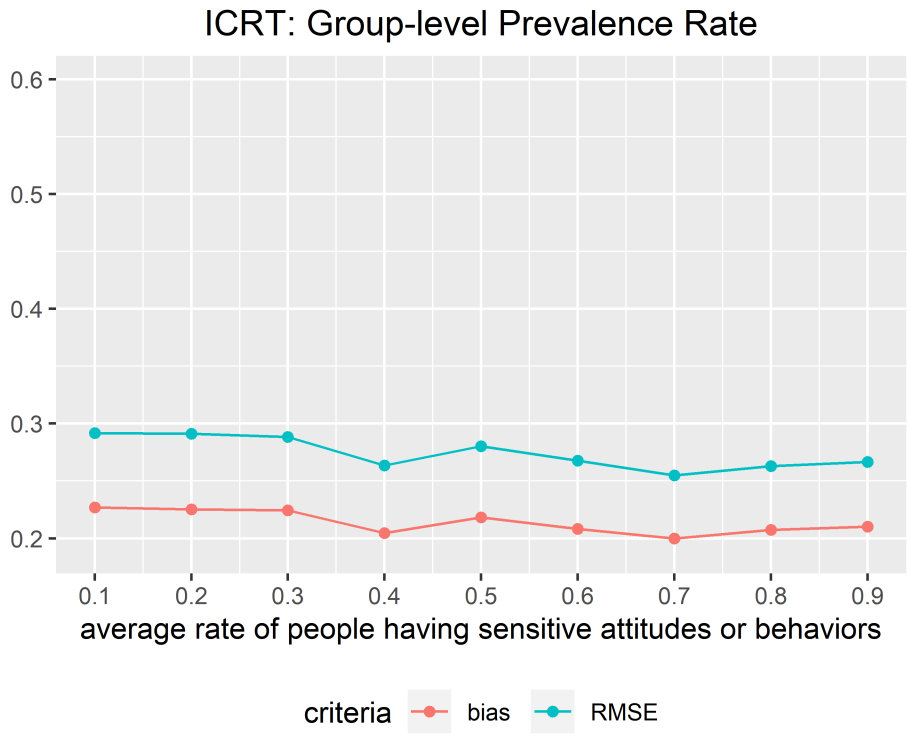

Table 7: RIRT Results for Varying the Overall Prevalence Rate

	correlation ($\rho(\theta_i, \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
$E(\Phi(\theta_i)) = 0.1$	0.957	0.227	0.292
$E(\Phi(\theta_i)) = 0.2$	0.958	0.225	0.291
$E(\Phi(\theta_i)) = 0.3$	0.958	0.224	0.288
$E(\Phi(\theta_i)) = 0.4$	0.965	0.205	0.264
$E(\Phi(\theta_i)) = 0.5$	0.961	0.218	0.280
$E(\Phi(\theta_i)) = 0.6$	0.964	0.208	0.268
$E(\Phi(\theta_i)) = 0.7$	0.967	0.200	0.255
$E(\Phi(\theta_i)) = 0.8$	0.965	0.207	0.263
$E(\Phi(\theta_i)) = 0.9$	0.964	0.210	0.267

Opt-out Option and the Rate of Opting out

Figure 8 and Table 8 describe the ICRT simulations examine the effects of opting out. Different from the RIRT results, the advanced ICRT model accounting for opt-out behavior performs worse than the simple

model in terms of bias and RMSE. That said, the correlations between the estimates and the latent variable of sensitive attitude/traits are quite similar suggesting that this added bias does not strongly impact the results.

At the same time, providing an opt-out option in the LE design substantively means that a researcher would not obtain any information regarding the respondent's sensitive trait. It is because different from the RIRT design, the researcher asks only one LE question. Therefore, what the advanced model does is essentially guessing the missing values using the opted out respondent's auxiliary trait and other personal attributes that the researcher thinks affect her sensitive trait. In general, ICRT is more complex and is susceptible to more uncertainty than RIRT. Thus, the model's attempts to estimate individuals' sensitive attitude or behavior when there is little information about them are unlikely to work very well. In such cases, simply dropping those individuals and running a simple ICRT model can help us obtain more reliable estimates. At the same time, a researcher cannot have the target estimates for the respondents who opted out when she drops the opted out responses, which is another drawback of providing the opt-out option in the first place.

Figure 8: ICRT Results for Varying the Overall Opt-out Rate

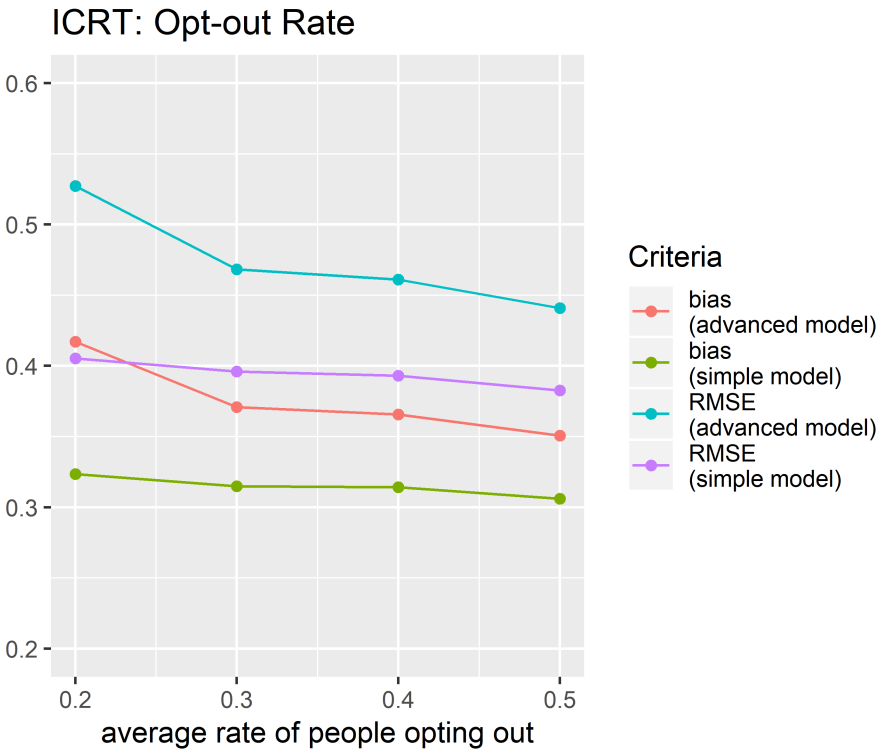

Table 8: ICRT Results for Varying the Overall Opt-out Rate

		correlation ($\rho(\theta_i, \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
advanced model	$E(\pi_i) = 0.2$	0.916	0.417	0.527
	$E(\pi_i) = 0.3$	0.923	0.371	0.468
	$E(\pi_i) = 0.4$	0.926	0.366	0.461
	$E(\pi_i) = 0.5$	0.927	0.351	0.441
simple model	$E(\pi_i) = 0.2$	0.918	0.324	0.405
	$E(\pi_i) = 0.3$	0.922	0.315	0.396
	$E(\pi_i) = 0.4$	0.923	0.314	0.393
	$E(\pi_i) = 0.5$	0.929	0.306	0.383

Ceiling Effect

Figure 9 and Table 9 show the results from our simulations exploring the ceiling effect on the (slightly modified) ICRT model. As can be seen, the ICRT model with the parameter τ handles the deviation of responses quite well. Indeed, the model's performance gets worse as more people choose "all but one are true" when they actually agree with "all items are true." However, as can be seen in Figure 9, the amount of increased bias/RMSE is relatively small. For example, as the rate of people deviating from choosing "all items are true" to "all but one items are true" increases from 10% to 90%, the bias and RMSE increases by approximately 0.1. In other words, even in worst case scenario that almost no one chooses the choice "all items are true," incorporating τ in the ICRT model mitigates the concern well. Therefore, if one is interested in using the ICRT design but is concerned about ceiling effects, incorporating τ is strongly recommended.

Figure 9: ICRT Results for Varying the Degree of Ceiling Effect

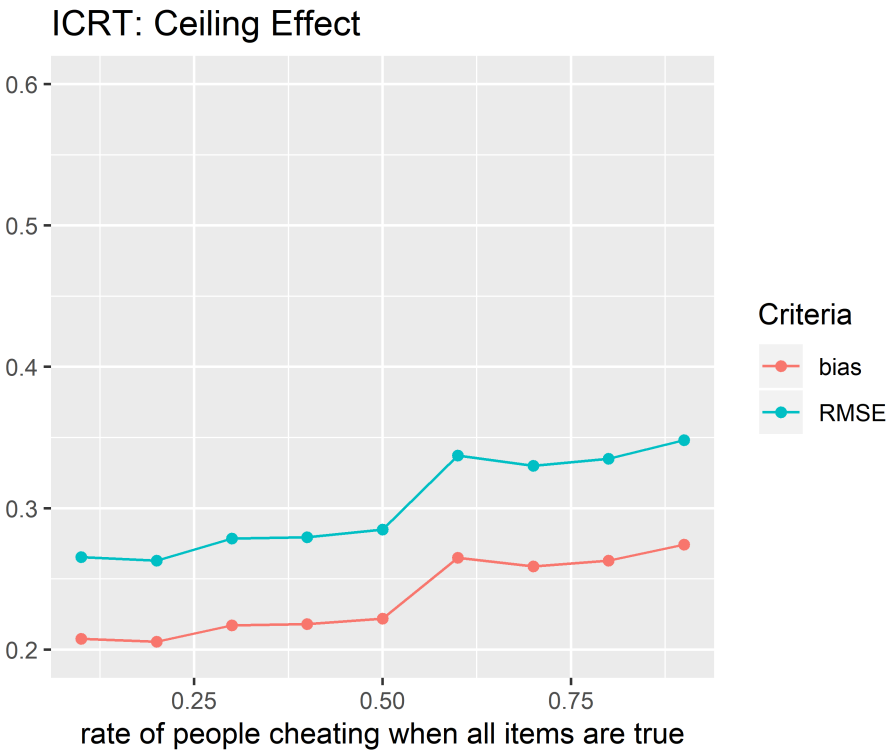

Table 9: ICRT Results for Varying the Degree of Ceiling Effect

	correlation ($\rho(\theta_i, \hat{\theta}_i)$)	bias ($E(\lvert\theta_i - \hat{\theta}_i\rvert)$)	RMSE ($(E((\theta_i - \hat{\theta}_i)^2))^{1/2}$)
$\tau = 0.1$	0.965	0.207	0.265
$\tau = 0.2$	0.965	0.206	0.263
$\tau = 0.3$	0.961	0.217	0.279
$\tau = 0.4$	0.960	0.218	0.280
$\tau = 0.5$	0.959	0.222	0.285
$\tau = 0.6$	0.965	0.265	0.337
$\tau = 0.7$	0.961	0.259	0.330
$\tau = 0.8$	0.960	0.263	0.335
$\tau = 0.9$	0.959	0.274	0.348

Comparisons and Conclusions

Randomized Item Response Theory and the Item Count Response Technique both provide difficult to obtain insights into sensitive questions. Both methods produce low levels of bias and are easily interpretable in many applications. Whereas traditional surveys involving direct questioning are subject to social desirability and low response rates, these methods often elicit honest responses and provide reliable information about the issues researchers care about. In most cases, our results show that both methods provide reliable estimates and are also able to address concerns that are sometimes outside the scope of original models (e.g., opt-out rate and ceiling effects). That said, our discussion and simulations also demonstrate that these methods should be seen as alternatives - each more suitable in unique scenarios - rather than competitors.[19]

Perhaps the clearest indication of their status as alternatives instead of competitors is the way the methods approach understanding the sensitive issue under study. While ICRT focuses on examining a single sensitive behavior, RIRT focuses on aggregating multiple behaviors into a single underlying sensitive trait. Thus, it is clear that the choice between the two should be driven, in large part, by whether one is interested in a specific behavior (use ICRT) or a related pattern of behaviors (use RIRT).

For example, in a case where a specific behavior has significant legal implications, using ICRT may be preferred as it provides estimates for the specific behavior instead of a general measure of an underlying trait. In contrast, in a setting where no single behavior is of specific importance, but consistent patterns of behavior have significant implications, for example, unhealthy pre-natal behaviors, RIRT may be preferred.

That said, it may be possible to combine multiple ICRT list experiment responses using additional IRT modeling and derive measures of a underlying trait motivating all of the behaviors in question. This model would, in effect, combine the insights from ICRT and RIRT models and allow researchers to examine both specific behaviors and general traits at the individual level To our knowledge, no such model currently exists.

Other important consideration that will determine whether one uses RIRT or ICRT include sample size (RIRT performs better in smaller samples (in particular, $N < 100$ in our simulations)), the availability and reliability of randomization devices (ICRT does not require randomization) and whether one is particularly concerned about ceiling effects (RIRT using crosswise RRD is immune to these effects).

That said, numerous other factors remain to be examined before a definitive recommendation of the effectiveness of either method can be given. For instance, both methods can fall victim to systematic non-compliance where, for example, respondents might simply randomly pick responses or always pick the first response. To our knowledge, there is no previous work that systematically compares RIRT and ICRT under such conditions.

In addition, work still needs to be done to understand the effects of providing respondents with an "opt-out" option where they refuse to answer a given question. While one might assume that refusal is positively correlated with the sensitive behavior (i.e., another "design effects" (Blair and Imai 2012, Li 2019), this assumption is not guaranteed and the effects of simply dropping refusals from analysis are largely unknown.

Ultimately, regardless of which method is used, it is clear that the rapid development of methods for assessing sensitive issues has unlocked entirely new possibilities for applied researchers. Indeed, both the methods examined here are easy to implement in an applied setting and require only minimal changes to overall survey design. However, our analysis makes clear that, before doing so, researchers should consider the individual features of each method and determine which is better suited for their application.

References

Ahlquist, John S. 2017. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators." *Political Analysis* 26: 34-53.

[19] In particular, RIRT is quite robust even with a model misspecification problem of not addressing the effects of auxiliary personal attitudes on the target sensitive trait. If we include these attributes in RIRT estimation, its performance enhances considerably and the estimates become much more accurate. The practical question is, how one can identify and measure such attributes in RIRT.

Bararesi, Lucio, Giancarlo Diana, and Pier Francesco Perri. 2014. "Horvitz-Thompson Estimation with Randomized Response and Nonresponse." *Model Assisted Statistics and Applications* 9: 3-10.

Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou. 2015. "Design and Analysis of the Randomized Response Technique." *Journal of the American Statistical Association* 110 (511): 1304-19.

Blair, Graeme, Winston Chou, and Kosuke Imai. 2019. "List Experiments with Measurement Error." *Political Analysis*.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20: 47-77.

Bockenholt, Ulf, and Peter G.M. van der Heijden. 2007. "Item Randomized-response Models for Measuring Noncompliance: Risk-return Perceptions, Social Influences, and Self-protective Responses." *Psychometrika* 72(2): 245-62.

de Jong, Martjn G., and Rik Pieters. 2019. "Assessing Sensitive Consumer Behavior Using the Item Count Response Technique." *Journal of Marketing Research*.

Fox, J.-P. 2005. "Randomized Item Response Theory Models." *Journal of Educational and Behavioral Statistics* 30(2): 1-24.

Fox, J.-P., M. Avetisyan, and J. van der Palen. 2013. "Mixture Randomized Item-response Modeling: A Smoking Behavior Validation Study." *Statistics in Medicine* 32(27): 4821-37.

Fox, J.-P., Duco Veen, and Konrad Klotzke. 2018. "Generalized Linear Mixed Models for Randomized Responses." *Methodology* 15: 1-18.

Hogligner, Marc, and Ben Jann. 2018. "More is Not Always Better: An Experimental Individual-level Validation of the Randomized Response Technique and the Crosswise Model." PLoS ONE 13(8): e0201770.

Li, Yimeng. 2019. "Relaxing the No Liars Assumption in List Experiment Analyses." *Political Analysis*.
Miller, J. 1984. "A New Survey Technique for Studying Deviant Behavior." Ph.D. Dissertation, Sociology Department, The George Washington University.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2015. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60(3): 783-802.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60 (139): 63-9.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. "Two New Models for Survey Sampling with Sensitive Characteristic: design and analysis." *Metrika* 67: 251-63.