Survey Research Methods (2022)
Vol. 16, No. 1, pp. 45-60
doi:10.18148/srm/2022.v16i1.7824
European Survey Research Association

Non-Compliance with Indirect Questioning Techniques: An Aggregate and Individual Level Validation

Thomas Krause
University of Stuttgart
Germany

Andreas Wahl
University of Stuttgart
Germany

Indirect questioning techniques are widely discussed and used as methods to avoid or reduce the effects of social desirability in interview situations on sensitive topics. Nevertheless, current evaluation studies suggest that indirect questioning techniques have a bigger compliance problem than evaluation studies based on the "more is-better" principle would suggest. In our study, we investigate the extent to which question compliance problems can be identified for a variant of the Randomized Response Technique, for the Crosswise Model and Triangular Model. By means of an aggregate and an individual level validation, we examine the response patterns of the participants. Contrary to the actual empirical application context of sensitive topics, we use a non-sensitive question that cannot be distorted by social desirability bias. The resulting "same-is-best" rationale differs from most evaluation studies to date, which work according to the "more-" or "less-is-better" principle. Our analyses are based on the data of a convenience sample of 1277 students in the form of an online survey experiment. The results suggest that the indirect questioning techniques show substantial weaknesses in terms of compliance and encourage further individual level evaluations.

Keywords: crosswise model; triangular model; randomized response technique; compliance; indirect questioning techniques

1 Introduction

The correspondence between social reality and respondents' answers in surveys represents a challenge for empirical social research from a variety of perspectives. Possible problems arise, for example, from a lack of information on the part of the respondent, a lack of motivation for cognition, or problems regarding the understanding of the question. However, these are mostly non-intentional sources of error that are not directly associated with the content-related target dimension of the question. The situation is different for questions on sensitive topics. Sensitive (question) topics in surveys are further understood to be questions that contain behaviours, characteristics or views that violate (at least perceived) social norms, traditional customs, or laws.

In addition to non-content-related sources of bias, sensitive questions include aspects that are directly related to the target dimension of the question. These additional sources are social desirability and the desire for social recognition (Gove & Geerken, 1977). This type of bias is widely referred to as "social desirability bias" (Fisher & Katz, 2000). Here, respondents conceal behaviours, views, and attributes that they perceive as undesirable in order to gain (or not lose) social recognition, or report socially desirable beliefs, characteristics, or actions, even though they do not share, possess, or have actually performed the action in question. Through false or exaggerated and understated statements, respondents try to avoid a bad impression or sanctions in the interview situation and try to present themselves as positively as possible (Paulhus, 2002). Survey research has been working on these content-related sources of bias for several decades, at least since the 1940s (Hyman, 1944; Wolter, 2019). One solution strategy are indirect questioning techniques (IQT) (Tourangeau & Yan, 2007).

Warner (1965) presented with the Randomized Response Technique (RRT) a technique that guarantees the confidentiality of individual answers by means of adding a random component to the responses. The resulting anonymization should encourage respondents to answer more honestly. The Crosswise Model and the Triangular Model (Yu, Tian, & Tang, 2008) can be understood as further developments of the RRT.

Many studies were able to uncover differences between the prevalences asked directly and values estimated using such indirect procedures. Current validation studies, however, increasingly state that these higher prevalences are not actually more reliable values (e.g. Coutts & Jann, 2011; Höglinger, Jann, & Diekmann, 2016; Kirchner, 2015; Meis-

Contact information: Andreas Wahl, Pfaffenwaldring 19, 70569 Stuttgart, (e-mail: andreas.wahl@eni.uni-stuttgart.de)

ters, Hoffmann, & Musch, 2020; Walzenbach & Hinz, 2019). If this were the case, then indirect questioning techniques would generate measurement artifacts. This brings aspects of question understanding and compliance with question instructions into focus. This study intends to further investigate whether respondents follow the question instructions, or whether indirect questioning techniques have a compliance problem.

2 State of research and research questions

2.1 Randomized Response Technique

One of the best-known indirect questioning techniques is the Randomized Response Technique (RRT). It goes back to Warner (1965) and is mainly utilised when questions on sensitive topics are asked. The RRT is intended to prevent socially desirable answering behaviour by guaranteeing additional anonymity (Wolter & Preisendörfer, 2013). In Warner's original version of 1965, respondents are asked two contradictory questions at the same time: "Do you carry the sensitive attribute?" (question A) or "Don't you carry the sensitive attribute?" (question B). Which of the two questions the respondents should answer is determined with a randomization device. This device can be a dice, a coin, a deck of cards or something similar. Because one problem of the Warner-RRT is the low efficiency and/or the high variance of the estimated prevalence (Wolter, 2012), the original design has been revised; meanwhile, there are various variants of the RRT (see Wolter (2012) or Hoffmann (2014)). One of these variants is the RRT with an unrelated question (RRTuq; Greenberg, Abul-Ela, Simmons, and Horvitz (1969)). In this version, question B is not the contradictory version of question A, instead the question is an independent one. For example, one could ask the sensitive question A and ask about a respondent's month of birth as question B. As is the case for all other RRT variants as well, two things are necessary for the unrelated question version to function: First, the respondents must keep secret to which question the randomization device refers them to and second, the researcher must know the probability distribution of the device (Korndörfer, Krumpal, & Schmukle, 2014).

Besides "physical" devices, such as dice etc., other random number generators are possible, like the house number distribution. This distribution follows Benford's law, according to which smaller numbers occur more frequently than larger numbers. The probability that respondents live in houses, whose house numbers begin with the numbers one to four, amounts to approx. 70% (Kundt, 2014). It follows that respondents answer the sensitive question in the unrelated question design (question A) with a probability of 70% and about 30% of them answer the unrelated question (question B). Because the respondents keep secret which of the two questions they answered, it is unknown to anyone else, which question has been answered (Jann, Jerke, & Krumpal, 2012). With this additional anonymity, respondents are expected to be more willing to answer honestly to questions on sensitive topics, since the randomization device allows them to avoid any negative consequences (Jerke & Krumpal, 2013). For the researcher, however, remains the possibility to estimate the prevalence of the sensitive issue in the population "even though no direct link between the observed answers and the variable of interest exists on the individual level" (Jann et al., 2012, p. 33). The following formulas (from Wolter (2012)) can be used to calculate the prevalence, its variance and consequently the standard error and the confidence interval (formulas in Equation 1 and Equation 2; notation in Table 1):

Table 1
Notation

$\hat{\pi}$	prevalence
π_B	prevalence of the unrelated question (RRTuq)
$\hat{\lambda}$	measured total percentage of Yes answers (RRTqu); measured total percentage of option A answers (CM/TM)
p	probability to answer question A (RRTuq); prevalence of the unrelated question (CM/TM)
n	sample size

$$\hat{\pi}_{\mathrm{UQT}} = \frac{\hat{\lambda} - (1-p) \times \pi_B}{p} \tag{1}$$

$$\mathrm{Var}\left(\hat{\pi}_{\mathrm{UQT}}\right) = \frac{\hat{\lambda} \times (1-\hat{\lambda})}{n \times p^2} \tag{2}$$

2.2 Crosswise Model and Triangular Model

The Crosswise Model (CM) as well as the Triangular Model (TM) are variants of the RRT discussed above and have decisive advantages over it and its other variants: First, studies prove that the two techniques are more comprehensible than other indirect questioning techniques (Hoffmann, Waubert de Puiseau, Schmidt, & Musch, 2017). That is especially the case for the RRT, where the interviewing process is considered relatively complicated and therefore can lead to confusion to such an extent that this has a negative effect on the functionality of the technique (Jerke & Krumpal, 2013). Second, unlike other indirect procedures, no random number generator is needed to indicate which question the respondents should answer. This makes these methods particularly interesting for use in online questionnaires or for self-administered questionnaires, since no interviewers need to be present. A third point only applies to the Crosswise Model because a self-protective answer cannot be given, as there is no response strategy that respondents can fall back on to ensure that the sensitive issue is not present ((Coutts

& Jann, 2011; Coutts, Jann, Krumpal, & Näher, 2011)). As a result, there is no “right” answer in the CM, and thus no distortion in favor of the self-protective answer.

The CM uses two questions that are asked at the same time: one of them is the sensitive question, while the other one is a non-sensitive, unrelated question. Both of them can be answered with either “Yes” or “No”. However, there are two necessities regarding the unrelated question: First, it’s crucial that the prevalence, or the occurrence of the “Yes” answer for the question is known and second, the prevalence must not be 0.5. For example, one could use the house number distribution (see above) and ask about the first digits accordingly, so that the prevalence is not 0.5 (Kundt, 2014). What’s special about the CM is that respondents are asked to answer both questions simultaneously and tick either option A or option B in accordance to the technique’s rules.

In the CM, option A should be ticked if both answers are “No” or “Yes”. If the answers differ (“Yes” to one question, “No” to the other) then the respondents should check option B. Because both options in the CM always include a “Yes” answer, there is no way for a respondent to deflect to a self-protective answer, where one would never be associated with the sensitive issue.

Since the respondents answer two questions at the same time, it remains unknown to which question they answered “Yes” or “No”. It also remains unknown whether the sensitive issue applies to the person being questioned or not. However, because the prevalence of the unrelated question is known, it is possible to calculate an estimate of the true prevalence of the sensitive issue in the population. This means, the respondents do not need to fear (social) sanctions because nobody knows if the individual respondent answered the sensitive question with “Yes”. If the sensitive issue should apply for a respondent, then additional anonymity can be assured. The following formulas (Equation 3 to Equation 4) can be used to calculate the prevalences and its variances for the CM (from: Yu et al. (2008); notation see Table 1):

$$\hat{\pi}_{\text{CM}} = \frac{\hat{\lambda} + p - 1}{2p - 1} \tag{3}$$

$$\text{Var}\,(\hat{\pi}_{\text{CM}}) = \frac{\hat{\pi}_{\text{CM}} \times (1 - \hat{\pi}_{\text{CM}})}{n - 1} + \frac{p \times (1 - p)}{(n - 1) \times (2p - 1)^2} \tag{4}$$

The difference between the Crosswise Model and the Triangular Model is minor. Again, two questions are asked at the same time, both of which can be answered with “Yes” or “No”. Of those two questions, one is an unrelated question, whereas the other question carries the sensitive subject. Again, the prevalence of the unrelated question must not be 0.5. As with the CM two options are presented, so that the two questions can be answered simultaneously.

Unlike the CM, in the TM, option A should only be ticked if the answer to both questions is “No”. Can one or both be answered with “Yes”, the respondents are asked to check option B. Because ticking option A means that the respondent does not carry the sensitive issue, the TM, comes with a self-protective answer. Therefore, the Crosswise Model has a slight advantage over the Triangular Model. Because of slightly different checking options available, the formulas for the TM differ a little from the CM formulas. They are as follows (Equation 5 to Equation 6) (from: Yu et al. (2008); notation see Table 1):

$$\hat{\pi}_{\text{TM}} = 1 - \frac{\hat{\lambda}}{1 - p} \tag{5}$$

$$\text{Var}(\hat{\pi}_{\text{TM}}) = \frac{\hat{\pi}_{\text{TM}} \times (1 - \hat{\pi}_{\text{TM}})}{n - 1} + \frac{p \times (1 - \hat{\pi}_{\text{TM}})}{(n - 1) \times (1 - p)} \tag{6}$$

2.3 State of research and research question

Many previous validation studies test indirect questioning techniques under the more- or less-is-better assumption. In general, strong validation studies, where the true prevalence is known, are rather rare because it is difficult to find true values for sensitive topics (Tourangeau & Yan, 2007). The state of research is also rather heterogeneous: Above all, the RRT has been the focus of research so far. The meta-study by Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) shows, for example, that the RRT tends to deliver more valid results than direct questions. However, the functionality of the RRT is clearly influenced by the sensitivity of the topic: Depending on how sensitively the respondents perceive the topic, the more likely it is that the RRT delivers more valid results. Similar results were found in a meta-analysis by Schnell and Thomas (2021), where they observe more pronounced effects for the CM than for direct questions, suggesting more valid results. However, the authors raise concerns because these differences are highly dependent on the target population, meaning that the CM might not be suitable for the general public. For the TM there is no such study available so far.

Studies in this area, with a wide range of topics, show that IQT are somewhat superior to direct questions: de Jong, Pieters, and Fox (2010), Krumpal (2012), Simon, Striegel, Aust, Dietz, and Ulrich (2006) or Solomon, Jacobson, Wald, and Gavin (2007) and Rosenfeld, Imai, and Shapiro (2015) show that more valid results can be obtained using the Randomized Response Technique. A similar picture with respect to the Crosswise and Triangular Model is presented by Jann et al. (2012), Korndörfer et al. (2014), Shamsipour et al. (2014), Hoffmann and Musch (2019), Erdmann (2019), Jerke and Krumpal (2013) or Hoffmann, Meisters, and Musch (2020). All of those studies show that more valid results can be achieved with the CM or TM. In addition, for the CM there is also evidence from a strong validation study: Hoffmann, Diedenhofen, Verschuere, and Musch (2015) show "that the CWM [Crosswise Model; the authors] is convincingly capable of obtaining valid prevalence estimates of sensitive attitudes and behaviours" (Hoffmann et al., 2015, p. 409). What is true for all techniques, however, is that the less- or more-is-better assumption cannot always be upheld. Other studies show that there is either no advantage of using indirect methods, or that they produce less valid results: Ostapczuk, Musch, and Moshagen (2009), Holbrook and Krosnick (2010), Coutts and Jann (2011), Kirchner, Krumpal, Trappmann, and von Hermanni (2013), Wolter and Preisendörfer (2013), John, Loewenstein, Acquisti, and Vosgerau (2018), Höglinger and Jann (2018) or Walzenbach and Hinz (2019) and Götze and Wahl (2020).

Now there are several reasons why IQT produce less valid results than direct questions, or why they cannot correctly replicate the known true value. That is firstly, because the postulated questions are not perceived as sensitive by the respondents. In this case, the techniques are deprived of their advantage, of making sure that the sensitive attribute cannot be traced back to the respondents. Therefore, an indirect method should not produce a better prediction than a direct question, as in such cases no socially desired answering behaviour needs to be feared. Secondly, it can be a confidence problem on the part of the respondents. If these distrust the indirect questioning techniques, this will falsify the results. This can happen in two ways: a) because they either always choose the safe answer, or b) because they do not believe in the assured anonymity and deliberately do not follow the instructions. Both of which lead to false answers. A third reason is that the respondents simply cannot follow the instructions at all. And if the respondents do not understand the instructions, there is but a small chance that they will apply these techniques correctly.

This third problem is the focus in more recent studies, in particular the studies by Höglinger and Diekmann (2017), Höglinger and Jann (2018), Meisters et al. (2020) and Schnapp (2019), but also Walzenbach and Hinz (2019). The key message is that the validation of indirect questioning techniques on the basis of the less- or more-is-better assumption is not always appropriate. Höglinger and Diekmann (2017) as well as Schnapp (2019) show an overreporting with the CM when asking a question for which the prevalence is known to be close to zero. The CM severely overestimates this prevalence, suggesting that it is likely to have a false positive bias. This means, that interpreting higher prevalences as a more valid result is not always appropriate.[1] Furthermore, if respondents do not understand the instructions for the indirect questioning techniques, the chance of false positive rates also increases (Höglinger & Jann, 2018). What makes matters worse is that such false positives can also occur when questions are asked that aren't necessarily sensitive (Walzenbach & Hinz, 2019).[2] In either case the higher estimated prevalences compared to the direct questions cannot be clearly interpreted as a validity criterion "since random ticking resulting from respondent confusion about the question format cannot be ruled out as an alternative explanation" (Walzenbach & Hinz, 2019, p. 1). More likely, the reason for such strange results is that a compliance problem exists. This is shown by Meisters et al. (2020): They demonstrate that the explanation of the indirect questioning technique can influence the estimated prevalences; the better/more detailed the explanations are, the more likely the techniques are to work. This is where our study is based on. The goal is to examine the different indirect questioning techniques to see if the respondents (can) apply them correctly.

2.4 Analytic background on compliance

As a general rule, comprehension, retrieval, judgment, and response selection according to Tourangeau, Rips, and Rasinski (2000) are used as ideal-typical steps in models for response behaviour. In the following, we will show that several factors are relevant for answering sensitive questions, which can lead to suboptimal response patterns.

First, for direct sensitive questions, it is made explicit that social desirability is the central distorting factor. The answers to direct questions on sensitive topics are mainly influenced by two aspects from a (simplified) causal perspective (see Figure 1). The true value of the respondents in regard to the target dimension (path a) and the social desirability factor (path c), which can become relevant with specific values of the respondents (path bc). After the comprehension of the question, the respondents have to retrieve, judge, and map their value on the response categories. Social desirability is primarily relevant in the step of response selection. If the actual value of the respondents (or their mapping on the response categories) goes against the perceived social norm, social desirability favours that the "true answer" is edited.

Indirect questioning techniques aim to eliminate the influence of social desirability when answering survey questions as much as possible. This is done by adding a random com-

[1] Schnapp (2019) argues to adjust the CM for random ticking patterns. However, even with this corrections, overestimations are still present.

[2] The authors used "blood donation" for the non-sensitive question. This seems unproblematic at first. However, the question about donating blood contains a socially desirable answer, because it can be assumed that to donate blood is socially desirable. Thus, if the CM eliminates the desirability bias, then the prevalence for donating blood should be lower with the CM than with the direct question. In the study, however, higher values were obtained through the CM. As further analysis by the authors showed, this is not only due to socially desirable answering behaviour, but also due to random ticking.

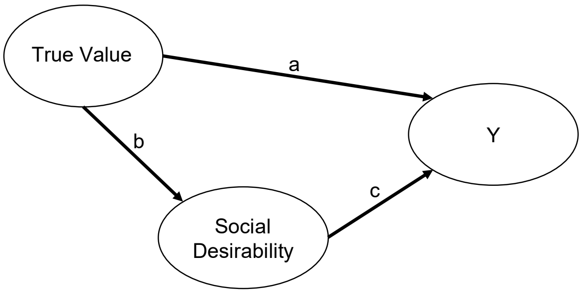


Figure 1. Direct Question Format

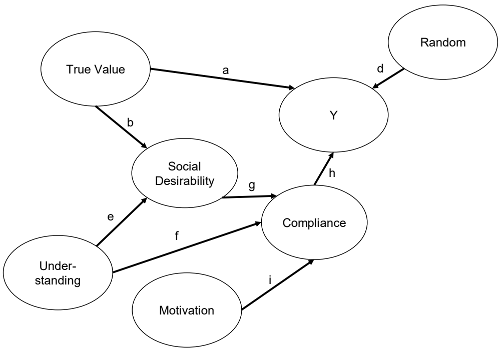


Figure 2. Indirect Question Format

ponent to the individual respondent's answer, which is intended to protect the individual respondent and should make the truthful answer more attractive (Becker, 2006; Esser, 1986). Since the addition of the random component is not a purely arithmetical process but influences the formulation, instructions, and the respondent's way of answering questions, additional factors and paths must be considered. From a conceptual perspective, the influence of social desirability is removed (no path c in Figure 2) by opening a path between the target value y and the random component (path d). Social desirability can therefore at best influence the compliance with the indirect question (path g). Because the distribution of the random component in the aggregate is known, it can be corrected mathematically.

Because the complexity of indirect questioning techniques is higher, it must be assumed that the four-step scheme of Tourangeau et al. (2000) is more elaborate. That is the case, because IQT require the retrieval of information for several questions. Furthermore, the respondents also need to combine this information in a specific way (Jerke, Johann, Rauhut, & Thomas, 2019). Therefore, more effort is necessary in comprehension, retrieval, judgement, and response selection. This complexity in turn brings factors to the front that are not so pronounced in direct question settings, namely: comprehension, motivation (and in part social desirability), and ultimately compliance with the question instructions.

Since the survey question is answered by additional steps or a combination of answer categories, question comprehension plays a significant role. The respondents can only follow and act on the instructions if they understand them (path f) and are motivated to follow them (path i). However, this alone is not a sufficient condition for actual compliance. Even then, respondents can still prefer maximum protection strategies in the interview situation, which in no case can lead to negative social sanctions (path g). This in turn influences compliance through social desirability. At the same time, the influence of social desirability is causally dependent on understanding the indirect questioning technique (path e) as well as the actual value of the respondents (path b).

In addition to these protective strategies, residual motivational factors can also be relevant. The effort associated to comply with the question can simply be refused. A lack of compliance can consequently be based on misunderstanding the instructions, problems in understanding the randomization/anonymization strategy, or unwillingness to follow the instructions (Höglinger et al., 2016) due to motivation or in some cases remaining effects of social desirability.

These influencing factors have not yet been given sufficient attention in research. The reason behind this is, that the detection and quantification of the understanding and the causally downstream compliance is usually not possible because the actual value of the respondents (path a) is unknown. These quantities are not separable in surveys whose primary goal is to capture the aggregated values of respondents. Even comparative validation studies that follow the less/more-is-better principle cannot separate actual "true" values from noncompliance. Similar to the correction for the random component, the true values of the respondents would have to be known in order to estimate the influence of other factors. In the following chapter, we will discuss our analytical considerations for our study design that makes it possible to compensate for this imbalance in order to afterwards derive our working hypothesis.

3 Hypotheses, method and data

3.1 Study rationale and working hypotheses

As a requirement to control for the question understanding and compliance it is necessary to actually know the true value of a respondent, therefore a second measurement is needed. Independent behavioural observations, the selection of persons with known "true" value[3], or a direct (second) question,

[3] Alternatively, a specific value can be created experimentally for validation purposes Moshagen, Hilbig, Erdfelder, and Moritz (2014).

which must not be distorted by the effects of social desirability, are suitable for this purpose. In this study, the last variant was used. By means of a direct question, which was (as far as possible) non-threatening, the actual values of the respondents could be ascertained. Because of this, it is possible to conceptually control all directly influencing quantities on the reported value y itself, except compliance. However, since the random component is known in the aggregate, its influence can be controlled for by taking into account the relevant random deviations (path d). Therefore, the influence of compliance (path h) can be isolated and quantified. Hereby, we are not able to distinguish between the two remaining causes for non-compliance (understanding and motivation). Nonetheless, we believe that our discussion of the causal relationships is instructive and important for interpreting the results and guiding further research.

Due to the survey design as a randomized questionnaire experiment (1) the Randomized Response Technique (unrelated question variant), (2) the Triangular Model, and (3) the Crosswise Model as well as (4) a direct question can be directly compared with each other. This study can therefore, on one hand, be referred to as an aggregate level validation study (Höglinger & Jann, 2018). Although this form of validation is still superior to simple comparative studies, it does not reach the high standard of individual level validation. However, since the actual values of the respondents are also known, this study also uses a strong(er) form of validation (Moshagen et al., 2014): We can directly compare the prevalence estimate of the indirect methods with the "true" value for some of the respondents. This study is therefore on the other hand an individual level validation study (Lensvelt-Mulders et al., 2005). This form of validation is superior to studies, which are only based on the less/more-is-better principle.

The validation logic of "less/more-is-better" is replaced by the principle "same-is-best" in this paper. Since in our study a non-threatening question is used for validation, a match between the direct question and the indirect question would indicate a high level of compliance with the questioning technique under investigation. If, on the contrary, our considerations and previous studies regarding compliance problems are correct, discrepancies between the values obtained can be expected.

We exclude the influence of social desirability (path g) as far as possible through the non-sensitive question, whereby motivation (path i) and understanding (path f) (in addition to the true value) come to the fore as relevant factors. If the "same-is-best" approach shows differences between the techniques, it must be assumed that the respondents do not sufficiently comply with an IQT. Since it must be assumed that social desirability with regard to a sensitive "true" value will only have a negative influence on compliance, a lack of compliance, which already occurs with non-sensitive characteristics, is also a problem for sensitive questions. Our testable working hypotheses are derived from this.

- H1: Differences in the prevalence estimates between the direct question and the indirect questioning techniques are present.
- H2: These differences are biased towards the 50% threshold.

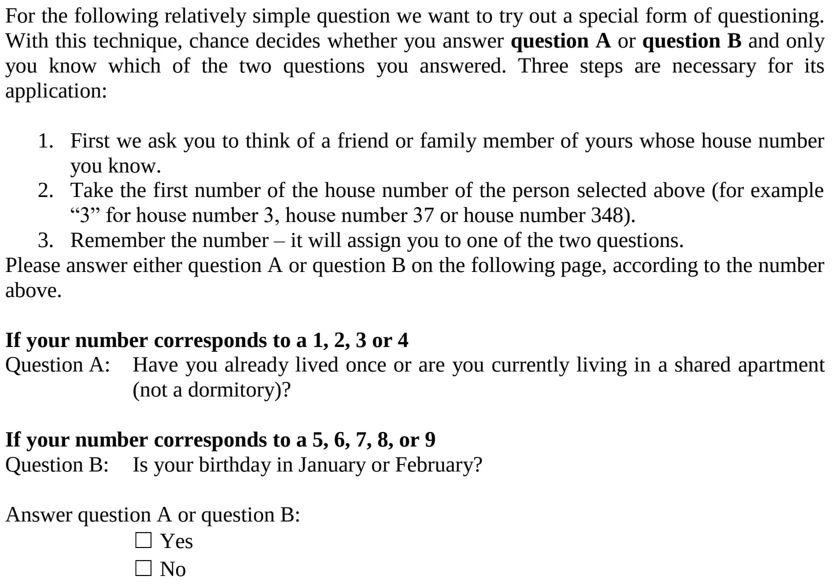
If these are correct, it must be assumed that the differences are due to a lack of compliance and more answers can be observed that resemble a random or arbitrary ticking pattern. If the respondents do not comply with the instructions of the IQT, due to a lack of comprehension (Höglinger et al., 2016) or motivation, and provide random/arbitrary answers, the resulting prevalence estimates should be biased towards 50%. Systematic deviations of the estimated proportion values in the direction of the socially desired characteristic can be largely excluded due to the non-sensitivity of the topic of our question.

3.2 Data

The examined data is based on an online survey, which was carried out in the context of a teaching research project at the University of Stuttgart. The participants were recruited from German university students, who were invited to participate in more than 300 different mailing lists, newsletters, and forwarded messages by secretariats. The survey was conducted from June 17, 2019 to July 14, 2019 using the survey software "Unipark". After the first two weeks, an additional reminder message was sent out. This convenience sample was chosen not only for economic research cost reasons, but also because of the experimental design of the survey. Since our aim is not to infer prevalence rates in a population, but to compare prevalence estimates between four randomized experimental groups, the use of a convenience sample is warranted.

A total of 1685 unique visits were reached, of which 1277 respondents could be analyzed after refusals (388) and early dropouts (20). These respondents were randomly distributed over four questionnaire versions. Each of which comprised a total of 48 question items. The topics covered aspects of study progress, mental health and drug issues, as well as socio-demographic variables. The indirect questioning techniques were varied across the four questionnaire variants so that each experimental group answered a different sequence of indirect (or direct) question formats. The average survey duration was just under 32 minutes.

In order to be able to test whether the indirect questioning techniques have a compliance problem, we needed a question that, firstly, is not perceived as sensitive and, secondly, is as simple as possible. The following figures show the indirect questioning techniques in their exact wording (see Figure 3, Figure 4 and Figure 5). We decided to ask the participants if they ever lived or currently live in a shared apartment. This



For the following relatively simple question we want to try out a special form of questioning. With this technique, chance decides whether you answer **question A** or **question B** and only you know which of the two questions you answered. Three steps are necessary for its application:

1. First we ask you to think of a friend or family member of yours whose house number you know.
2. Take the first number of the house number of the person selected above (for example "3" for house number 3, house number 37 or house number 348).
3. Remember the number – it will assign you to one of the two questions.

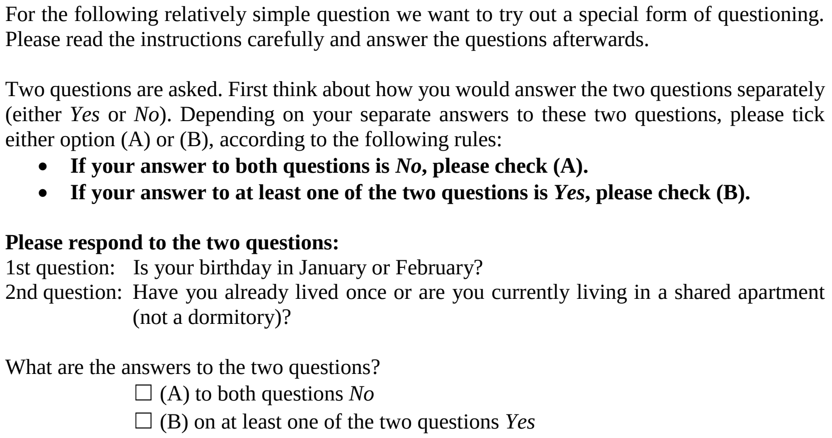
Please answer either question A or question B on the following page, according to the number above.

If your number corresponds to a 1, 2, 3 or 4
Question A: Have you already lived once or are you currently living in a shared apartment (not a dormitory)?

If your number corresponds to a 5, 6, 7, 8, or 9
Question B: Is your birthday in January or February?

Answer question A or question B:
□ Yes
□ No

Figure 3. Unrelated Question Randomized Response Technique (RRTuq)



For the following relatively simple question we want to try out a special form of questioning. Please read the instructions carefully and answer the questions afterwards.

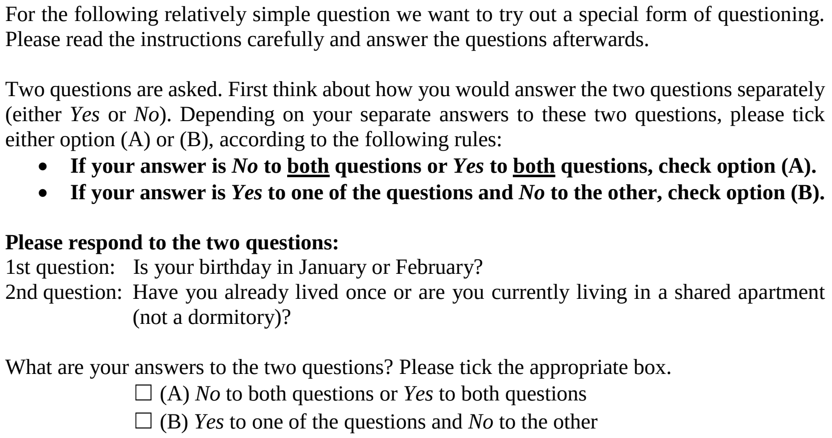
Two questions are asked. First think about how you would answer the two questions separately (either *Yes* or *No*). Depending on your separate answers to these two questions, please tick either option (A) or (B), according to the following rules:

- **If your answer to both questions is *No*, please check (A).**
- **If your answer to at least one of the two questions is *Yes*, please check (B).**

Please respond to the two questions:
1st question: Is your birthday in January or February?
2nd question: Have you already lived once or are you currently living in a shared apartment (not a dormitory)?

What are the answers to the two questions?
□ (A) to both questions *No*
□ (B) on at least one of the two questions *Yes*

Figure 4. Triangular Model



For the following relatively simple question we want to try out a special form of questioning. Please read the instructions carefully and answer the questions afterwards.

Two questions are asked. First think about how you would answer the two questions separately (either *Yes* or *No*). Depending on your separate answers to these two questions, please tick either option (A) or (B), according to the following rules:

- **If your answer is *No* to both questions or *Yes* to both questions, check option (A).**
- **If your answer is *Yes* to one of the questions and *No* to the other, check option (B).**

Please respond to the two questions:
1st question: Is your birthday in January or February?
2nd question: Have you already lived once or are you currently living in a shared apartment (not a dormitory)?

What are your answers to the two questions? Please tick the appropriate box.
□ (A) *No* to both questions or *Yes* to both questions
□ (B) *Yes* to one of the questions and *No* to the other

Figure 5. Crosswise Model

question appears to be suitable for three reasons: First, our sample is a highly educated one and it can be assumed that this question is not too cognitively demanding. Second, a question about a living situation that can be answered with "Yes" and "No" is in itself unproblematic, since it does not specifically ask where exactly someone lives. These answer categories rule out the possibility that an interviewee has to reveal a living situation that he or she does not want to disclose. Third, questions about shared apartments for students are unproblematic, since these arrangements are generally used by students and can therefore be considered a common phenomenon. As an empirical test, to see if the assumption of non-sensitivity holds, we calculated point biserial correlations of the direct question about shared living with question items of a social desirability scale (KSE-G, Kemper, Beierlein, Bensch, Kovaleva, and Rammstedt (2012)). The correlations are small, with a range of the estimators between $[-0.13; 0.02]$. Additionally, none but one of the correlations is statistically significant ($\forall p > 0.028$, see Table A1 in Appendix). This supports the assumption that the question about shared apartments can be classified as non-sensitive and therefore all question variants should yield the same results (i.e. "same-is-best").

The varied direct/indirect questions about (current or past) shared living were asked right at the beginning of the online questionnaire. As can be seen from the question wording in Figure 3, Figure 4 and Figure 5, the three variants of the indirect questions were framed as a practice run for the questioning techniques in order to counteract irritation due to the non-sensitivity of the topic. In contrast to the CM and TM, the RRTuq needed a randomization device that guided the participants to the question they should answer. We decided to make use of Benford's law (see above). This means that nearly 70% of the participants in the RRTuq variation of the questionnaire had to answer question A.

In addition, in the following analyses, we also consider self-reported compliance and perception of the survey questions as alternative explanatory factors that can potentially explain a lack (or even the presence) of correspondence between experimental groups. In these questions, respondents indicated the extent to which they followed the question instructions, how focused they were in answering the questions, whether they answered honestly, and whether they trusted the anonymity assurances (all 5-point rating scale, treated as quasi-continuous measures)[4].

For all the indirect questioning techniques to work, it's crucial that the prevalence, or the occurrence of the "Yes" answer for the unrelated question is known. Here, we ask for the birth month of the respondents: "Is your birthday in January or February?". The prevalence must not be 0.5 and is approximately 0.167 (2 out of 12 months = 0.167)[5].

4 Results

4.1 Aggregate level validation

There are some differences between the prevalence rate for the aggregate level validation, which considers all respondents from the four questionnaire variants with valid values ($N = 1043$).[6] The direct question (DQ) provides a prevalence estimate of 0.591 (Std. Err. = 0.031, $N = 255$), the RRTuq a value of merely 0.504 (Std. Err. = 0.046, $N = 242$), the Crosswise Model an estimate of 0.523 (Std. Err. = 0.044, $N = 295$) and the Triangular Model an estimated value of 0.548 (Std. Err. = 0.037, N = 251). With regard to the absolute amount, the estimates of all indirect questioning techniques are lower than those of the non-sensitive DQ (see Figure 6). All estimates of the prevalence of the indirect questioning techniques are thus shifted towards the 50% threshold.

The significance of these group differences was assessed based on logistic regression models adapted to indirect questioning techniques (van den Hout, van der Heijden, & Gilchrist, 2007), which include the group difference between the questions asked via indirect methods and the direct question as a dichotomous predictor. The model estimations (of the so-called logistic randomized response regression) were performed using the RRlog function of the R package "RReg version 7.1" (Heck & Moshagen, 2018). The difference between DQ and RRTuq is not significant in a likelihood ratio test restricting the prevalence rate to be equal in both formats ($\chi^2(1) = 2.485$, $p = 0.115$, $N = 497$). Also, the prevalence estimates of the CM ($\chi^2(1) = 1.658$, $p = 0.198$, $N = 550$) and the TM ($\chi^2(1) = 0.833$, $p = 0.361$, $N = 506$) do not differ significantly from the prevalence determined by the direct question.

In addition to the bivariate analysis between question technique and prevalence estimation, we also tested alternative explanatory factors for group differences in multivariable models (Table 2). In these models, subjectively reported assessments on response behaviour during the online survey were taken into account. The respondents self-reported on

[4] followed instructions: Do you think that you have followed the specific survey methods correctly in each case?; focused answering: How high was your concentration during the survey?; honest answer: With what degree of honesty did you answer our questions?; trust in anonymity: How much trust do you have in our measures to ensure the anonymity and privacy of the participants in this survey?.

[5] We assume an equal distribution for the birth months of the respondents, as there is no reason to assume that the birth month is a factor that prevents someone from taking part in the survey. We also know that there is no equal distribution of birth months in Germany, although the deviation from one can – in most cases – be ignored. Therefore, the prevalence of 0.167 is only an approximation.

[6] A basic descriptive breakdown of the examined experimental groups can be found in the Appendix Table A3

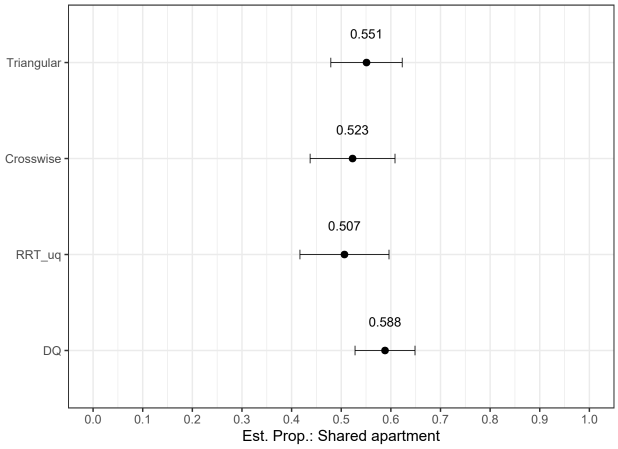


Figure 6. Prevalence estimates of the aggregate level validation with 95% confidence interval. No. of Obs.: DQ = 255; RRTuq = 242; Crosswise = 295; Triangular = 251.

compliance with the survey instructions, on focus while answering, on their honesty in answering the questions, and on subjective trust in the anonymity of the survey. Since these are direct questions about behaviour during the survey, these measurements are to some degree problematic and are presumably subject to the effects of social desirability themselves. Nonetheless, the multivariable analyses show that the lack of significant differences between the RRTuq, CM, TM and the DQ are stable even when the subjectively reported behaviour of the respondents is statistically controlled for. This result supports that the non-existent differences between the questioning techniques are not only due to differences in the response behaviour of the experimental groups. Nevertheless, it must be considered that these control variables are only subjectively reported response behaviour. An actual (not perceived) misunderstanding of the questioning technique or (not reported) lack of motivation cannot be ruled out.

4.2 Individual level validation

One problem of an aggregate level validation is that the actual value of the respondents examined with the indirect procedures is unknown. Thus, "only" randomly assigned groups are compared, instead of a comparison between the direct and the indirect question with one and the same person. For our individual level validation, however, the actual value of the respondents is known.

We used a question that all participants (regardless of variant) had to answer at the beginning of the survey: "In what type of apartment do you currently live?". One of the possible answer categories[7] was "In a shared apartment (not in a dormitory)". For this analysis we only keep cases that gave this answer. Consequently, we know the true prevalence of respondents that are living in a shared apartment: 100%. If subsequently, the questioning techniques used for our non-sensitive question are reliable, then they should yield this prevalence.[8]

Using the answer of another question as a "true" value poses a problem, as the answers to this question could also be untrue. Ultimately, it cannot be ruled out that respondents noted a different value in the question used as filter. However, because this is a sample where living in a shared apartment is an everyday occurrence, the question is not of a sensitive nature, and it doesn't afford a major cognitive effort, the measured value should be reliable. Untrue answering behaviour cannot be ruled out at this moment, but we are quite confident that what we intended to measure has been measured.

Another criticism to this approach is that we only consider respondents, who we know have the characteristic in question. Therefore, we are only able to assess if the IQT lead to false-negatives (i.e. underestimate the true value). An analysis to assess the opposite, however, is not possible because we only know if respondents *currently* live in a shared apartment, but not whether they have lived in a shared apartment before. Thus, we do not have the right true value to make a comparison, as the non-sensitive question includes past living arrangements. Hence, we can only consider false-negatives in our paper.

For the individual level validation, the results can be read off directly from Figure 7.1. The 95% confidence interval contains the true value of the subgroup for the RRTuq, the TM and the direct question[9]. Only the confidence interval of the CM does not contain the "true" value. The point estimate of the prevalence of 86% is also well away from the true value of 100%. The estimates for the RRTuq and the TM are 85.2% and 97%, respectively. Close to the actual value for the TM, but far away, with a wide confidence interval for the RRTuq.

This is different from the results above, but since the actual value of the subgroup is known, this result represents an even stronger form of validation. Due to the small number of cases and the associated low cell numbers, these results could not be reliably investigated in multivariable models.

However, the two-sided mean comparisons from Figure 7.1 hardly reach the liberal threshold of 50% or the stricter one of 80% (Urban & Mayerl, 2018) for statistical power (see Table A2 in Appendix). Only the test for the CM achieves a higher chance of detecting an existing difference instead of not detecting it, with 57% power. Due to the small number of

[7] The possible categories were: a) With my parents/with relatives, b) In a shared apartment (not a dormitory), c) In a student dormitory, d) In an apartment, alone, e) In an apartment with the partner, f) Other.

[8] A basic descriptive breakdown of the examined experimental groups can be found in the Appendix Table A4

[9] For the DQ, however, two respondents gave inconsistent responses, which explains the calculated prevalence value of 98.8%.

Table 2
Multivariable logistic randomized response regression (aggregate level validation

	RRTuq vs DQ			CM vs DQ			TM vs DQ		
	Estim.	SE	p-Value LRT	Estim.	SE	p-Value LRT	Estim.	SE	p-Value LRT
Constant	−0.268	1.390	0.847	0.128	1.394	0.927	1.525	1.331	0.250
not DQ	−0.361	0.224	0.110	−0.285	0.218	0.191	−0.182	0.195	0.350
followed instruction	0.000	0.205	0.998	0.077	0.198	0.698	0.006	0.172	0.969
focused answering	0.040	0.154	0.795	0.012	0.159	0.941	0.051	0.151	0.734
honest answer	0.197	0.262	0.454	−0.022	0.267	0.934	−0.324	0.260	0.211
trust in anonymity	−0.112	0.118	0.337	−0.013	0.116	0.911	0.042	0.114	0.711
No. of Obs.	497			550			506		
LL	-335.028			-376.977			-337.915		

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

cases in the individual level validation, we decided to lower the significance level to 0.1 in order to increase the statistical power of the tests. But here, too, the same pattern of results emerges. The CM deviates significantly from the "true" value, while the confidence intervals of RRTuq and TM still contain the value assumed to be true (see Figure 7.2). But even for this significance level, the tests for TM and RRTuq only reach a power of 40.8% and 49.2% (see also Table A2 in Appendix).

Since the direction of the deviation from the "true" value is fixed for our case (less than 1.0), a one-sided significance test is appropriate. For a one-sided test with 90% confidence interval, all tests achieve a higher statistical power than 50% (RRTuq: 63.6%; Crosswise: 81%; Triangular: 55.3%). As a result, the Crosswise Model ($t = -2.1671$, $df = 103$, $p = 0.01627$, $CI =]-\infty, 0.943]$), the RRTuq ($t = -1.6418$, $df = 60$, $p = 0.05293$, $CI =]-\infty, 0.969]$) and also the Triangular Model ($t = -1.4225$, $df = 86$, $p = 0.07925$, $CI =]-\infty, 0.997]$) differed significantly from the reference value.

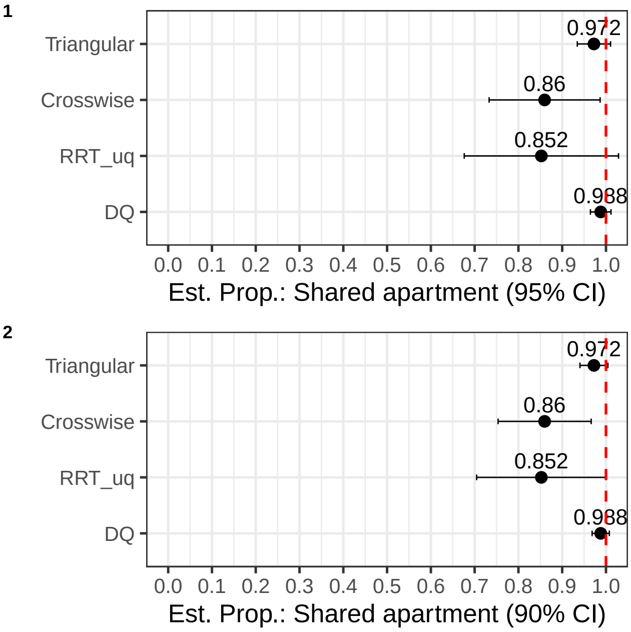


Figure 7. Prevalence estimates of the individual level validation with two-sided confidence intervals.
No. of Obs.: DQ = 83; RRTuq = 61; Crosswise = 104; Triangular = 87.

5 Discussion and conclusion

Our individual and aggregate level validation point in different directions. In the aggregate, no significant differences can be found between the questioning techniques, even if the average point estimates, as expected, deviate in the direction of the 50% threshold. For the individual level, on the other hand, significant deviations from the true value in the direction of the second working hypothesis can be seen. Here, we replicate the negative findings of Höglinger et al. (2016) and Walzenbach and Hinz (2019) regarding the CM. However, the RRTuq and the TM don't show any significant differences from the reference value at the 0.05 level. But it has to be noted that the RRTuq provides an even worse prevalence estimate (85.2%) than the CM (86%) in the individual level validation. Only the broad confidence interval prevents the identification of a significant difference in this case. After our power analysis, we set the significance level at 0.1 with a one-tailed test, which resulted in acceptable levels of statistical power for all IQT. Since the number of cases was low due to the design-related reduction of the sample, and the tests based on it had too little statistical power, lowering the significance level seems reasonable. Based on these tests, we observed differences to the reference value for the RRTuq and the TM also. This indicates compliance problems for all indirect questioning techniques. With the uniform direction of the deviation from the reference value and the significant differences in the individual level validation, our two working hypotheses cannot be rejected and must be regarded as provisionally confirmed.

Since the topic of the question was deliberately chosen so that no effects of social desirability should occur, these results cannot be attributed to social desirability, but rather to a lack of understanding of the question instructions or a lack of motivation to follow them. Even though no empirical distinction can be made between these two causes, our results show that the indirect questioning techniques already exhibit compliance problems for non-sensitive questions.[10] Since we are dealing with a self-recruited student sample with a high level of formal education, it can also be expected that these problems will be even more pronounced in a more heterogeneous and generally less motivated sample. However, without further investigation, this must be regarded as speculative.

Compared to the RRTuq and the CM, the TM shows small deviations from the reference value, which even in the one-sided tests are only significant at the 0.1 level. However, this comparatively good performance is counteracted by the fact that a central weakness of the TM does not come into play for non-sensitive questions. It is criticized for having an evasion strategy, meaning respondents can choose a "safe" answer, which doesn't lead to social sanctions. Since a non-sensitive question was used here, this possible source of bias could not be considered. Therefore, the comparatively good performance can possibly only be maintained in settings with neutral topics and should not be overstated.

A further limitation of this study is that for the individual level validation, we could only consider respondents who we were sure that they have the attribute in question. This only allowed us to examine false-negative response patterns. If, on the other hand, it had been possible to consider persons who certainly do not possess the attribute, this would have placed our results on a broader foundation. However, this was not possible due to the available data. The situation is similar in regard to the number of cases available. It would certainly have been desirable if the examined subgroups had been larger. With less uncertainty about the prevalence estimates, the results for RRTuq, TM and CM would have been clearer in our opinion. However, the relatively small number of cases in combination with the low efficiency was accordingly accompanied by proportionately larger confidence intervals. Through lowering of the significance level and the use of one-sided tests, we were able to counteract this problem, but at the cost of a higher probability of type I errors. Further individual level validation studies with a higher number of cases are therefore still advisable.

Despite these limitations, our analyses supports critical views on compliance with IQT. In our view, our results allow two further conclusions:

1. At this point an individual level validation with direct access to a "ground truth" (Meisters et al., 2020, p. 2) seems indispensable for further studies. The objective should be to take into account socially desirable, undesirable, and neutral characteristics and behaviours. In addition, high, low and prevalence rates close to 50% should be contrasted. Only such an experimental design allows a comprehensive evaluation of the interplay between false-negative and false-positive response patterns, the subject matter of the research question, and the relative prevalence in the population.

2. The mental processing and response generation processes involved in answering indirect questions must be given more attention to. As the results of Meisters et al. (2020) suggest, the investigation of heterogeneous effects in subpopulations (e.g. age or education level) or different administrative modes of question formulation can help investigate these processes. Nevertheless, in addition to purely quantitative comparative studies, a combination with qualitative survey methods is also possible and desirable. Methods such as the think aloud method or cognitive interviews (see e.g. Jerke et al. (2019)) can be used profitably within the context of neutral question topics that are not distorted by the effects of social desirability.

In conclusion, it must be noted that more recent evaluation studies (e.g. Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; Meisters et al., 2020; Walzenbach & Hinz, 2019), as well as our results, suggest that the elegance and logical consistency of indirect questioning techniques do not automatically lead to the desired results. Here, practical aspects such as trust, understanding, and the motivation of the respondents in the interview process play an important role. Whether or under which circumstances the logical stringency of the questioning technique and practical empirical applicability of indirect questioning techniques can be reconciled, however, does not seem to have been conclusively answered yet.

References

Becker, R. (2006). Selective response to questions on delinquency. *Quality and quantity*, *40*(4), 483–498.

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), 169–193.

Coutts, E., Jann, B., Krumpal, I., & Näher, A.-F. (2011). Plagiarism in student papers: Prevalence estimates using special techniques for sensitive questions. *Jahrbücher für Nationalökonomie und Statistik*, *231*(5-6), 749–760.

[10] Schnapp (2019) argues that non-intrinsically motivated participants should comply less often with the questions instructions, which should accentuate biased results. As our study is thematically tailored to our sample, and we did not use incentives, we can be relatively confident that we eliminate non-compliance because of lack of motivation as much as possible.

de Jong, M., Pieters, R., & Fox, J.-P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, *47*(1), 14–27.

Erdmann, A. (2019). Non-randomized response models: An experimental application of the triangular model as an indirect questioning method for sensitive topics. *Methods, Data, Analyses*, *13*(1), 139–167. doi:10.12758/MDA.2018.07

Esser, H. (1986). *Können Befragte lügen? Zum Konzept des "wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung*. Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-.

Fisher, R. J., & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. *Psychology and Marketing*, *17*(2), 105–120.

Götze, A., & Wahl, A. (2020). Psychische Gesundheit: Eine heikle Thematik in der empirischen Umfrageforschung? Zur Validierung des Crosswise Modells. *Schriftenreihe des Instituts für Sozialwissenschaften*, *49*. doi:10.18419/opus-11083

Gove, W. R., & Geerken, M. R. (1977). Response bias in surveys of mental health: An empirical investigation. *American Journal of Sociology*, *82*(6), 1289–1317. doi:10.1086/226466

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, *64*(326), 520. doi:10.2307/2283636

Heck, D. W., & Moshagen, M. (2018). RRreg: an R package for correlation and regression analyses of randomized response data. *Journal of Statistical Software*, *85*(2), 1–29. doi:10.18637/jss.v085.i02

Hoffmann, A. (2014). *Indirekte Befragungstechniken zur Kontrolle sozialer Erwünschtheit in Umfragen* (Dissertation, Heinrich-Heine-Universität Düsseldorf, Düsseldorf). Retrieved from https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=32837

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology*, *62*(6), 403–414. doi:10.1027/1618-3169/a000304

Hoffmann, A., Meisters, J., & Musch, J. (2020). On the validity of non-randomized response techniques: An experimental comparison of the crosswise model and the triangular model. *Behavior Research Methods*, *52*(4), 1768–1782. doi:10.3758/s13428-020-01349-9

Hoffmann, A., & Musch, J. (2019). Prejudice against women leaders: Insights from an indirect questioning approach. *Sex Roles*, *80*(11-12), 681–692. doi:10.1007/s11199-018-0969-6

Hoffmann, A., Waubert de Puiseau, B., Schmidt, A. F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods*, *49*(4), 1470–1483. doi:10.3758/s13428-016-0804-3

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the crosswise-model RRT. *Political Analysis*, *25*(1), 131–137.

Höglinger, M., & Jann, B. (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. doi:10.1371/journal.pone.0201770

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. In *Survey research methods* (Vol. 10, pp. 171–187).

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

Hyman, H. (1944). Do they tell the truth? *Public Opinion Quarterly*, *8*(4), 557–559. Retrieved from http://www.jstor.org/stable/2745311

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. *Public Opinion Quarterly*, *76*(1), 32–49. doi:10.1093/poq/nfr036

Jerke, J., Johann, D., Rauhut, H., & Thomas, K. (2019). Too sophisticated even for highly educated survey respondents? A qualitative assessment of indirect question formats for sensitive questions. *Survey Research Methods*, *13*(3), 319–351. doi:10.18148/srm/2019.v13i3.7453

Jerke, J., & Krumpal, I. (2013). Plagiate in studentischen Arbeiten: Eine empirische Untersuchung unter Anwendung des Triangular Modells. *methoden, daten, analysen*, *7*(3), 347–368. doi:10.12758/mda.2013.017

John, L. K., Loewenstein, G., Acquisti, A., & Vosgerau, J. (2018). When and why randomized response techniques (fail to) elicit the truth. *Organizational Behavior and Human Decision Processes*, *148*, 101–123. doi:10.1016/j.obhdp.2018.07.004

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G). *1869-0491*, *2012/25*.

Kirchner, A. (2015). Validating sensitive questions: A comparison of survey and register data. *Journal of Official Statistics*, *31*(1), 31–59. doi:10.1515/jos-2015-0002

Kirchner, A., Krumpal, I., Trappmann, M., & von Hermanni, H. (2013). Messung und Erklärung von Schwarzarbeit in Deutschland – Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit. *Zeitschrift für Soziologie*, *42*(4), 291–314. doi:10.1515/zfsoz-2013-0403

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology*, *45*, 18–32.

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-semitism in germany: A comparison of randomized response and direct questioning. *Social Science Research*, *41*(6), 1387–1403. doi:10.1016/j.ssresearch.2012.05.015

Kundt, T. C. (2014). Applying 'benford's law' to the crosswise model: Findings from an online survey on tax evasion. *SSRN Electronic Journal*. doi:10.2139/ssrn.2487069

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research. *Sociological Methods & Research*, *33*(3), 319–348. doi:10.1177/0049124104268664

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? *PLOS One*, *15*(6), e0235403. doi:10.1371/journal.pone.0235403

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology*, *61*(1), 48–54. doi:10.1027/1618-3169/a000226

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, *39*(6), 920–931. doi:10.1002/ejsp.588

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun (Ed.), *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Erlbaum.

Rosenfeld, B., Imai, K., & Shapiro, J. N. (2015). An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science*, *60*(3), 783–802. doi:10.1111/ajps.12205

Schnapp, P. (2019). Sensitive question techniques and careless responding: Adjusting the crosswise model for random answers. *Methods, Data, Analyses*, *13*(2). doi:10.12758/mda.2019.03

Schnell, R., & Thomas, K. (2021). A meta-analysis of studies on the performance of the crosswise model. *Sociological Methods & Research*, 004912412199552. doi:10.1177/0049124121995520

Shamsipour, M., Yunesian, M., Fotouhi, A., Jann, B., Rahimi-Movaghar, A., Asghari, F., & Akhlaghi, A. A. (2014). Estimating the prevalence of illicit drug use among students using the crosswise model. *Substance Use & Misuse*, *49*(10), 1303–1310. doi:10.3109/10826084.2014.897730

Simon, P., Striegel, H., Aust, F., Dietz, K., & Ulrich, R. (2006). Doping in fitness sports: Estimated number of unreported cases and individual probability of doping. *Addiction (Abingdon, England)*, *101*(11), 1640–1644. doi:10.1111/j.1360-0443.2006.01568.x

Solomon, J., Jacobson, S. K., Wald, K. D., & Gavin, M. (2007). Estimating illegal resource use at a Ugandan Park with the randomized response technique. *Human Dimensions of Wildlife*, *12*(2), 75–88. doi:10.1080/10871200701195365

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

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

Urban, D., & Mayerl, J. (2018). *Angewandte Regressionsanalyse: Theorie, Technik und Praxis* (5., überarbeitete Auflage). doi:10.1007/978-3-658-01915-0

van den Hout, A., van der Heijden, P. G., & Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics & Data Analysis*, *51*(12), 6060–6069. doi:10.1016/j.csda.2006.12.002

Walzenbach, S., & Hinz, T. (2019). Pouring water into wine: Revisiting the advantages of the crosswise model for asking sensitive questions. *Survey Methods: Insights from the Field (SMIF)*. doi:10.13094/SMIF-2019-00002

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69. doi:10.1080/01621459.1965.10480775

Wolter, F. (2012). *Heikle Fragen in Interviews*. doi:10.1007/978-3-531-19371-7

Wolter, F. (2019). A new version of the item count technique for asking sensitive questions: Testing the performance of the person count technique. *Methods, Data, Analyses*, *13*(1). doi:10.12758/MDA.2018.04

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research*, *42*(3), 321–353.

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*(3), 251–263. doi:10.1007/s00184-007-0131-x

Appendix
Tables

Table A1
Point biserial correlation between DQ and SD-Scale items

Item-id: shortened Itemtext	point biserial corr	p-value
sd1: taken advantage of someone in the past	−0.087	0.158
sd2: always friendly and polite to others	0.010	0.868
sd3: only help people if I expect to get something	0.021	0.738
sd4: always remain objective	0.021	0.708
sd5: occasionally thrown litter away	−0.135	0.028
sd6: always listen carefully	0.010	0.872

Note. Items match the social desirability scale (KSE-G) from Kemper, Beierlein, Bensch, Kovaleva, and Rammstedt (2012)

Table A2
Power Analysis for Individual Level Validation

RRTuq	
Cohen's d	0.210
N	61
power for 95% twosided	0.365454
power for 90% twosided	0.491943
power for 90% onesided	0.636461
Crosswise	
Cohen's d	0.212
N	104
power for 95% twosided	0.574137
power for 90% twosided	0.694326
power for 90% onesided	0.809714
Triangular	
Cohen's d	0.153
N	87
power for 95% twosided	0.290402
power for 90% twosided	0.408792
power for 90% onesided	0.553358

Table A3
Descriptive Statistics for Aggregate Level Validation (N = 1043)

	DQ (N=255)		RRT (N=242)		CM (N=295)		TM (N=251)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
age	22.56	4.12	22.07	3.98	22.37	3.70	22.64	4.30
sexratio	0.28	0.45	0.23	0.42	0.32	0.47	0.32	0.47
semester	5.08	2.86	4.84	2.94	5.31	3.27	5.18	2.86
focused answering	3.94	0.71	3.95	0.72	4.01	0.62	3.94	0.64
honest answer	4.82	0.40	4.78	0.47	4.78	0.42	4.79	0.41
intrest in survey	3.79	0.81	3.71	0.82	3.65	0.89	3.77	0.81
trust in anonymity	4.18	0.90	4.05	0.98	4.04	0.95	4.15	0.81
followed instructions	4.60	0.54	4.68	0.50	4.64	0.55	4.56	0.63
sd1	2.60	1.16	2.56	1.15	2.63	1.17	2.58	1.14
sd2	3.38	0.93	3.34	0.94	3.33	0.94	3.36	0.97
sd3	1.94	0.85	2.15	0.95	2.06	0.90	2.08	0.96
sd4	3.01	0.95	2.96	0.99	2.86	1.02	2.89	1.00
sd5	2.07	1.33	2.06	1.27	2.14	1.32	2.17	1.35
sd6	3.79	0.78	3.69	0.81	3.72	0.77	3.68	0.83

Table A4
Descriptive Statistics for Individual Level Validation (N = 335)

	DQ (N=83)		RRT (N=61)		CM (N=104)		TM (N=87)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
age	22.28	3.49	22.00	3.71	22.27	2.83	22.13	3.14
sexratio	0.30	0.46	0.24	0.43	0.27	0.45	0.33	0.47
semester	5.41	3.10	4.54	2.96	5.95	3.53	5.15	2.90
focused answering	3.99	0.76	3.92	0.74	3.97	0.51	3.90	0.73
honest answer	4.86	0.35	4.74	0.44	4.75	0.44	4.71	0.46
intrest in survey	3.73	0.75	3.84	0.69	3.77	0.92	3.77	0.84
trust in anonymity	4.22	0.84	4.02	0.97	4.07	0.98	4.22	0.71
followed instructions	4.63	0.51	4.64	0.55	4.67	0.51	4.55	0.59
sd1	2.71	1.22	2.67	1.11	2.70	1.19	2.61	1.12
sd2	3.40	0.97	3.38	0.95	3.36	0.98	3.48	0.95
sd3	1.94	0.83	2.10	0.77	2.13	0.89	2.15	1.02
sd4	3.05	0.90	3.03	1.00	2.85	1.01	2.92	1.01
sd5	2.33	1.52	2.21	1.36	2.12	1.32	2.32	1.37
sd6	3.75	0.84	3.51	0.77	3.63	0.76	3.72	0.83
currently living in... (filter)	2.00	0.00	2.00	0.00	2.00	0.00	2.00	0.00