

**Nonresponse Bias in Randomized Controlled Experiments in
Criminology Putting the Queensland Community Engagement Trial
(QCET) Under a Microscope**

Author

Antrobus, Emma, Elffers, Henk, White, Gentry, Mazerolle, Lorraine

Published

2013

Journal Title

Evaluation Review

Version

Accepted Manuscript (AM)

DOI

[10.1177/0193841X13518534](https://doi.org/10.1177/0193841X13518534)

Rights statement

Emma Antrobus et al, Nonresponse Bias in Randomized Controlled Experiments in Criminology: Putting the Queensland Community Engagement Trial (QCET) Under a Microscope, Evaluation Review, 2013, Vol 37(3-4) 197-212. Copyright 2014 The Authors. Reprinted by permission of SAGE Publications.

Downloaded from

<http://hdl.handle.net/10072/69723>

Griffith Research Online

<https://research-repository.griffith.edu.au>

Non-Response Bias in Randomized Controlled Experiments in Criminology: Putting the Queensland Community Engagement Trial (QCET) Under a Microscope

Emma Antrobus, Henk Elffers, Gentry White, & Lorraine Mazerolle

A paper submitted to the Special Issue of *Evaluation Review: Advances in Randomized Experiments and Quasi-Experiments in Criminology and Criminal Justice*

Running header: Non-Response Bias

Author Information:

Emma Antrobus, PhD, Research Fellow, Institute for Social Science Research and ARC Centre of Excellence in Policing and Security (CEPS), The University of Queensland, St Lucia, Queensland, Australia.

Email: e.antrobus@uq.edu.au

Henk Elffers, PhD, Professor, Empirical Research into Criminal Law enforcement, VU University of Amsterdam, The Netherlands and Visiting Fellow at the School of Criminology and Criminal Justice, Griffith University, Mt Gravatt, Queensland, Australia.

Email: HElffers@nscr.nl

Gentry White, PhD, Senior Research Fellow Mathematical Sciences School, Science and Engineering Faculty, Queensland University of Technology, Brisbane, Queensland, Australia and Honorary Senior Research Fellow Institute for Social Science Research and ARC Centre of Excellence in Policing and Security (CEPS), The University of Queensland, St Lucia, Queensland, Australia.

Email: gentry.white@uq.edu.au

Lorraine Mazerolle, PhD, Professor, Institute for Social Science Research and ARC Centre of Excellence in Policing and Security (CEPS), The University of Queensland, St Lucia, Queensland, Australia

Email: l.mazerolle@uq.edu.au

Corresponding Author:

Dr Emma Antrobus

Research Fellow

Institute for Social Science Research and

ARC Centre of Excellence in Policing and Security (CEPS)

The University of Queensland

Level 4, GPN3 Building, Campbell Road

ST LUCIA QLD 4072

Australia

Tel: 61 7 3346 9306 Fax: 61 7 334 67646

Email: e.antrobus@uq.edu.au

Acknowledgements:

The research reported in this paper was funded, in its entirety, by the Australian Research Council (ARC) Centre of Excellence in Policing and Security (CEPS). The authors thank the team of researchers from the University of Queensland (Institute for Social Science Research) and Griffith University who participated in a variety of ways to bring this trial to fruition. The partnership between the research team and the Queensland Police Service is particularly acknowledged. The views expressed in this material are those of the authors and are not those of the Queensland Police Service. Responsibility for any errors of omission or commission remains with the authors. The Queensland Police Service expressly disclaims any liability for any damage resulting from the use of the material contained in this publication and will not be responsible for any loss, howsoever arising, from use or reliance on this material.

Abstract

Objectives: The goal of this paper is to examine whether or not the results of the Queensland Community Engagement Trial (QCET) – a Randomized Controlled Trial that tested the impact of procedural justice policing on citizen attitudes towards police – were affected by different types of non-response bias.

Method: We use two methods (Cochrane and Elffers methods) to explore non-response bias: first, we assess the impact of the low response rate by examining the effects of non-response group differences between the experimental and control conditions and pooled variance under different scenarios. Second, we assess the degree that item response rates are influenced by the control and experimental conditions.

Results: Our analysis of the QCET data suggest that our substantive findings are not influenced by the low response rate in the trial. The results are robust even under extreme conditions, and statistical significance of the results would only be compromised in cases where the pooled variance was much larger for the non-response group and the difference between experimental and control conditions was greatly diminished. We also find that there were no biases in the item response rates across the experimental and control conditions.

Conclusion: RCTs that involve field survey responses – like QCET – are potentially compromised by low response rates and how item response rates might be influenced by the control or experimental conditions. Our results show that the QCET results were not sensitive to the overall low response rate across the experimental and control conditions and the item response rates were not significantly different across the experimental and control groups. Overall, our analysis suggests that the results of QCET are robust and any biases in the survey responses do not significantly influence the main experimental findings.

Keywords: experiments, non-response bias, survey response rates

Introduction

Randomized control trials (RCTs) are a powerful tool for comparing the effect of two different treatments on a population. In a well-designed and executed experiment, the observed differences between the treatment and control group can be attributed, with little doubt, to the treatment. In the case of RCTs, it is assumed that the random allocation of subjects to either the treatment or control groups controls for, by design, the effects of other factors. While researchers can do their best to eliminate bias and minimize random error in the design and execution of an experiment, the problems of response bias can be a serious issue for randomized control trials – particularly those conducted in field settings – when the measurements are made from post-intervention surveys.

Post intervention surveys offer an important way for researchers to measure differences between treatment and control groups in field experiments. Both control and experimental condition participants are asked to complete a survey and their responses are then used to assess whether or not differences exist between the experimental and control group respondents. Response bias can emerge, however, from low response rates. Indeed, post intervention surveys administered under randomized field trial conditions with low response rates can introduce systematic bias in two ways: first, if the effect of the experimental condition influences the propensity to respond the responses are biased. Second, item response rates can be biased if they are influenced by whether or not the respondent is in the control and experimental conditions. Both of these response bias problems potentially emerge in post intervention surveys with low overall response rates.

In this paper we examine and test for two types of response bias in the context of the Queensland Community Engagement Trial (QCET): a randomized field trial that sought to assess the impact of procedural justice policing on citizen attitudes to police using survey methods (see Mazerolle Bennett, Antrobus, & Egghins, 2012). The QCET is a randomized controlled trial of the impact of procedurally just policing (see Mazerolle, Antrobus, Bennett, & Tyler, 2013; Mazerolle, et al., 2012). It involved surveying drivers after they had come into contact with police officers, who administered a random breath test (RBT) using either a procedurally just script or as routine. The participants received a postage paid survey after the RBT to complete and return on their own. As a result of this method, the non-response rate in QCET was high (nearly 85%). In the first instance, we consider the effects of general non-response bias (or the propensity to respond) on the substantive QCET findings. Using the Cochran Method, we explore the effect of low response rates on the statistical significance of QCET results by comparing two sample means of individual constructs, where the construct itself is an arithmetic mean of several items. In the second case, we use the Elffers Method to consider whether or not the QCET effects found in previous published papers (see Mazerolle et al. 2012; 2013) contained differential non-response bias, where item response rates varied by experimental condition between treatment and control. We discuss the potential response biases under these two different models. We conclude that is highly unlikely that the response rate affected the substantive findings of the QCET.

The Queensland Community Engagement Trial (QCET) Design and Response Rates

The QCET was an empirical test of the effects of procedural justice through a routine police-citizen encounter (RBT) (see Mazerolle et al., 2012, 2013). Research consistently shows that if people perceive police officers' behavior to be procedurally fair then they tend to see the police as legitimate, and thereby are more willing to obey the law (see Kane, 2005; Tyler, 1990, 2003; Tyler and Huo, 2002, Wolfe, 2011). The trial tested this theory using a randomized field trial design to examine whether the existing RBT procedure, in conjunction with procedural justice components, could increase perceptions of procedural justice and therefore police legitimacy. The trial involved 60 stationary RBT operations, which were randomly allocated to a control or experimental condition, and surveyed almost 21,000 drivers about their perceptions of and their encounters with the police during the RBT.

In the experimental condition, RBTs were conducted using elements of procedural justice in addition to the standard RBT procedure. Police officers provided drivers with information about the purpose of RBT and the impact of drink-driving, crime prevention strategies, and gave drivers the opportunity to ask questions (see Mazerolle et al. 2012 for more detail). In the control condition, RBTs were conducted following the normal police operating procedures. The experimental and control groups were compared to determine whether the intervention had changed drivers' views of drink-driving and the police, and whether there were differences in perceptions of fairness, respect, compliance, satisfaction, confidence and trust, both in terms of the specific RBT encounter, as well as more general perceptions of the police.

However, the response rates for the study were quite low, due to a number of factors, not least the inability of researchers to track drivers who had received a survey in order to follow up with them. Of the 20,985 surveys distributed to drivers, there were only 2746 valid responses (response rate = 13.09 percent). The response rate for both groups was similar: 12.21 percent of the experimental (N = 1,097) and 13.74 percent of the control condition drivers (N = 1,649) returned surveys. As noted in Mazerolle et al. (2012), more surveys were distributed to drivers in the control condition (400 per operation, as opposed to 300 per operation in the experimental condition). This was largely due to the need to generate a minimum of 30 responses per RBT operation where, after conducting a pilot test, we estimated a worse-case scenario of a 10% response rate for the actual trial. Given the increased time taken for the experimental encounters, we knew we could only generate a maximum of 300 traffic stops per experimental operation in a regular eight hour shift.

The Problem of General Non-Response Bias

The integrity of survey responses, and the conclusions drawn from them, is dependent on the voluntary participation of respondents. Research shows that participant surveys used in experimental research in criminology have a response rate of around 60 to 70% (based on studies reported in the Lum matrix; see Lum, Koper, & Telep, 2011). In contrast, surveys of the general public tend to have lower response rates than targeted surveys, with response

rates of 10% not uncommon (Baruch & Holtom, 2008). This variation in response rates can be explained by several factors, with features such as survey length, multiple contacts, incentives, and pre-notification about the survey all increasing the propensity for a respondent to participate in the survey (Cook, Heath, & Thompson, 2000; Fox, Crask, & Kim, 1988). In cases where these options are unavailable to researchers however, lower response rates not only reduce statistical power, but can introduce systematic bias when the effect of the experimental condition on the propensity to respond might not have “cancelled out” between the treatment and control groups.

Non-response is addressed by noting that the statistics of interest are sample means. Extending the results in Cochran (1963), the true sample difference can be evaluated under different scenarios. Non-response for a sample can be considered as a stratified sample, where the two strata are the response group and the non-response group. In this case the true sample mean is

$$\bar{Y} = w\bar{Y}_R + (1-w)\bar{Y}_{NR} \quad (1)$$

where w is the response rate, or the proportion of subjects who responded. Consider the difference between a treatment and control group:

$$\bar{Y}_T - \bar{Y}_C = \left[w_T \bar{Y}_{TR} + (1-w_T) \bar{Y}_{TNR} \right] - \left[w_C \bar{Y}_{CR} - (1-w_C) \bar{Y}_{CNR} \right] \quad (2)$$

In the case where $w_C \approx w_T = w$, that is, the response rates are approximately equal for the treatment and control groups, (2) simplifies to

$$\begin{aligned} \bar{Y}_T - \bar{Y}_C &= w(\bar{Y}_{TR} - \bar{Y}_{CR}) - (1-w)(\bar{Y}_{TNR} - \bar{Y}_{CNR}) \\ &= w\Delta\bar{Y}_R - (1-w)\Delta\bar{Y}_{NR} \end{aligned} \quad (3)$$

Considering the issue of non-response bias, researchers are more interested in determining whether or not the differences in sample means for the non-response group would alter the statistical significance of the results. Differences in sample means for the treatment and control groups in an experiment are assessed using a two-sample t -test. Evaluating the test statistic for this test requires both the difference in sample means for the non-response data, and the pooled variance for the difference. The difference for the entire sample is calculated using (2), and the pooled variance estimate is derived from the properties of the variance for the sum of random variables. From (2) the difference between treatment and control for the entire sample is the weighted sum of the differences for the response and non-response groups. The resulting variance estimate is

$$\begin{aligned} S_p^2 &= w^2 S_{pR}^2 \sqrt{\frac{1}{n_{TR}-1} + \frac{1}{n_{CR}-1}} + (1-w)^2 S_{pNR}^2 \sqrt{\frac{1}{n_{TNR}-1} + \frac{1}{n_{CNR}-1}} \\ &= w^2 S_{pR}^{2*} + (1-w)^2 S_{pNR}^{2*} \end{aligned} \quad (4)$$

and the final form of the test statistic is

$$t = \frac{w\Delta\bar{Y}_R + (1-w)\Delta\bar{Y}_{NR}}{\sqrt{w^2 S_{pR}^{2*} + (1-w)^2 S_{pNR}^{2*}}} \quad (5)$$

where the degrees of freedom for the Welch's t -test (Welch, 1947) are determined using the Satterwaite approximation (Satterwaite, 1946). The experimenter can make reasonable estimates of the non-response difference in means and pooled variance by considering their likely values in comparison to the response group. The experimenter can then calculate the resultant t statistics under a variety of scenarios. Although the truth is not known, expert opinion and experience can inform the experimenters as to the likelihood of these various scenarios and provide some insight into the robustness of the experimental results.

The Problem of Differential Item Response Rate Bias

Violating the assumption used in (3) that the rate of response is approximately equal for the treatment and control groups can be misleading. In that case the exact formula given in (2) can be extended to evaluate the equivalent t -test, though the computations can be cumbersome and the assumptions around the correlation between strata become tenuous, further compounding the confusion. In this case the literature provides several proposals (often elaborate) to correct this selection bias. Winship and Mare (1992) summarized their review of the various methods proposed, with a warning that different methods yield different results and that there is not enough known about robustness of the correction methods proposed. This is especially acute for the Heckman estimator (Heckman, 1979), used largely in an econometric regression context for panel data. While experimental criminology as a discipline can profit from the Heckman (1979) and Winship and Mare (1992) type of approach, the utility of these in experimental criminology may be sometimes questionable, particularly in studies where the unit of analysis is a place or area, because RCT sample sizes in some situations can be rather small, often having only several dozen cases (e.g., see Weisburd & Gill, 2013). In this instance formal modeling is particularly useful if we have sufficient statistical power to estimate the characteristics of the modeled response process. For smaller sample sizes, however, we often do not have the statistical power to estimate the non-response process parameters needed.

In the QCET, perceptions of police were predicted to be significantly improved in the experimental condition than in the control condition as a result of officers using a procedurally just approach. Drivers who have very negative opinions of police may reasonably be less likely to choose not to complete a questionnaire about their perceptions of the police than drivers with positive opinions. In both cases, the difference between the observed drivers in the two samples (experimental/control condition, randomly assigned) is not an unbiased estimate of whether the experimental condition has a greater impact on increasing perceptions of police in the population. There may be fewer drivers with negative views of police in the control condition, but we may not see evidence of that because negative drivers would be less likely to respond than positive drivers. Hence due to this non-

response, the observed fraction of negative drivers will be artificially low in the control condition. This would also be the case in the experimental condition, but if there are fewer negative drivers in this condition the bias will be less. Thus the net result may underestimate the difference between both treatments' effects.

We argue that in experiments like QCET, it is often helpful to analyze the possible influence of differential item response in the case of dichotomous responses. For that purpose, a three step procedure, tailored to the problem at hand, is proposed: (a) specify a (parameterized) very simple model for differential item response rate, usually having one or a few unknown response likelihood parameters in it, (b) calculate what bias is generated by this non-response process, and derive a formula for bias size in terms of the unknown response parameter, and (c) study whether response bias is problematic for reasonable assumptions about the size of the response parameters. This approach has been proposed earlier by Elffers (1982, 2001) in the context of tax fraud and voting polls. The method should indeed be tailored each time, as, of course, the tenability of a simple differential response rate model (ad a.) is critically dependent on the problem at hand, just as the decision (ad c.) whether the resulting bias (ad b.) is problematic or not is essentially context dependent. Therefore, we will illustrate this three-step method in a reanalysis of the QCET RCT that tested the impact of procedural justice (experimental condition) on citizen perceptions of police legitimacy in Queensland (Mazerolle et al., 2012).

Assessing General Non-Response Bias in QCET (Cochrane Method)

Turning now to the QCET data, we can implement (2) to construct a set of operating characteristic curves showing the results for the resulting *t*-test under various scenarios for the non-response results. The response rates for the experimental and control groups are 12.12% and 13.74%; these are reasonably close enough to allow the use of the assumption that they are approximately equal and we can use the overall response rate of 13.09% in our calculations. From (5) it is evident that the true value of the test statistic *t* depends on both ΔY_{NR} and S_{pNR}^{2*} .

In cases where there is significant non-response within a study, experimenters can make reasonable estimates of the difference in means and pooled variance for non-responders based on their experience (e.g., similar studies with similar effects) and responses for those participants who did respond (actual responses). For example, if we consider the pooled variance of non-responders as being either equal to responders' pooled variance, or multiplied by some factor, we can then calculate the resulting threshold values for statistical significance for the difference between treatment and control for non-responders (based on a test statistic value of 2 – the critical *t* value when degrees of freedom are greater than 50). Below this threshold, the overall results combining the responders' and non-responders' results would no longer be statistically significant (i.e., the experimental and control groups

would not be significantly different). This is useful as a mental exercise for the researcher and can be used *a priori* to evaluate possible outcomes for experiments as a part of the design process.

Figure 1 shows the results for the non-response bias model for the construct of perceptions of procedural justice within the QCET, considering various values for non-responders' difference between the experimental and control groups, and the resultant pooled variance. These results show that if the pooled variance for the non-responders is equal to that for responders, the difference between the control and experimental groups would need to be less than -0.013 – that is, the results would have to indicate a negative effect for the treatment – in order to make the overall results no longer statistically valid. In the case where the non-responders' pooled variance was 4 times that of responders', the threshold is 0.007; if the pooled variance were 9 times that of responders', the threshold would be 0.027. Given the extreme case where pooled variance of non-responders were 9 times that of responders (which is an almost unthinkable large variance, and unlikely in and of itself), the observed difference would have to be almost a tenth that of the difference between the experimental and control responders' results in order to compromise the statistical validity of the results.

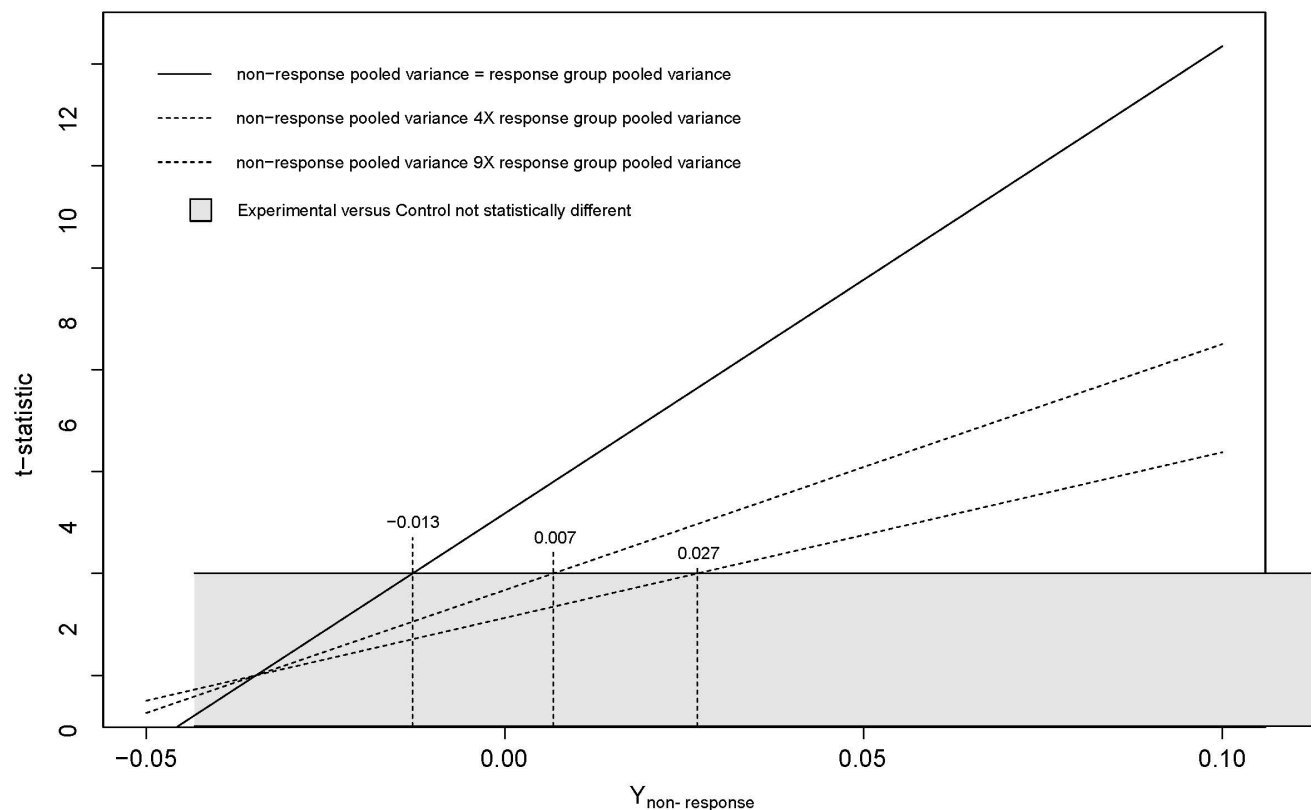


Figure 1. The Effects of $Y_{\text{non-response}}$ on test statistics

Assessing Differential Item Response Rates in QCET (Eiffers Method)

In this section we apply the three step tailored method for investigating item response bias in the dichotomous case. First we have to propose a non-response process model (*step a*).

The study addresses drivers subjected to an RBT. Let ξ be the probability that somebody would complete a questionnaire given that they have a positive attitude toward the police and they were in the treatment group. Moreover, it seems reasonable to suggest that the actual treatment will influence response rates: being treated with the procedurally just script (experimental condition) might increase the feeling with the respondent that because the police treated them fairly and spent time explaining their actions, they feel obliged to reward that by completing a questionnaire, whereas in the “normal mode” (control condition) people do not have such an incentive to complete the questionnaire. However, it may be the opposite: as the experimental condition takes longer, people may become irritated by the time lost and hence feel that the police have already taken too much of their time, so a lower response rate could be expected. It seems reasonable therefore to account for a differential effect on response rates: let us say that the ratio of response rates between a person being treated by the control condition and by the experimental condition is θ , an unknown parameter. There is no differential effect of treatment if $\theta = 1$.

Let us now look at the dependent variable. The QCET study examined a number of dependent variables, but for this example we consider the perceived use of procedural justice of police during the encounter as the dependent variable of interest. It seems reasonable to suppose that the actual perceived procedural justness of the police will influence response likelihood. People who judge the police to be procedurally just, will, presumably, tend to complete a questionnaire more often than people who have a negative perception of police officers’ treatment of them. Let us denote the ratio between the probability of completing a questionnaire in the negative and positive group as ρ , the differential (for dependent variable) non-response factor or distortion factor. There is no distortion effect on response if $\rho = 1$, as response rate is not affected by people’s attitudes (only treatment). Notice that it would be possible to suppose that there is also an interaction effect between the treatment and the perception of procedural justice, in which case we would need separate ρ parameters for both conditions, but we refrain from that here. To summarize, we model response probability with three parameters, ξ , θ , and ρ : ξ is the probability of responding for people with a positive attitude towards the police in the experimental condition, θ , and ρ being multiplicative factors that account for responsive tendency differences, ρ picking up the difference between the experimental and control group, θ catering for the difference between people with a negative and positive attitude. For completeness we also sum up the other symbols used in this model: n_1 is the number of respondents approached in the experimental condition, of which then m_1 do respond. π_1 is the fraction people that, when getting the experimental treatment, would end up with a positive attitude (i.e. the parameter of interest), p_1 is the fraction in the experimental sample reporting a positive attitude. The same parameters with subscript 2 denote the parallel entities in the control group. Figure 2 gives a schematic representation of the response process.

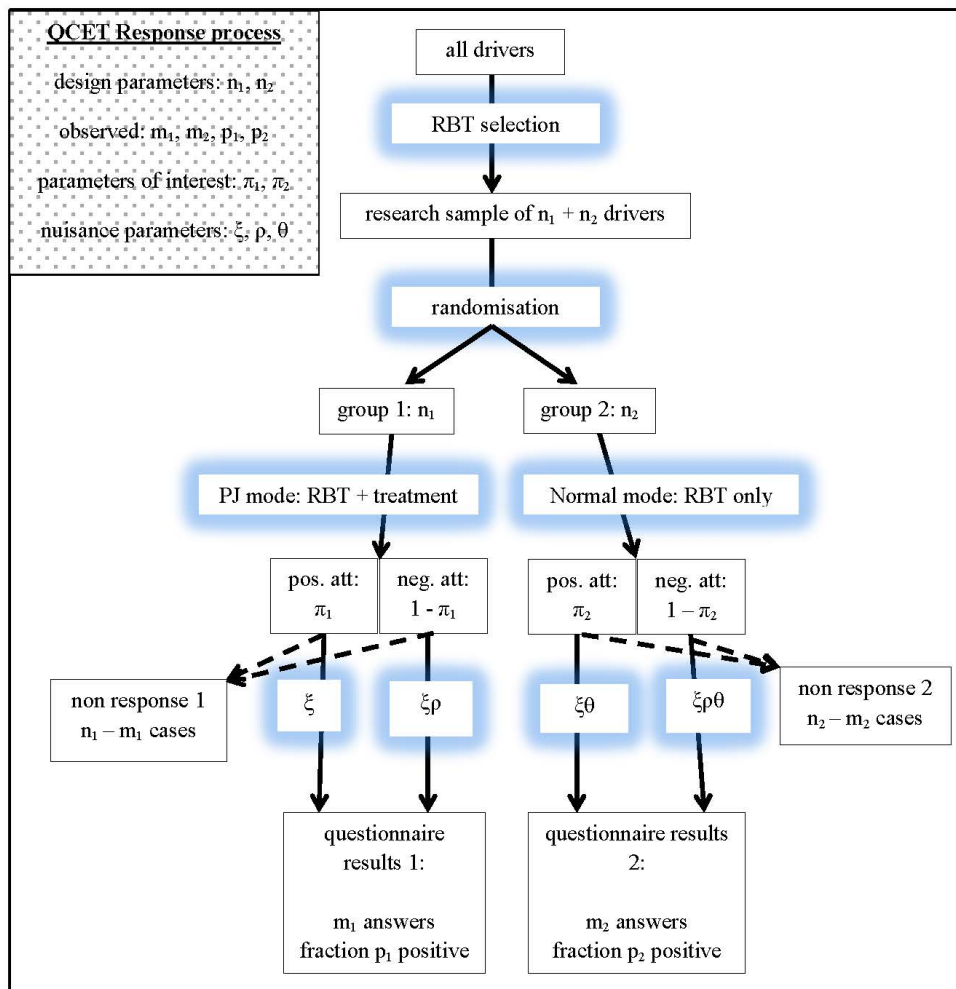


Figure 2. The QCET response process

Notice that the θ and ξ parameters are not generating response bias. The ξ – parameter tells us that we must question a lot of people to get enough data, while $\theta \neq 1$ additionally tells us that we have to work even harder in one condition to get enough responses than in the other

condition. However, sample estimators in both conditions produce unbiased estimates for the respective population characteristics in the two mode strata. It is only the ρ factor that generates bias. Nevertheless, we specify the full model, including the other non-response parameters, to make clear what exactly is meant by the distortion factor. In the next step (*step b*), we compute the effect of different values of ρ on the estimates.

Let e_i be the expected fraction of people with a positive perception of procedural justice of the police in each condition i ($i = 1, 2$). It holds that:

$$\begin{aligned} e_i &= \pi_i \xi n_i / \{ \pi_i \xi n_i + (1-\pi_i) \rho \xi n_i \} \\ &= \pi_i / \{ \pi_i + (1-\pi_i) \rho \} \end{aligned} \quad [1]$$

Which is, as predicted, not dependent on θ and ξ .

Solving this for π_i gives

$$\pi_i = \rho / \{ 1/e_i - (1-\rho) \} \quad [2]$$

Substituting the observed fraction of positive answers p_i for e_i , we can derive a relation between the difference between the population strata fractions $\pi_1 - \pi_2$ and the observed p_1 and the unknown differential response distortion parameter ρ

$$\pi_1 - \pi_2 = \rho * \{ 1/(1/p_1 - (1-\rho)) - 1/(1/p_2 - (1-\rho)) \} \quad [3]$$

We can use this formula to estimate to what extent the actually observed difference between the positive fractions $p_1 - p_2$ should be corrected on behalf of distortion by ρ , for various values of ρ .

In the concrete example of the QCET, we have observed fractions $p_1 = 0.85$ and $p_2 = 0.73$, in which case [3] comes down to what is depicted in Figure 3 (see below), for values of ρ between $\frac{1}{2}$ and 2.

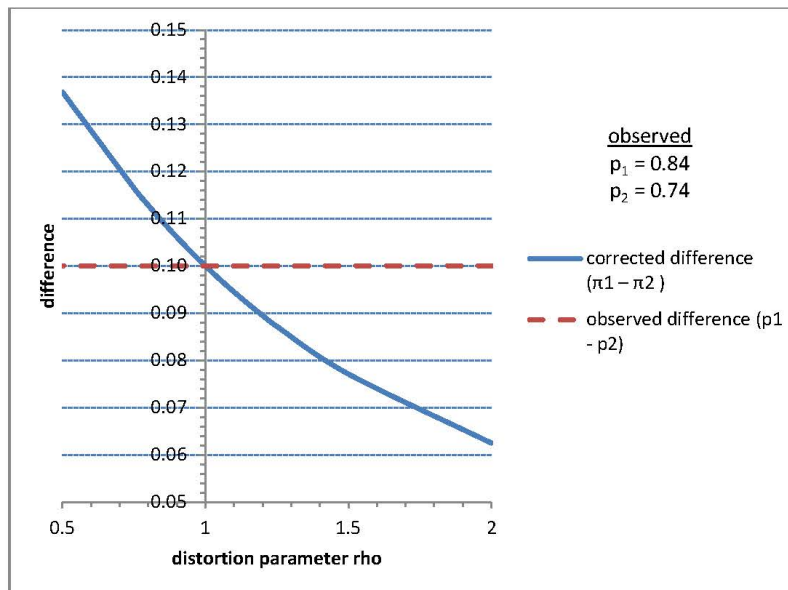


Figure 3. Correction on estimated difference between fractions of people perceiving police as legitimate for differential non response

The third step in the procedure (*step c*) is now to make up our minds about whether or not the size of the corrections as a function of ρ is so large that they would influence what conclusions should be drawn from the analysis. Ultimately, this is dependent on what size of ρ we are ready to accept as possible in the case at hand. In the above sample, we see that even rather serious distortions (non-response being twice as likely for people having a positive perception of police as for those having a negative perception, or vice versa) have only a moderate influence on $\pi_1 - \pi_2$: the observed fraction difference is now 0.12, and the correction changes it to 0.07 when $\rho = 2$, hence still endorsing the conclusion that the experimental condition results in a higher perception of procedural justice. In accordance with previous applications of the method (Elffers, 1982, 2001), we suggest that a distortion larger than a factor 2 is rather unlikely, and hence we conclude that the QCET results as published earlier (see Mazerolle et al, 2012, 2013) are robust against moderate distortion by non-response.

Discussion and Conclusions

Non-response bias can be a serious problem in many field randomized controlled trials that use surveys to assess differences between experimental and control group participants. Low response rates, such as those found in the QCET, should be of concern to researchers in criminology, particularly where experimental methods are used. Low response rates and the resulting increased potential for differential attrition across conditions in randomised field experiments can reduce the robustness of the findings of a typically very robust methodology. From the QCET, it becomes clear that mail surveys, while useful and frequently used in other methods within criminology and other social sciences, pose a particular problem for

randomised experiments as they have the potential to insert unnecessary bias into an otherwise highly rigorous design. Without the ability to follow up non-responding participants, there is greater potential for certain types of people to choose not to respond to a mailed survey. Finally, differential attrition is a serious issue, and should be given more attention in relation particularly to randomized experiments. Not paying sufficient attention to these issues can have serious implications for policy and practice. If we cannot be certain that our results are not biased, we have no place making recommendations based on such results.

In this paper, we have presented two ways for examining non-response bias in order to assess the degree of distortion that may be present in results published about QCET (see Mazerolle et al, 2012; 2013). In the QCET, the overall response rate was 13.09%. The central goal of this paper was to scrutinize one of the key QCET outcome measures (procedural justice) to assess whether or not the published results may be biased in some way due to the low response rate. We subjected the QCET data to testing response bias in two different ways: first, we used the Cochran Method to scrutinize the QCET data and assess the general impact of low response rates on the statistical significance of results. Second, we followed the Elffers approach for examining whether or not differences in the control and experimental responses influenced the way surveyed citizens responded positively or negatively (a dichotomous measure) in the questionnaire to the items that comprised a measure of perceptions of procedural justice. Overall, our results show that the size of the QCET distortion was moderate for both the test of bias using the Cochran method, and the Elffers dichotomous analyses. This is of course a satisfactory result for the original QCET authors (see Mazerolle et al., 2012, 2013), as well as a sign that non-response bias indeed may be a problem, but certainly is not always a grave, fatal flaw.

References

- Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, *61*(8), 1139-1160. doi: 10.1177/0018726708094863
- Cochran, W. G. (1963). *Sampling Techniques* (2nd Ed.). Wiley.
- Cook, C., Heath, F., & Thompson, R. L. (2000). A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys. *Educational and Psychological Measurement*, *60*(6), 821–836. doi: 10.1177/00131640021970934
- Elffers, H. (1982) Non-randomness of samples and bold assumptions, *Proceedings 2nd Dutch Sociometric Society Congress*, Amsterdam, Vrije Universiteit, 3-11.
- Elffers, H. (2001) Selectieve respons is lang niet altijd desastreus [Selective nonresponse is not at all always detrimental]. *Kwantitatieve methoden*, **22**, (nr. 68), p. 49-55.
- Fox, R. J., Crask, M. R., & Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, *52*(4), 467-491. doi: 10.1086/269125
- Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*:153–161.
- Kane, R. J. (2005). Compromised Police Legitimacy as a Predictor of Violent Crime in Structurally Disadvantaged Communities. *Criminology*, *43*(2):469–498.
- Lum, C., Koper, C., & Telep, C. W. (2011). The Evidence-Based Policing Matrix. *Journal of Experimental Criminology*, *7*(1), 3–26.
- Mazerolle, L., Antrobus, E., Bennett, S., & Tyler, T. (2013). Shaping citizen perceptions of police legitimacy: A randomized field trial of procedural justice. *Criminology*, *51*(1), 33–64. doi: 10.1111/j.1745-9125.2012.00289.x
- Mazerolle, L., Bennett, S., Antrobus, E., & Eggins, E. (2012). Procedural justice, routine encounters and citizen perceptions of police: Main findings from the Queensland Community Engagement Trial (QCET). *Journal of Experimental Criminology*, *8*, 343–367. doi: 10.1007/s11292-012-9160-1
- Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *American Economic Review*, *80*:319–323.
- Satterwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *(2)*:110–114.
- Tyler, T. R. (1990). *Why People Obey the Law*. Yale University Press.

Tyler, T. R. (2003). Procedural justice, legitimacy, and the effective rule of law. In Tonry, M. H., editor, *Crime and Justice: A Review of Research*, volume 30. University of Chicago Press, Chicago, IL.

Tyler, T. R. and Huo, Y. J. (2002). *Trust in the Law: Encouraging Public Cooperation with the Police and the Courts*. Russell Sage Foundation, New York, NY.

Welch, B. L. (1947). The generalization of "Student's" problem where several different population variances are involved. *Biometrika*, (34):28–35.

Winship, C. and Mare, R. D. (1992). Models for Sample Selection Bias. *Annual Review of Sociology*, 18:327– 350.

Wolfe, S. E. (2011). The effect of low self-control on perceived police legitimacy. *Journal of Criminal Justice*, 39(1):67 – 74.