Prying Open the Black Box of Causality: A Causal Mediation Analysis Test of Procedural Justice Policing

Krisztián Pósch

LSE Law, Society and Economy Working Papers 23/2017
London School of Economics and Political Science
Law Department

Prying Open the Black Box of Causality: A Causal Mediation Analysis Test of Procedural Justice Policing

Krisztián Pósch*

Abstract: This paper reviews causal mediation analysis as a method for estimating and assessing direct and indirect effects in experimental criminology and testing procedural justice theory by examining the extent to which procedural justice mediates the impact of contact with the police on various outcomes. Causal mediation analysis permits one to better interpret data from a field experiment that has suffered from a particular type of implementation failure. Data from a block-randomised controlled trial of procedural justice policing (the Scottish Community Engagement Trial) was analysed. All constructs were measured using surveys distributed during roadside police checks. The treatment implementation was assessed by analysing the treatment effect consistency and heterogeneity. Causal mediation analysis and sensitivity analysis were used to assess the mediating role of procedural justice. The results suggest that the treatment effect was consistent and fairly homogeneous, indicating that the systematic variation in the study is attributable to the design. Moreover, procedural justice acts as a mediator channelling the treatment's effect towards normative alignment (NIE=-0.207), duty to obey (NIE=-0.153), sense of power (NIE=-0.078), and social identity (NIE=-0.052), all of which are moderately robust to unmeasured confounding. The NIEs for risk of sanction and personal morality were highly sensitive, while for coerced obligation and sense of power they were non-significant. This paper shows that causal mediation analysis is a versatile tool that can salvage experiments with systematic yet ambiguous treatment effects by allowing researchers to "pry open" the black box of causality. Most of the theoretical propositions of procedural justice policing were supported. Future studies are needed with more discernible causal mediation effects.

* PhD Candidate in the LSE Department of Methodology. I would like to thank Jonathan Jackson for many insightful comments and suggestions for an earlier version of this paper. I would like to also thank Sarah MacQueen and Ben Bradford for providing the dataset for the analysis.

I. INTRODUCTION

A recurring feature of social scientific research is that the majority of tests of cause-and-effect relations estimate whether a treatment affects an outcome – that is to say, they address the first order question – but they leave unexplored any underlying processes that may transmit the putative effect. Impact evaluations in criminology tend to focus on whether a desired outcome was achieved, not on how that outcome was produced (Famega et al. 2017), thus leaving incomplete our understanding of the mechanisms through which effects take place. This exclusive focus on whether a treatment affects certain outcomes can lead to research staying uninformed about crucial intermediate variables and can overlook the mechanisms which influence the scrutinised outcome. For instance, a number of randomised controlled trials (RCTs) have tested the efficacy of hot-spots policing, but the lack of assessment of how an antecedent variable (X) transmits its effect (at least partially) through an intervening (mediator) variable (M) to the outcome (Y) means that we do not know how and why hot-spots policing works. This failure to assess causal mediation limits the power and purchase of explanatory frameworks (Bullock et al. 2010; Imai et al. 2011). Without greater focus on how certain outcomes are produced, the chance of Type III and Type IV errors increase and researchers can erroneously conclude that the observed (null-) findings emerge due to the success or failure of the tested theory (Hassell and Lovell 2014).

This paper offers causal mediation analysis as a tool to address this "black-box" view of implementation and causality (Fagan 2017). The contribution of this article is threefold. First, it addresses the strong assumptions and limitations of the traditional approach to mediation analysis (the product method, see Baron and Kenny 1986). This traditional approach has been widely used in observational research, especially in the literature of structural equation modelling, where direct and indirect effects are routinely estimated (Mackinnon 2008; Mackinnon et al. 2013). However, most users of this method are unaware of the strong and often unattainable underlying assumptions for estimating indirect effects, which if not met might lead to unreliable and unsound estimates.

Second, it introduces causal mediation analysis, a technique developed by Imai and colleagues (Imai et al. 2010a 2010b; Imai et al., 2011) that seeks to overcome the limitations mentioned above to produce potentially causally interpretable results. Also presented are sensitivity analyses techniques that can be used to assess the robustness of results to unmeasured confounding. Unlike in previous criminological work, where causal mediation analysis has been used in a longitudinal research context (Walters 2015 2016), it is here employed in an experimental setting; moreover, this paper also goes beyond a recent review of applied literature on causal mediation in criminology (Walters and Mandracchia 2017) by (a) presenting a versatile statistical technique and (b) utilising the potential outcome framework to outline fundamental causal assumptions and describe new definitions of direct and

indirect effects. Moreover, it recommends two sensitivity analysis methods that can be easily used in most applied settings.

Third, it provides a demonstration of how to test causal mediation effects by drawing on data from the Scottish Community Engagement Trial (ScotCET) (MacQueen and Bradford 2015). ScotCET was designed to estimate the effect of procedurally just policing on people's experience of procedural justice. However, this RCT produced findings contrary to expectations, where those who received the designed procedurally just treatment reported lower levels of perceived fairness of the police. In such instances, qualitative process evaluations can address what went wrong during implementation (Haberman 2016; MacQueen and Bradford 2017), but these endeavours are retroactive, only focusing on startling cases, and thus they may suffer from verification bias. Moreover, such problematic datasets with unusual results are often discarded without proper statistical tests having been carried out on treatments' effects. This paper shows how to test whether value can be extracted by focusing on treatment constancy and treatment effect heterogeneity, that is, by assessing whether the systematic variation in the dataset is attributable to the research design. To foreshadow the results, an assessment of the treatment constancy and treatment effect heterogeneity shows that the unintended negative treatment effect in ScotCET was produced by the treatment assignment.

Finally, using ScotCET, this article adopts causal mediation analysis to test a fundamental assumption of the theory of procedural justice policing: namely, that the perceived procedural justice of the police channels the impact of previous experiences with the police towards various desirable outcomes, such as police legitimacy, social identity, or sense of power. By shifting the focus from the total effect of the treatment to the indirect (mediated) effects, experiments with systematic but ambiguous treatments can become interpretable, and the initial theory thus can become testable. The causal mediation analysis results here provide qualified support for the theory of procedural justice policing, showing that procedural justice mediates the impact of previous experiences with the police on police legitimacy, sense of power, and social identity with moderate levels of robustness to unmeasured confounding.

II. SCOTTISH COMMUNITY ENGAGEMENT TRIAL (SCOTCET)

The overarching goal of ScotCET was to test procedural justice theory in the context of roadside checks, where drivers were stopped by the police for vehicle safety checks and alcohol testing. ScotCET was fielded during the Festive Road Safety Campaign in the December of 2013 and January of 2014 in Scotland. The design block randomised ten matched pairs of police units to minimise bias across delivery units. After the roadside checks, more than 12,000 questionnaires were handed out to drivers from which 511 were returned after the start of treatment period (176 from the treatment and 335 from the control group). The study provided police officers in the treatment group with a series of talking points with

the aim of communicating procedurally just messages, while officers in the control group would carry on with their usual behaviour during these police encounters. As already noted, ScotCET produced the opposite effect to that intended: those who received the treatment reported lower levels of perceived procedural justice compared to the control group (MacQueen and Bradford 2015).

In a retroactive qualitative process evaluation, MacQueen and Bradford (2017) conducted nine group interviews with police officers who had taken part in the experiment. This follow-up revealed several issues that may have impacted negatively on the treatment implementation. ScotCET coincided with a period of heightened anxiety among officers due to a substantial and unpopular organisational reform in the Scottish police force. Moreover, the participating officers had not been properly briefed regarding the purpose of the study. They had received opaque instructions, assumed that the experiment would have a negative impact on their interactions with members of the public, and felt that the prompts and questionnaire had been assembled by out-of-touch researchers. The focus groups revealed unanimous signs of discontent and negativity towards the experiment. It is conceivable that this had a diffuse negative effect on the officers' attitudes and behaviour during encounters in the treatment groups, which may explain (at least partially) the contradictory findings (MacQueen and Bradford 2017).

Despite these problems and apparent failure of implementation, the authors maintained that the treatment effect was still interpretable due to the robustness of the study design.[1] Yet police officers reportedly differed in how they had carried out the treatment – some recited the provided messages verbatim, some completely disregarded the prompts, and some only handed out the questionnaires (MacQueen and Bradford 2017). To test the consistency of the treatment effects for each outcome variable, one model was fitted for the whole dataset accounting for the block randomisation with clustered robust standard error, with the treatment and covariates[2] included as explanatory variables. In addition, separate models were fitted for each of the matched pairs. Due to an insufficient sample size for one pair, only nine matched pairs were included in the analysis. Figure 1 shows the results for procedural justice as the outcome variable, with by and large similar effect sizes across the different pairs of delivery units – the differing results for the fourth, seventh and eighth pair can be attributed to normal sampling variability. Similar results emerged for the other outcome variables of interest, which appears to confirm the consistency of the treatment's effect. Treatment-effect heterogeneity was also examined, which revealed minimal design heterogeneity after controlling for covariates and minor covariate heterogeneity, reinforcing the robustness of the

[1] "The data gathered successfully captured the outcome of the experimental intervention, and the robust design and internal validity of the ScotCET experiment assures us that the negative effects observed within our experiment group can be directly attributed to the intervention, or factors associated with the intervention." (MacQueen and Bradford, 2016, 4.pp.)

[2] As detailed in the Preliminary remarks, these variables were gender, age, housing status, employment, and whether a breath test was conducted.

results (further details and discussion regarding treatment consistency and treatment effect heterogeneity can be found in the Appendix).

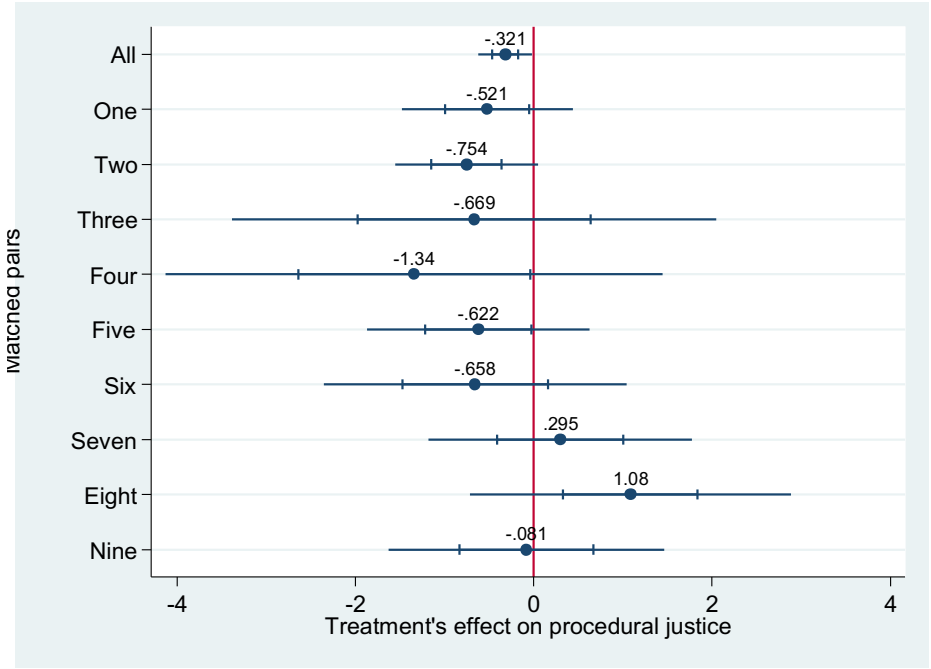

Figure 1 Treatment effect consistency for procedural justice

Although the checks of treatment consistency and heterogeneity provided strong evidence regarding the internal validity of the treatment's effect, it is still very difficult to give a proper definition for the treatment. Therefore, this article proposes a mere descriptive interpretation, assuming that the treatment induced systematic variation in the public's experiences with the police. For the treatment group, these experiences were more negative compared to the control group. Thus, this paper tests a fundamental question found in the procedural justice literature: whether the impact of a person's previous positive/negative experiences with the police is channelled through procedural justice to affect certain outcome variables (e.g., legitimacy).

III. PROCEDURAL JUSTICE POLICING

Procedural justice policing is an extensively discussed topic in criminology (Tyler et al. 2015). The perspective posits that, when thinking about the police, people tend to form their attitudes to a great degree on whether police officers seem to act in fair, neutral, and respectful ways. In Western countries, these considerations appear to be more influential than instrumental concerns, such as the effectiveness of the police or beliefs about the unequal distribution of policing outcomes. Attitudes towards the procedural fairness of the police are thought be influenced by legal

socialisation (e.g., Trinkner and Tyler 2016) and direct/vicarious contact with the police (e.g., Bradford et al. 2009; Tyler et al. 2014).

Importantly for this paper, procedural justice theory has a number of mediational layers, most notably a causal path linking (a) police behaviour in a police-citizen encounter, (b) the subjective judgement of the citizen regarding the procedural fairness of the officer, and (c) citizen judgements regarding police legitimacy. ScotCET was designed as a partial replication of the Queensland Community Engagement Trial (QCET), which found that when officers followed a "procedurally fair" script, citizens tended to view their experience as more procedurally just, and that this experience of procedural justice in turn predicted police legitimacy (Mazerolle et al. 2013). Of note is that Mazerolle and colleagues used a traditional path analysis, concluding that the experience of procedural justice mediated some of the treatment effect of police behaviour on legitimacy.

Another prominent claim regarding procedural justice is that it is one of the key (if not the key) bellwethers of how people think about other desirable outcomes regarding the police, and of these outcomes, perhaps the legitimacy of the police stands out the most. Following Hough, Jackson and Bradford (2013, also Huq et al. 2017), in this paper it will be assumed that legitimacy of the police consists of two parts. Firstly, moral alignment with the police taps into the idea that the police respect key societal values regarding how authority should be exercised, where these values' congruence is shown by officers acting in normatively appropriate ways. Secondly, duty to obey encapsulates people's willing consent to follow police orders. Bottoms and Tankebe (2012) have voiced some concerns regarding the validity of this conceptualisation, arguing that for some people this duty to obey might stem from coercion by and fear of the police, instead of normative considerations. Thus, as a theoretical innovation, questions were included for this coerced/prudential aspect of obligation to obey. This coerced obligation to obey is assumed to be the "yin" to free duty to obey's "yang"; it aims to measure forced aspects of obligation to obey the police. It is expected that, similar to legitimacy, procedural justice will influence this aspect of obedience, but in the opposite direction.

Apart from legitimacy of the police, procedural justice of the police is also presumed to have an impact on how people rate their own social standing. Police officers are representatives not only of the state, but the communities they serve (Bradford 2014), and if the police treat someone fairly, with respect, and provide citizens with a voice, those citizens will feel empowered (a strengthened sense of power) and that they belong in that particular community (emboldened social identity) (Mentovich 2012). In contrast, it is conceivable that these procedurally just signals do not affect how people perceive the police's capabilities and power over them (unchanged power distance between them and the police).

Finally, personal morality and perceived risk of sanction have also been included in the analyses here as possible outcomes (Jackson et al. 2012). I speculate that procedural justice will not mediate the treatment's impact on either of these,

hence, similar to power distance, citizens may only be influenced directly by their previous experiences with the police.

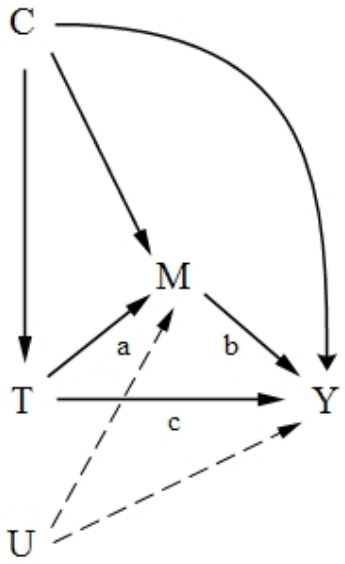

Figure 2 Outline of a mediation model with a single mediator

IV. CAUSAL MEDIATION ANALYSIS: CLASSICAL DEFINITIONS OF DIRECT AND INDIRECT EFFECTS

In this article, I hypothesise that perceived treatment by the police (X) is expected to shape respondents' attitudes (regarding procedural justice) (M), which in turn is expected to influence – among other things – their views on the legitimacy of the police (Y). Because traditionally, X refers to any kind of (even observed) variable, this paper will denote the antecedent variable as T, which indicates the randomised treatment. In addition, it is conventional to control for a vector of pre-treatment covariates C (see Figure 2). Using the traditional decomposition of the product method, and as depicted by Figure 2, 'c' stands for the direct effect of T on Y, while the product of 'a' and 'b' (i.e., the estimates of T's effect on M, and M's effect on Y) stands for the indirect effect of T that goes through M towards Y. This approach is generally referred to as the product method as an indication of how the indirect effect is derived.

However, several criticisms have emerged regarding the applicability of the product method. First, the product method is only capable of identifying[3] direct and

[3] Identifiability here – and throughout the paper – means that an (causal mediation) effect is consistently estimable. It follows that identification is a necessary, but not sufficient requirement, which precedes the actual statistical estimation and refers to the ability to obtain the effects of interest (Manski 2007; Keele

indirect effects if the linearity assumption holds (Imai et al., 2010b; Jo 2008). This means that for non-linear (e.g., multinomial) models the indirect effect cannot be computed relying on the product method. The second caveat is usually referred to as no-interaction assumption or effect homogeneity. This prescribes that there cannot be an interaction between the treatment and the mediator which affects the outcome. The absence of interaction is important, because it permits the effect decomposition and also provides a good indication for effect homogeneity (i.e. the causal effects are constant across cases) (Kline 2015). In the presence of an interaction (e.g. between the treatment and procedural justice in this paper), the method of identification of the direct and indirect effects breaks down as it becomes unclear how to calculate the total effect. Yet, the lack of interaction is not sufficient, because effect homogeneity needs to apply to each individual case, which is an untestable (and highly unlikely) assumption.

A further limitation concerns the applied literature rather than the method itself. Similarly to other causal techniques, causal mediation analysis relies on no unmeasured confounder assumptions, which are usually addressed by the random assignment of participants to treatment and control group(s). In other words, if we randomly assign people to a treatment or control group, we can safely assume that they will not differ across important and influential measured *and* unmeasured characteristics (e.g. age, education, previous experience with the police), and hence the exogeneity assumption is met. However, even if the treatment T is randomly assigned, the mediator-outcome relationship is not randomised, which might result in people self-selecting for their mediators independent from the treatment and due to an unmeasured confounder U (depicted in Figure2). This U can generate biased direct and indirect effects, thus producing unreliable results. This issue has been mostly overlooked, partly because it was not discussed in the classic article by Baron and Kenny (1986); although it was discussed in an earlier paper by one of the authors (Judd and Kenny 1981).

To further complicate matters, randomisation of the mediator, as proposed by some (Bullock et al. 2010; Spencer et al. 2005; Walters and Mandracchia 2017), is also not sufficient for assessing the indirect effect. When both the mediator and treatment are randomly assigned, the exogeneity assumption is satisfied for each, however, it does not apply to the combination of the two. In such cases, the treatment can causally affect the mediator, and the mediator can causally affect the outcome, however, the mediator does not transmit the effect of the treatment anymore due to its random assignment (Imai et al. 2010a; Keele 2015). Thus, this seems to be a germane problem in the literature as special design-based strategies need to be applied to assure that the mediator-outcome relationship is indeed causal (Imai et al. 2013; Imai et al. 2011; Pirlott and Mackinnon 2015). A careful selection of pre-treatment covariates might mitigate the possibility of an unmeasured

2015). Importantly, this is different from the model-based identification regularly used in the structural equation literature.

influential U, but it can rarely solve the issue altogether (VanderWeele 2015). Consequently, alternative definitions of direct and indirect effects need to be developed and different analytical strategies pursued to address these issues.

V. COUNTERFACTUAL DEFINITIONS OF THE DIRECT AND INDIRECT EFFECTS

In the following paragraphs the controlled direct effect, natural direct effect, and natural indirect effect are discussed as alternatives to the direct and indirect effects from the product method. These new alternative definitions rely on the potential outcome framework and counterfactual way of thinking (Pearl 2001; Robins and Greenland 1992). The language of conditional expectations is employed, to indicate that population average effects are, in fact, conditional expectations of the individual level effects. These counterfactual definitions are given assuming a binary treatment variable mirroring the one used in ScotCET.

For all of these counterfactual definitions, let us assume that we compare two hypothetical worlds where in the first world T is set to 0 (i.e. control) and in the second T is set to 1 (i.e. treatment) within the same individual at the same moment in time. Using ScotCET as an example, this would mean that the same person would have been exposed to both the trained messages and the usual police practice during the roadside check at the very same moment in time from the very same officer(s). Although in real life we can never know what would have happened to that individual had that person been assigned to the other group[4] instead of the observed one, hypothetically we can conceive these two separate counterfactual outcomes. It follows that counterfactual inference can never be derived for a single individual, only for a population.

The controlled direct effect (CDE) considers a specified value of M=m and captures the expected increase in Y when T changes from T=0 to T=1 (i.e. within the individual M is kept constant, while she receives both the control and treatment at the same time). This is a direct effect since the effect of T is not transmitted through M. The value of CDE might change depending on the chosen value of m, which also means that relying on CDE does not allow the decomposition of the total effect to direct and indirect effects. Still, setting the m to different values can provide policy relevant information, such as the number of meetings people on parole should attend in order to reduce their recidivism.

(1) $CDE(m)=E[Y(1,m)-Y(0,m)]$

The natural direct effect (NDE) is similar to the controlled direct effect, as it estimates the expected increase in Y when T changes from T=0 to T=1. However,

[4] This limitation is often referred to as the fundamental problem of causal inference (Holland 1986).

the NDE does not hold m constant, instead it permits m to take its value in the "natural" way for each individual if that individual had been assigned to the control condition. This modification allows for the decomposition of the effects. Provided that there is no T-M interaction, the CDE(m) and NDE will coincide when the CDE is controlled on the average value of M=m.

(1) $NDE=E[Y(1,M(0))-Y(0,M(0))]$

The natural indirect effect (NIE) does the opposite of NDE as it approximates the expected increase in Y when the treatment is kept at T=1, while M is freed to take its natural value of m for the treatment and the control group respectively. This is an indirect effect that captures the effect of T on Y which is transmitted through M:

(2) $NIE=E[Y(1,M(1))-Y(1,M(0))]$

Importantly, both the direct and indirect effect can be defined through holding M at T=1 for the direct effect, while holding Y at T=0 for the indirect effect, which will produce identical results in respect of the total effect:

(3) $NDE_{alt}=E[Y(1,M(1))-Y(0,M(1))]$

(4) $NIE_{alt}=E[Y(0,M(1))-Y(0,M(0))]$

Crucially, these alternative decompositions will determine to where the effect of the potential T-M interaction term is assigned (Daniel et al. 2015; Muthen and Asparouhov 2015). Using the classic definition of NIE and NDE (Pearl 2001), the interaction term is assigned to the indirect effect, while for the NDE_{alt} and NIE_{alt} it is assigned to the direct effect. To avoid confusion, sometimes the words "total" and "pure" are added to the direct and indirect effects, where total indicates the added interaction effect. Therefore, the NIE is the total indirect effect (TNIE), while the NDE is the pure direct effect (PNDE). Conversely, the alternative definitions of NIE_{alt} and NDE_{alt} refer to the pure indirect (PNIE) and total direct effects (TNDE) respectively. Again, the sum of these effects is equal, adding up to the total effect TE (TE=TNIE+PNDE=PNIE+TNDE). As shown, the total effect (TE) can be decomposed as the sum of the NDE and NIE:

(6) $TE=E[Y(1)-Y(0)]=$
$E[Y(1,M(1))-Y(0,M(0))]=$
$\{E[Y(1,M(1))-Y(1,M(0))]\}+$

$$\{E[Y(1,M(0))-Y(0,M(0))]\}=$$
$$NIE+NDE=NIE_{alt}+NDE_{alt}=$$
$$TNIE+PNDE=PNIE+TNDE$$

As described above, the identification of the direct and indirect effects through the potential outcome framework does not posit the no-interaction assumption, which allows for the effect decomposition even in the presence of such an association. Moreover, it is nonparametrically identifiable, thus does not require the linearity assumption either, which permits more flexible modelling (Pearl 2001). For details regarding the estimation of the NDE and NIE please refer to the Appendix.

Assumptions of causal mediation analysis

In order to make causal claims based on the decomposition outlined above the sequential ignorability assumption needs to be satisfied (Imai et al. 2010a). This no unmeasured confounder assumption lists the different sources of U that can produce biased results and requires that, after controlling for all pre-treatment covariates C, there is no unmeasured confounder for:

a) The relationship between the treatment (T) and outcome (Y)
b) The relationship between the mediator (M) and outcome (Y)
c) The relationship between the treatment (T) and mediator (M) and,
d) There is no post-treatment mediator-outcome confounder (L) that was affected by the treatment

From these four assumptions, (a) and (c) constitute exogeneity assumptions usually applied to determine the average treatment effect in randomised experiments and are automatically satisfied in the case of random assignment of T. For (b) to be fulfilled M either needs to be as-if randomly assigned (using a special design) or assumed that it is as-if randomly assigned after controlling for T and C. To accomplish the final point (d), one needs to rely on a parsimonious model similar to Figure 2, as it posits that there cannot be other post-treatment confounders (essentially other mediators) that are not included in the model. In terms of the new definitions of the different direct and indirect effects assumptions, (a) and (b) are sufficient to derive the CDE(m)[5], while (a)-(d) are needed for the NDE and NIE. Finally, as with randomised experiments in general, the stable treatment unit value assumption also needs to be met.

[5] Notably, the usual regression-based models will no longer be sufficient, other approaches, such as marginal structural models, structural nested models and so on, can be used to derive the CDE (Coffman and Zhong 2012; Lepage et al. 2016; Moerkerke et al. 2015).

Sensitivity analysis for a single mediator

Similarly to other techniques in the causal inference literature, causal mediation analysis also relies on untestable and non-refutable assumptions (Manski 2007). Although, the strong claims of the sequential ignorability assumption cannot be directly tested on the observed data, sensitivity analyses can be utilised that permit researchers to quantify the robustness of their findings and assess the influence of unmeasured confounders. Critically, even if the treatment was randomised, the ignorability of mediator M should be studied through evaluating whether there is a reasonable chance that omitted variable U might invalidate the results. However, in most cases sensitivity analyses will not provide easily discernible results, rather a range of values that will indicate the plausibility of the results. As there are no established benchmarks upon which one could decide on the absolute robustness of results, inferences must be informed by previous findings from the field and should be compared with the impact of other measured confounders. There are several different sensitivity analysis techniques (Ding and Vanderweele 2016), here, two will be discussed; these techniques work especially well with continuous mediators and are capable of gauging the robustness of the NDE and NIE.

The first technique (Imai et al. 2010a; Imai et al. 2011; Imai and Yamamoto 2013) fits two regressions, one for M and the other for Y with a T-M interaction. One can take the error terms (ε) from these regressions and specify a correlation between them denoted by ϱ. Since the error terms incorporate the impact of U, the value of ϱ will relatively increase if there is an influential U that affects both M and Y. Conversely, ϱ will comparatively decrease in the absence of an influential U. Thus, the sensitivity of the mediation results can be tested by systematically increasing the correlation between the two εs and evaluating the extent to which the estimates are altered. Accordingly, the direct and indirect effects will be the functions of the parameter ϱ, and the higher value it takes will imply relatively more robust results. A mathematically equivalent, but perhaps more intuitive way of reporting the results, is to consider the R-squared statistics and interpret the results in terms of U's explanatory power. There are two R^2s worthy of interest. The R^2 for the residual variance shows the proportion of previously unexplained variance that is explained by U. Alternatively, the R^2 for the total variance represents the same, but for the proportion of the original variance. In the case of the R^2s, higher values will indicate relatively lower sensitivity to the violation of the sequential ignorability assumption compared to results from similar studies.

The other sensitivity analysis technique is called the left out variable error method (LOVE) (Cox et al. 2013; Mackinnon and Pirlott 2016), which assesses the extent to which an unmeasured variable U would have to affect the association between M and Y in order for the observed association to be attributable to this confounding alone. This approach classifies the error due to U as a misspecification error and applies correlation techniques for bias detection. Therefore, LOVE relies

on the correlation between T-M, T-Y, and M-Y to approximate the correlation between U-Y and U-M. The average of the U-Y and U-M correlation corresponds to a correlation coefficient that would make the observed mediated effect zero. As in the previous case, a higher coefficient will entail less sensitive results. The major advantage of this method is that it enables a less convoluted assessment of the effect of U on the M-Y relationship. However, this straightforwardness comes at price: unlike the previous sensitivity analysis, the LOVE technique does not include pre-treatment covariate Cs, which considerably limits its authenticity for the model under scrutiny. Nevertheless, the LOVE method can be still a powerful detector of bias and an easy check of the relationships between T, M, and Y.

VI. RESULTS: PRELIMINARY REMARKS

The items of procedural justice, moral alignment, duty to obey, coerced obligation, social identity, risk of sanction, and personal morality were all entered in a confirmatory factor analysis, and their confirmatory factor scores were derived and utilised in further analysis.[6] For power distance and sense of power, their shorthand single-item measures were used. In each causal mediation analysis model gender, age, housing status, employment, and whether a breath test had been conducted[7] were included as covariates. For the sake of brevity the list of measures is included in the Appendix (Table 3/a), further details can also be found in Macqueen and Bradford (2015). The table and discussion of the correlational results were also added to the Appendix.

Unlike in observational studies, where the goal of including control variables is to remove potentially spurious relationships between the independent and dependent variables, in experiments the inclusion of covariates aims to improve the efficiency of the analysis by eliminating nuisance variance (i.e. variation that is clearly not attributable to the treatment) (Coffman and Zhong 2012; Mutz and Pemantle 2016). Yet, and unlike with classical randomised experiments, for the M-Y relationship, causal mediation analysis requires the logic of controls to be applied, because the mediators are not randomised. Thus, it is reasonable that one would only select a limited set of covariates for the model for the mediator, and a broader set of pre-treatment control variables for the model for the outcome. In the analysis presented here, however, the same list of pre-treatment variables was kept for both models.

The potential outcome framework applied throughout this article encourages the use of falsification tests. Relying on DAGs (Directed Acyclic Graphs), one can depict the expected causal relationships, where not only the presence, but the absence of causal pathways (effects) is also relevant. The presence of an unexpected causal effect can be a sign of the influence of unmeasured confounders and the

[6] The results of the confirmatory factor analysis are available from the author upon request.
[7] The police are required to determine the need to conduct a breath test before a stop is initiated.

failure of the identification strategy pursued (Keele 2015). In accordance with this, Figure 3 portrays the expected relationships between the treatment, mediator, and the outcome variables.

In all models the treatment was binary, and the mediator and outcome variables were continuous. The "mediation" R package (Tingley et al. 2014) was used with interaction allowed between the treatment and the mediator, and 1000 bootstraps specified for more precise standard errors. The negative direction of the majority of the effect sizes reflects the unexpected negative effect of the treatment.

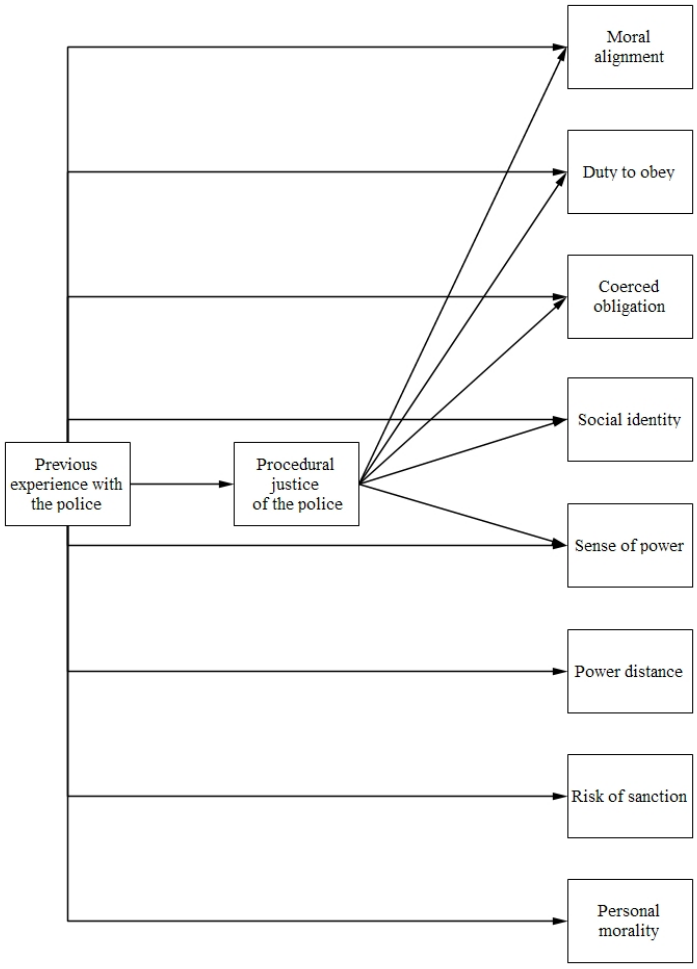

Figure 3 DAG based on the hypotheses of this article

Causal mediation analysis

The causal mediation analysis results are displayed in Table 1. I take the model fitted for moral alignment (first row) to exemplify the interpretation of the results. The average natural indirect effect (NIE) of procedural justice is -0.207, which is

significant on the 5% level. This NIE mediates 84.2% of the total effect with a non-significant natural direct effect of -0.007. To nullify the average NIE the mean correlational coefficient between the error terms from the model for the mediator and outcome would need to be 0.6. This (ϱ=0.6) corresponds to 36% of the residual variance and 20% of the total variance of the model. Thus, this relationship seems to be less sensitive or, in other words, fairly robust to unmeasured confounding. By contrast, for the average NDE's effect to reach zero, this correlation coefficient would only need to approach 0.1, with the power to explain 1% of the residual variation and less than 1% of the total variation. Therefore, this result is highly sensitive to unmeasured confounding, which corresponds to its NDE value that is close to zero and non-significant. Finally, the left-out-variable value (LOVE) implies that on average an unmeasured confounder would need to have a 0.7 correlation with the mediator and outcome to make the average NIE non-significant.

Procedural justice as mediator	*Type*	*Average effect*	*Mediate %*	*Mean ϱ*	*Residual R^2*	*Total R^2*	*Mean LOVE*
Moral alignment	NIE	-0.207* [-0.384, -0.031]	84.2%	0.6	0.36	0.20	0.7
	NDE	-0.007 [-0.261, 0.240]		~0.1	0.01	~0.01	
Duty to obey	NIE	-0.153* [-0.297, -0.018]	34.9%	0.5	0.25	0.17	0.7
	NDE	-0.279* [-0.540, -0.008]		0.7	0.49	0.32	
Coerced obligation	NIE	0.068 [-0.021, 0.159]	18.9%	0.3	0.09	0.07	0.5
	NDE	-0.115 [-0.373, 0.130]		0.5	0.25	0.12	
Social identity	NIE	-0.052* [-0.108, -0.005]	16.9%	0.3	0.09	0.12	0.5
	NDE	-0.243* [-0.411, -0.080]		0.8	0.64	0.46	
Sense of power	NIE	-0.078* [-0.150, -0.003]	61.5%	0.5	0.25	0.15	0.6
	NDE	-0.038 [-0.154, 0.086]		0.2	0.04	0.02	
Power distance	NIE	-0.001 [-0.024, 0.209]	~1%	~0.1	0.01	~0.01	~0.1
	NDE	-0.251* [-0.434, -0.075]		0.8	0.64	0.53	
Risk of sanction	NIE	-0.039* [-0.095, -0.002]	17.6%	0.2	0.04	0.03	0.3

	NDE	-0.102 [-0.354, 0.140]		0.3	0.09	0.07	
Personal morality	NIE	-0.061* [-0.131, -0.005]	10.9%	0.2	0.06	0.04	0.3
	NDE	-0.468** [-0.701, -0.241]		0.9	0.81	0.59	

Table 1 Causal mediation analysis results with averaged NDE and NIE effects and sensitivity analyses

**p<0.05, **p<0.01*

It is worth comparing the results (Table 1) to the DAG (Figure 3) presented earlier. Procedural justice does seem to channel the effect of the treatment to moral alignment (as discussed in the previous paragraph), duty to obey (NIE=-0.153, $p<0.05$, Mediate %=34.9%, $\varrho=0.5$, $R^2_{residual}=0.25$, $R^2_{total}=0.17$, LOVE=0.7), social identity (NIE=-0.052, $p<0.05$, Mediate %=16.9%, $\varrho=0.3$, $R^2_{residual}=0.09$, $R^2_{total}=0.12$, LOVE=0.5), and sense of power (NIE=-0.078, $p<0.05$, Mediate %=61.5%, $\varrho=0.5$, $R^2_{residual}=0.25$, $R^2_{total}=0.15$, LOVE=0.6). The only inconsistency with Figure3 appears to be coerced obligation with a non-significant NIE (NIE=0.068, $p>0.05$, Mediate %=18.9%, $\varrho=0.3$, $R^2_{residual}=0.09$, $R^2_{total}=0.07$, LOVE=0.5), which is surprising given the moderately strong correlation ($r=-0.411$, $p<0.01$) between procedural justice and coerced obligation (see: Appendix/Table2/a). To further investigate this puzzling lack of indirect effect the models were fitted without the covariates (see: Appendix/Table4/a) where coerced obligation turns significant. It seems that the mediated relationship disappears after taking into account potentially influential pre-treatment covariates.

Continuing with the falsification checks and the juxtaposition of Table1 and Figure3, procedural justice is not a significant mediator for power distance (NIE=-0.039, $p>0.05$, Mediate %=~1%, $\varrho=\sim0.1$, $R^2_{residual}=0.01$, $R^2_{total}=\sim0.01$, LOVE=~0.01). However, the average NIE is significant for both risk of sanction (NIE=-0.039, $p<0.05$, Mediate %=17.6%, $\varrho=0.2$, $R^2_{residual}=0.04$, $R^2_{total}=0.03$, LOVE=0.3) and personal morality (NIE=-0.061, $p<0.05$, Mediate %=10.9%, $\varrho=0.2$, $R^2_{residual}=0.06$, $R^2_{total}=0.04$, LOVE=0.3). Seemingly, procedural justice's indirect effects for risk of sanction and personal morality can be considered as further failures of the initially proposed falsification checks. Nevertheless, the sensitivity analyses seem to indicate otherwise: for both models a 0.2 correlation between the error terms would make the NIE non-significant, with a corresponding explanatory power of 3-4% of the total variance. The LOVE scores of 0.3 imply that an unmeasured confounder with moderately strong correlation with the mediator and the outcome would be sufficient enough to nullify the indirect effects. By comparison the next weakest significant NIE of social identity has a ϱ value of 0.3 with 12% of the total variation required to be explained and a LOVE value of

0.5. This illustrates one of the major advantages of causal mediation analysis: instead of merely relying on p-values, one can assess the robustness of results using sensitivity analysis. In the current case, for instance, risk of sanction's and personal morality's NIEs seem to be highly sensitive to unmeasured confounding, while social identity with a comparable effect size appears to be only moderately sensitive.

Procedural justice	*Type*	*Effect size*	*Mediate %*	*Mean ϱ*	*Residual R^2*	*Total R^2*
Moral alignment	NIE(0)	-0.171* [-0.321, -0.026]	69.5%	0.5	0.25	0.13
	NIE(1)	-0.244* [-0.449, -0.037]	98.9%	0.7	0.49	0.25
	NDE(0)	0.029 [-0.231, 0.284]		0.1	0.01	0.01
	NDE(1)	-0.044 [-0.299, 0.213]		0.2	0.04	0.02
Duty to obey	NIE(0)	-0.130* [-0.260, -0.014]	29.7%	0.4	0.16	0.11
	NIE(1)	-0.176* [-0.345, -0.020]	40.2%	0.5	0.25	0.16
	NDE(0)	-0.256 [-0.514, 0.009]		0.7	0.49	0.32
	NDE(1)	-0.302* [-0.558, -0.031]		0.7	0.49	0.32
Coerced obligation	NIE(0)	0.090 [-0.028, 0.209]	25.8%	0.4	0.16	0.08
	NIE(1)	0.046 [-0.014, 0.120]	12.1%	0.2	0.04	0.02
	NDE(0)	-0.094 [-0.343, 0.153]		0.4	0.16	0.08
	NDE(1)	-0.137* [-0.400, -0.116]		0.6	0.36	0.17
Social identity	NIE(0)	-0.029 [-0.074, 0.001]	9.2%	0.1	0.01	0.01
	NIE(1)	-0.075* [-0.156, -0.006]	24.7%	0.4	0.16	0.11
	NDE(0)	-0.219* [-0.387, -0.054]		0.8	0.64	0.46
	NDE(1)	-0.295** [-0.472, -0.124]		0.8	0.64	0.46

Sense of power	NIE(0)	-0.080* [-0.156, -0.003]	63.4%	0.5	0.25	0.15
	NIE(1)	-0.076* [-0.154, -0.003]	59.7%	0.5	0.25	0.15
	NDE(0)	-0.040 [-0.157, 0.085]		0.3	0.09	0.05
	NDE(1)	-0.036 [-0.154, 0.087]		0.2	0.04	0.02
Power distance	NIE(0)	0.021 [-0.006, 0.062]	7.1%	0.1	0.01	0.01
	NIE(1)	-0.023 [-0.068, -0.008]	3.7%	0.1	0.01	0.01
	NDE(0)	-0.229* [-0.415, -0.059]		0.8	0.64	0.53
	NDE(1)	-0.273** [-0.454, -0.091]		0.8	0.64	0.53
Risk of sanction	NIE(0)	-0.033 [-0.089, 0.009]	14.5%	0.1	0.01	0.01
	NIE(1)	-0.046 [-0.127, 0.001]	20.7%	0.2	0.04	0.03
	NDE(0)	-0.095 [-0.348, 0.143]		0.3	0.09	0.07
	NDE(1)	-0.108 [-0.360, 0.141]		0.4	0.16	0.13
Personal morality	NIE(0)	-0.028 [-0.082, 0.008]	4.8%	0.1	0.01	0.01
	NIE(1)	-0.094* [-0.197, -0.007]	17.1%	0.3	0.09	0.07
	NDE(0)	-0.435** [-0.667, -0.206]		0.9	0.81	0.59
	NDE(1)	-0.501** [-0.737, -0.270]		0.9	0.81	0.59

Table 2 Causal mediation analysis results with the interaction's effect attributed either to the NIE or NDE, and sensitivity analyses

**p<0.05, **p<0.01*

Finally, the inclusion of the interaction effect needs to be discussed. Another improvement of causal mediation analysis is that it manages to resolve the inclusion of the interaction effect while still guaranteeing a meaningful decomposition. In Table1 the average NIE and NDE were included. By contrast, Table2 has the NIEs

and NDEs discussed in the methodological overview: NIE(1) corresponds to NIE, NDE(0) to NDE, while NIE(0) corresponds to NIE_{alt}, and NDE(1) to NDE_{alt}[8]. In other words, setting the value at 1, or for the treatment, will mean that the given effect fully incorporates the effect of the interaction. Taking moral alignment as an example, when the whole interaction is attributed to the NIE (NIE(1)), it has an effect size of -0.244, mediates almost fully the effect of the treatment (Mediate %=98.9%), with a ϱ=0.7 needed to make the indirect effect non-significant, with 49% of the residual, and 25% of the total variation explained. Conversely, if none of the interaction is attributed to the NIE (NIE(0)), it has an effect size of -0.171, procedural justice only mediates a little more than two-thirds of the treatment's effect (Mediate %=69.5%), with a mean ϱ=0.5, which coincides with the residual variance of 25%, and the total variance of 13%.

Even if it is difficult to determine where to assign the effect of the interaction, Table2 can help to inform the researcher about the presence/absence of an influential T-M interaction. As an example, for risk of sanction (NIE(0)=-0.033, NIE(1)=-0.046) and sense of power (NIE(0)=-0.080, NIE(1)=-0.076), the traditional product method would have probably provided very similar indirect effects, as the allocation of the interaction does not seem to hugely affect these models' NIE. However, for personal morality (NIE(0)=-0.028, NIE(1)=-0.094) and the already mentioned moral alignment, the product method would have provided very different results, which would not have accounted for the impact of the T-M interaction.

VII. DISCUSSION

Much empirical research in the social sciences is focused on identifying causal relationships, and this is especially true for experimental studies. Yet, most of these efforts only scrutinise the average causal effects, they are not concerned with underlying causal processes and mechanisms. This article has discussed causal mediation analysis as a promising statistical method to "pry open" this black box of causality. This approach goes beyond the traditional product method and can be applied to models with non-linear link functions and interactions, without positing the effect homogeneity assumption, while quantifying the potential influence of unmeasured confounders for the mediator-outcome relationship through sensitivity analyses (Imai et al. 2010a 2010b; Imai et al. 2011).

The potential outcome framework used in this article is a rigorous tool that makes modelling assumptions explicit and offers new definitions of direct and indirect effects, which can be identified based on whether particular assumptions are satisfied. Future research would benefit from considering each step of the

[8] As noted earlier, the different decompositions will refer to the same total effect. For instance, for moral alignment it will be: TE=-0.215=NDE(mean)+NIE(mean)=-0.007+-0.207=NIE(1)+NDE(0)=-0.244+0.029= NIE(0)+NDE(1)=-0.171+-0.044.

sequential ignorability assumption and gauging whether the proposed causal mediation models are identifiable. Sensitivity analysis techniques would provide further insight into the robustness of emerging results, and could make tenuous relationships more discernible. At times, when parts of the experimental community are pre-occupied with the "replication crisis" and "p-hacking", these sensitivity analysis techniques could be readily applied as further tests regarding the viability of results.

To exemplify the utility of causal mediation analysis, this paper chose to reanalyse the ScotCET dataset. The assessment of the treatment effect constancy and heterogeneity of ScotCET shows that the treatment effect is very similar across the matched pairs, and that there is minimal design and small covariate heterogeneity, indicating that the treatment effect is produced by the experimental design. Causal mediation analysis allows a change in the focus of the analysis, moving from the equivocal treatment effect to the mediated effect of procedural justice policing. With this technique, future laboratory and field experiments may pursue similar effect decomposition, testing the extent to which different intermediate constructs channel the treatment's effect to various outcomes.

The rich set of variables from the ScotCET dataset allowed a wide-scale test of the theory of procedural justice policing. Most of the results align with the a priori falsifications checks (Figure 3). Procedural justice appears to channel the impact of previous experiences with the police towards moral alignment, duty to obey, social identity, and sense of power. For power distance, there is no significant mediated effect, while personal morality and risk of sanction are highly sensitive to unmeasured confounding. Nevertheless, not all falsification checks can be verified: procedural justice only seems to transmit the effect of the treatment to coerced obligation to obey when the covariates are not included in the model. This and the correlational analysis (Appendix/Table2/a) imply that, contrary to theory, coerced obligation to obey is not closely related to felt obligation to obey, and possibly not informed by procedural justice when basic pre-treatment covariates are accounted for. Future studies in the literature should address this lack of relationship and attempt to clarify the theoretical position of prudential obligation in the procedural justice literature.

As with every method, causal mediation analysis faces certain challenges that need to be addressed. Even with a randomised treatment the sequential ignorability assumptions are very demanding. For instance, in the ScotCET example, there might be influential covariates that were not measured and thus not included in the models (e.g. earlier contact with the police, victimisation). Unlike in other fields, such as epidemiology, where dozens of pre-treatment covariates are regularly considered, in the social sciences it is usually very difficult to find exhaustive lists of such covariates (VanderWeele 2015). Moreover, the results of the sensitivity analyses cannot be assessed on their own, but only with regard to the list of pre-treatment covariates that are accounted for. Noticeably, some of the results become more robust to unmeasured confounding when the covariates are not included in

the models (see: Appendix/Table4/a). This means that the robustness of the results can be only determined in comparison to other variables in the models, unless sensitivity benchmarks have been established.

Another potential criticism of causal mediation analysis is that it requires the assumption that only a single mediator will channel a treatment's effects towards the outcome. Yet, in the social sciences theories often posit multiple pathways. In non-Western countries, for example, police effectiveness is usually considered alongside procedural justice (Bradford et al. 2014). However, this would violate assumption (d) of the sequential ignorability assumption, which does not allow the presence of further mediators. Hence the method presented here can only be applied to relatively simple models and other more complex solutions need to be pursued when multiple mediators are present (Daniel et al. 2015; VanderWeele and Vansteelandt 2014).

Finally, this study's treatment merits some discussion. Even though the diagnostics of treatment consistency and heterogeneity indicate that the treatment's effect is only attributable to the design, still without knowing exactly what transpired during the roadside encounters, only a descriptive interpretation can be provided, which renders any explanation of the direct effects ambiguous. Moreover, it is plausible that the treatment effect without the discussed implementation failure would have produced different results. As with other experimental results, multiple trials are needed to revisit the findings presented here. Yet, by relegating the treatment's effects and elevating the mediated effects, causal mediation analysis permitted a clarification regarding to what extent these experiences were carried by procedural justice, thus producing theoretically valuable findings.

APPENDIX:

Assessment of treatment effect consistency and treatment effect heterogeneity

As with procedural justice, the treatment's effect for the other outcome variables is fairly consistent across the experimental blocks (Figures 1a-8a). Due to normal sampling variability there are usually a couple of blocks that do not align with the main trend, but this is expected in such comparisons (in fact, their absence would be suspect). The standard errors on the figures are not particularly meaningful, as the number of participants across the different blocks strongly vary from one to another. The 95% and 68% confidence intervals are marked on each figure to provide some sense regarding the distribution of the potential "true" values of the treatment effect.

Another way to assess the robustness of the treatment effect is to consider the potential of treatment effect heterogeneity. The "FindIt" R package and Squared Loss Support Vector Machine (L2-SVM) (Imai and Ratkovic 2013) were used to assess this potential heterogeneity. In experimental research blocking guarantees that the treated and control groups are identical with respect to the influential covariates, thus they cannot affect the treatment effect (Imai et al. 2008). Two L2-SVM models were fitted for each outcome and were subsequently compared to each other, one with only the covariates and the other with the covariates and the blocking design considered. The two models showed only miniscule differences in the average treatment effect (Mean=0.010, Minimum=0.006 Maximum=0.015), which indicates the robustness of the design and that the inclusion of the covariates was sufficient to capture the differences across the blocks.

In addition, treatment effect heterogeneity was considered regarding the influential covariates, where they were allowed to take on interactions with the treatment and with one another. Here, only minor differences were registered compared to the average treatment effects without these interactions (Mean=0.019, Minimum=0.006, Maximum=0.038), with significant treatment-covariate interactions emerging only in the model for risk of sanction. Due to these relatively small changes, and to preserve consistency across the models, the covariates were included in their original form. The detailed results from these heterogeneity analyses can be found in Table 1/a.

	ATE	*Design heterogeneity differences in ATEs*	*Covariate heterogeneity differences in ATEs*	*Treatment-covariate interaction*
Procedural justice	-0.321	0.006	0.016	NA
Moral alignment	-0.277	0.015	0.035	NA
Duty to obey	-0.465	0.009	0.038	NA
Coerced obligation	-0.095	0.009	0.006	NA
Social identity	-0.274	0.007	0.033	NA
Sense of power	-0.306	0.015	0.012	NA
Power distance	-0.025	0.012	0.006	NA
Personal morality	-0.531	0.008	0.007	NA
Risk of sanction	-0.147	0.006	0.016	Treat*female -0.089 Treat*owner*employed -0.759
Overall average	-0.271	0.010	0.019	

Table 1/a Average treatment effect, design and covariate heterogeneity, and treatment-covariate interactions (NA = not applicable)

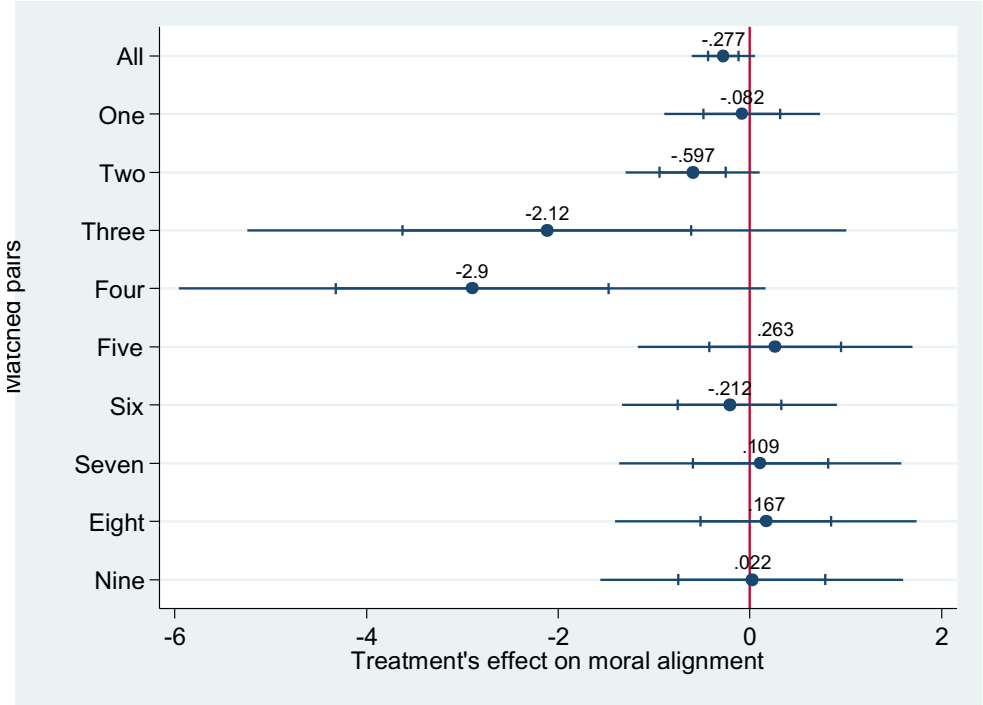

Figure 1/a Treatment effect consistency for moral alignment

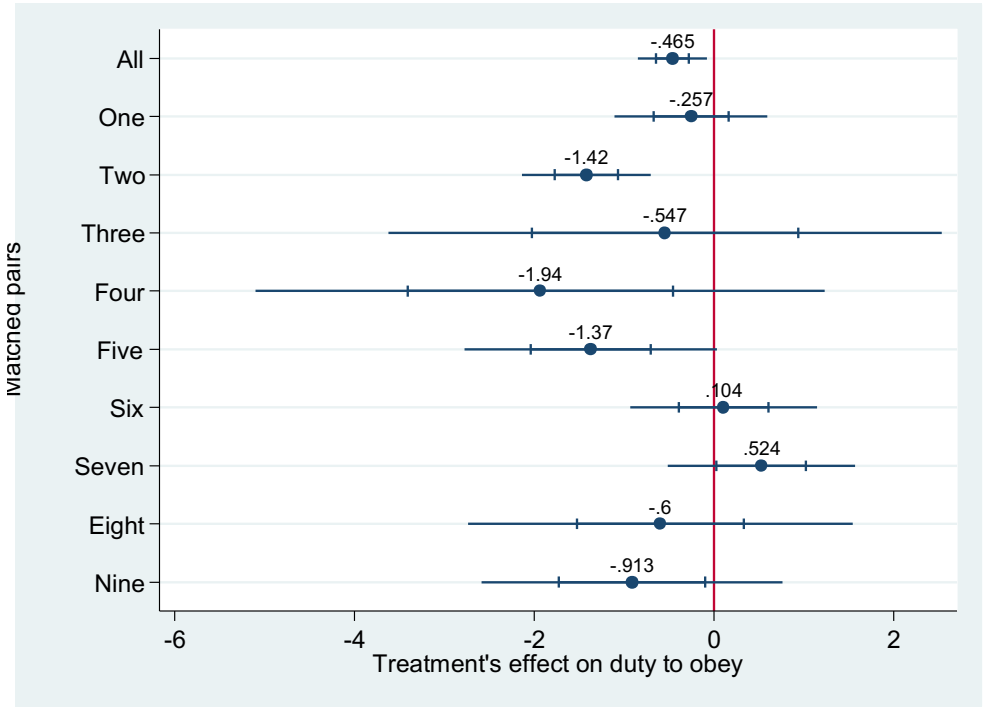

Figure 2/a Treatment effect consistency for duty to obey

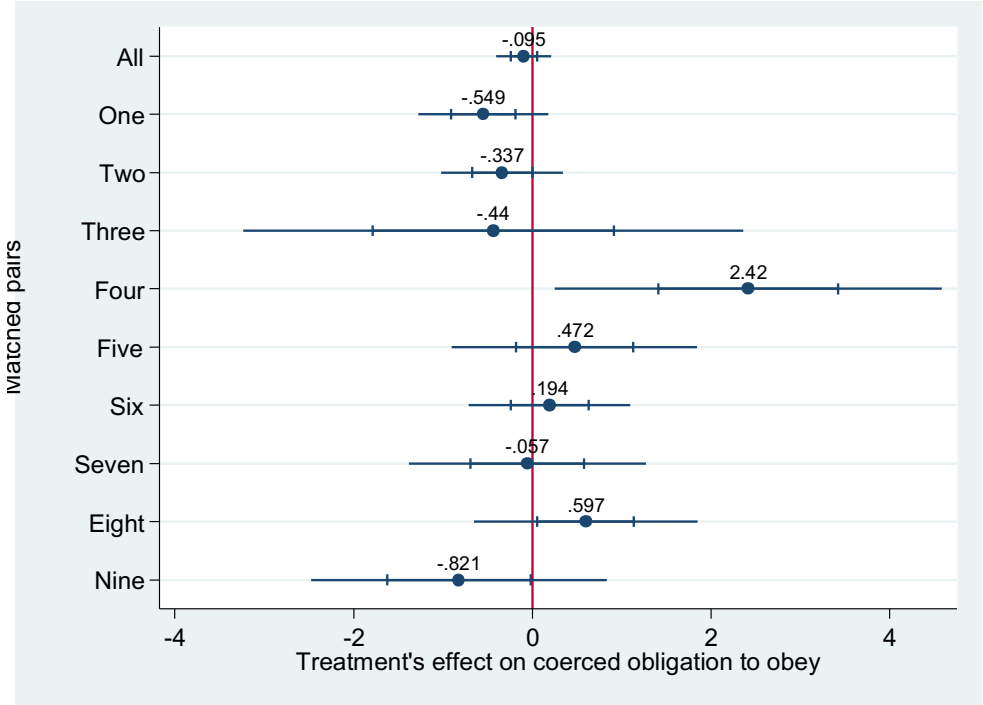

Figure 3/a Treatment effect consistency for coerced obligation to obey

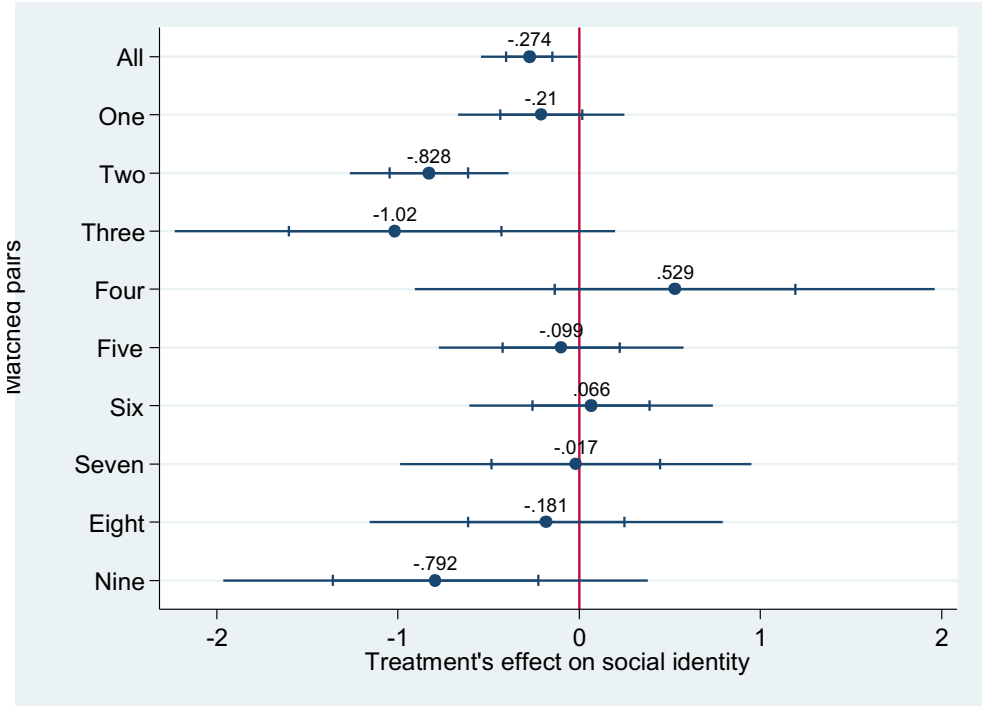

Figure 4/a Treatment effect consistency for social identity

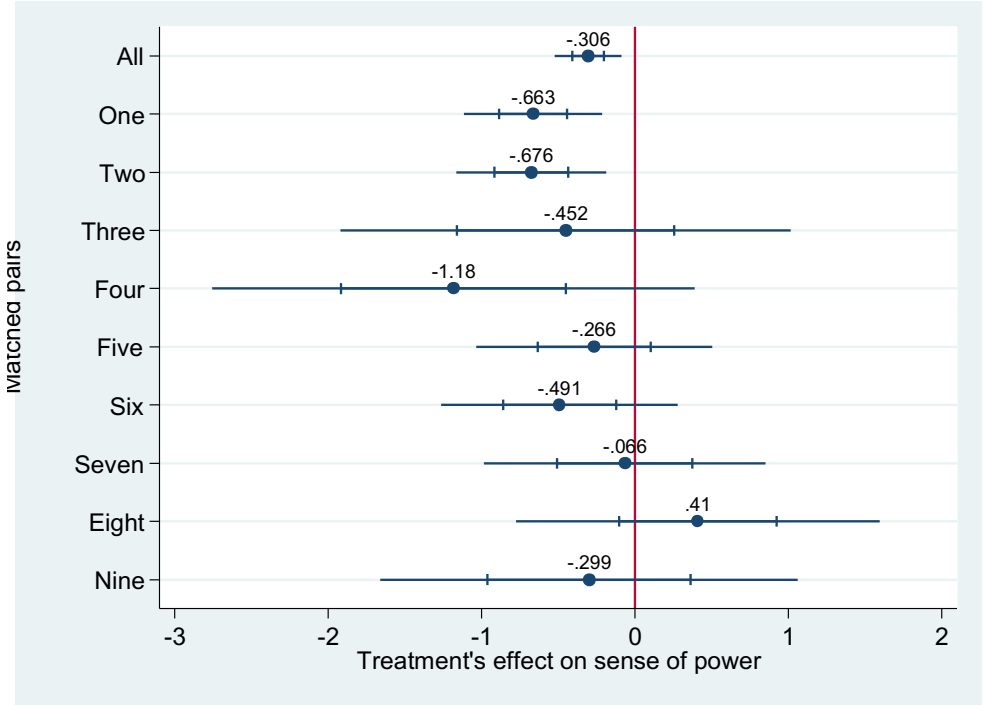

Figure 5/a Treatment effect consistency for sense of power

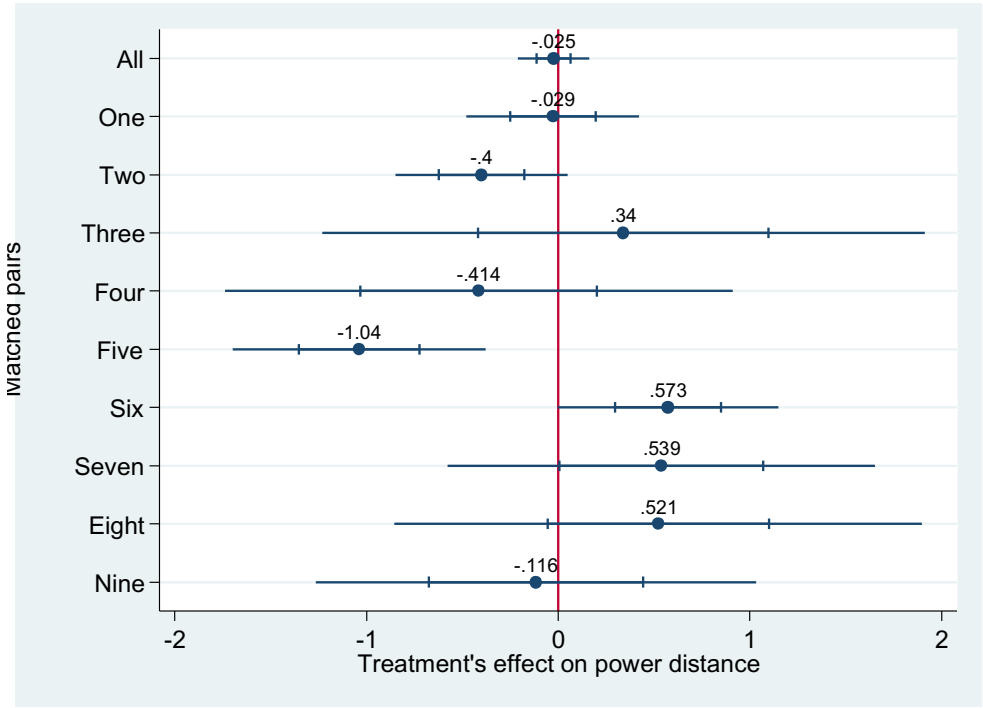

Figure 6/a Treatment effect consistency for power distance

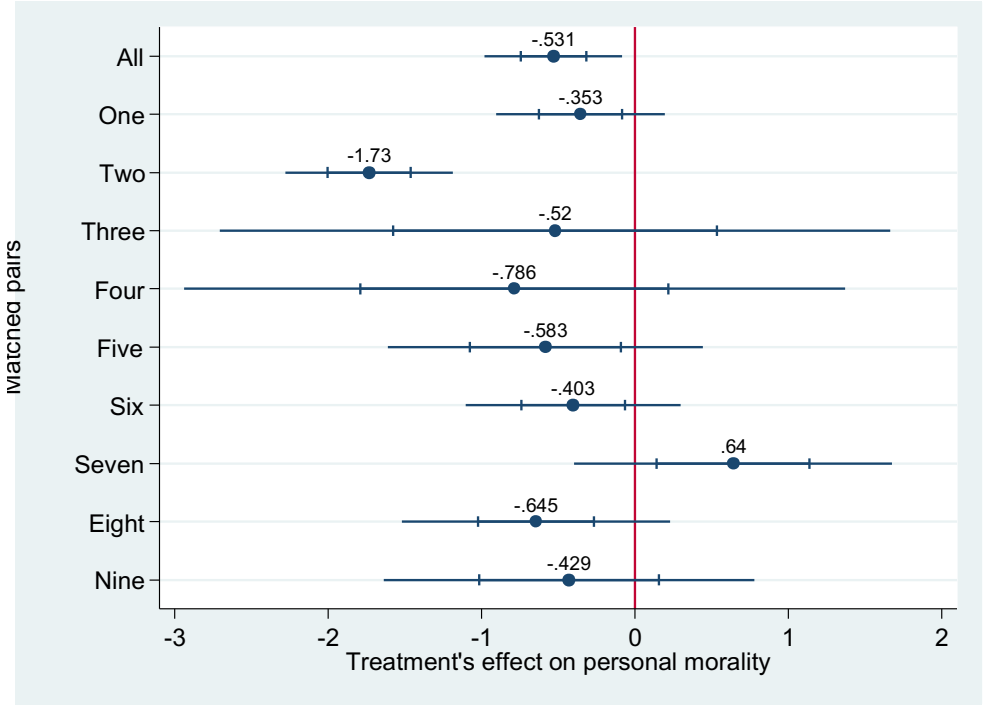

Figure 7/a Treatment effect consistency for personal morality

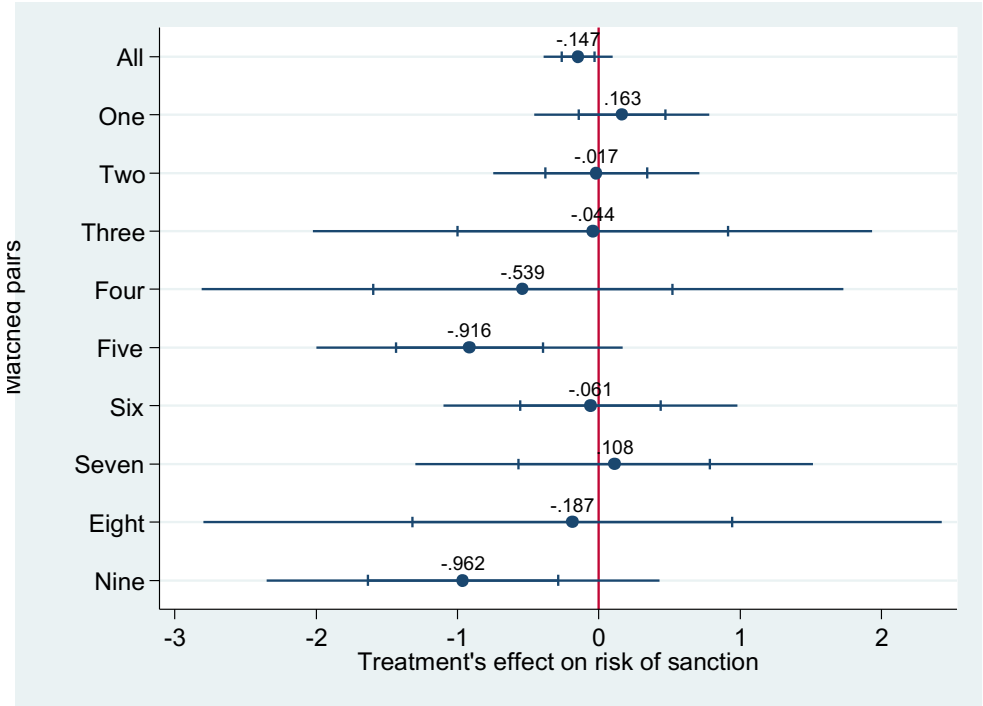

Figure 8/a Treatment effect consistency for risk of sanction

Estimation of the direct and indirect effects:

Following Baron and Kenny's (1986) seminal article, product method mediation analysis with a single mediator can be expressed as:

$$(1a) \quad M=\beta_0+\beta_1 t+\beta_2 c+\varepsilon_1$$
$$Y=\theta_0+\theta_1 t+\theta_2 m+\theta_3 c+\varepsilon_2$$

In the first equation, β_1 denotes the effect of the treatment on the mediator ('a' in Figure2) after taking into account the covariates (β_2) with the intercept (β_0) and error term (ε_1) In the second equation, θ_1 is the direct effect of T on Y ('c' in Figure2) after controlling for M (θ_2) ('b' in Figure2) and C (θ_3) with the constant (θ_0) and error terms (ε_2). The mediated (indirect) effect is the product of the coefficient of the treatment in the regression for the mediator (β_1) and the coefficient of the mediator in the regression for the outcome (θ_2).

To demonstrate how the new definitions of direct and indirect effects accommodate the interaction between T and M, θ_4 was added to the previous formula of the product method (1a), assuming the linearity of the effects. Also notice, that unlike (1a) the error terms are no longer present as they are expected to be $E(\varepsilon)=0$. In the equations, 't_0' refers to the control group, 't_1' refers to the treatment, while 't' refers to the effect of the treatment in the given equation. Provided that the previously discussed assumptions hold for the respective effects comparing t_0 and t_1, on average for the population, the following can be derived:

$$(2a) \quad CDE(t_1, t_0;m)=(\theta_1+\theta_4 m)(t_1-t_0)$$
$$(3a) \quad NDE(t_1, t_0; t_0)=(\theta_1+\theta_4(\beta_0+\beta_1 t+\beta_2 c))(t_1-t_0)$$
$$(4a) \quad NIE(t_1, t_0; t_1)=(\theta_2\beta_1+\theta_4\beta_1 t)(t_1-t_0)$$

From these formulas it can be easily discerned that when $\theta_4=0$, (2a) and (3a) coincide ($CDE(t_0,t_1;m)=NDE(t_0,t_1;t_1)=\theta_1(t_0-t_1)$), and (4a) is simplified to the traditional product method ($NIE(t_0,t_1;t_0)=\theta_2\beta_1(t_0-t_1)$). It follows that the product method is a special case of causal mediation analysis that produces valid estimates when the linearity and no interaction assumptions stand and the sequential ignorability assumption is satisfied (Imai et al. 2011).

As an alternative to these fully parametric models Imai et al. (Imai et al. 2011) have proposed a semiparametric estimation approach. Following their modelling strategy firstly, two regression models are fitted for the mediator and the outcome of interest, similarly to the parametric approach. Likewise, two sets of mediator (conditional on T and C) and outcome (conditional on M, T, and C) values are generated for every observation for each level of treatment $T=t_0$ and $T=t_1$. Again, in a similar vein, the effects are computed through averaging the differences between the predicted potential values. This approach is superior to the previous one in that it is applicable for any kind of link function, while the parametric one is

only applicable to a couple of special link functions (i.e. linear and binary logit with rare outcome variables) (VanderWeele 2015). Because of its flexibility, here the semiparametric approach was used but, notably, for linear outcome variables, the two approaches will generate almost identical results. Finally, both approaches recommend using resampling techniques, such as the nonparametric bootstrap or Monte Carlo approximation to correctly represent the prediction uncertainty of the estimates in these models.

Correlational results

The correlational results (Table2/a) show that the treatment had a weak negative association with the other variables. The correlation between treatment and personal morality emerged with the biggest magnitude ($r=-0.217$, $p<0.01$), followed by social identity ($r=-0.150$, $p<0.05$), duty to obey ($r=-0.144$, $p<0.01$), power distance ($r=-0.118$, $p<0.01$), moral alignment ($r=-0.114$, $p<0.05$), sense of power ($r=-0.113$, $p<0.05$), and procedural justice ($r=-0.103$, $p<0.05$). The association between perceived risk of sanction ($r=-0.085$, $p>0.05$) and coerced obligation ($r=-0.010$, $p>0.05$) did not reach statistical significance on the 5% level.

The mediator of interest, procedural justice, followed the expected pattern: it had a strong positive correlation with moral alignment ($r=0.698$, $p<0.01$), sense of power ($r=0.547$, $p<0.01$), and duty to obey ($r=0.463$, $p<0.01$), a moderately strong one with social identity ($r=0.298$, $p<0.01$), and a strong negative one with coerced obligation ($r=-0.411$, $p<0.01$), and weak negative one with power distance ($r=-0.074$, $p<0.05$). In addition, it had a weak and moderately strong positive relationship with risk of sanction ($r=0.150$, $p<0.01$) and personal morality ($r=0.309$, $p<0.01$), respectively.

It was presumed that there would be a negative association between coerced obligation, power distance and the rest of the variables. For coerced obligation this was found to be true on different levels of significance ($r=-0.033$- -0.411). Yet, coerced obligation had a non-significant negative bivariate relationship with duty to obey ($r=-0.061$, $p>0.05$), which raises questions as to whether the two variables are in fact two sides of the same coin. By contrast, for power distance the results were relatively obscure. Power distance had a weak positive relationship with duty to obey ($r=0.141$, $p<0.01$), social identity ($r=0.135$, $p<0.01$), and personal morality ($r=0.096$, $p<0.01$), and a non-significant one with moral alignment ($r=-0.043$, $p>0.05$), sense of power ($r=-0.015$, $p>0.05$), and risk of sanction ($r=0.176$, $p>0.05$). This indicates that further work is needed to determine the theoretical place of power distance in the model of procedural justice policing, while also implying that power distance and sense of power should be handled as separate constructs. Power distance and coerced obligation showed a moderately strong positive relationship ($r=0.300$, $p<0.01$).

Finally, the remaining variables had the anticipated significant positive bivariate relationships with one another with varying magnitudes (moral alignment: $r=0.149$-0.632, $p<0.01$; duty to obey: $r=0.141$-0.632, $p<0.01$; social identity:

r=0.173-0.387, p<0.01; sense of power: r=0.176-0.511, p<0.01; risk of sanction: r=0.147-0.290, p<0.01; personal morality: r=0.290-0.347).

Variable	*Treatment*	*Procedural justice*	*Moral alignment*	*Duty to obey*	*Coerced obligation*	*Social identity*
Procedural justice	-0.103*					
Moral alignment	-0.114*	0.689**				
Duty to obey	-0.144**	0.463**	0.632**			
Coerced obligation	-0.010	-0.411**	-0.359**	-0.061		
Social identity	-0.150*	0.298**	0.352**	0.356**	0.001	
Sense of power	-0.113*	0.547**	0.511**	0.387**	-0.248**	0.247**
Power distance	-0.118**	-0.074*	-0.043	0.141**	0.300**	0.135**
Risk of sanction	-0.085	0.150**	0.149**	0.147**	-0.033	0.173**
Personal morality	-0.217**	0.309**	0.347**	0.313**	-0.054	0.307**

Variable	*Sense of power*	*Power distance*	*Risk of sanction*
Power distance	-0.015		
Risk of sanction	0.176**	0.049	
Personal morality	0.300**	0.096**	0.290**

Table 2/a Correlational results
**p<0.05, **p<0.01, ***p<0.001*

List of constructs, measures, and response alternatives

Construct	*Items*	*Response alternatives*
Procedural justice	The police in Scotland make fair decisions. The police in Scotland listen to people before making decisions. The police in Scotland treat people with dignity and respect. The police in Scotland treat everyone equally.	1 – Hardly ever 2 – Not very often 3 – Some of the time 4 – Most of the time
Moral alignment	The police have the same sense of right and wrong as me. The police stand up for values that are important for people like me. I support the way the police usually act.	1 – Strongly disagree. 2 – Disagree. 3 – Neither agree nor disagree 4 – Agree 5 – Strongly agree
Duty to obey	I feel a moral obligation to obey the police. I feel a moral duty to support the decisions of police officers, even if I disagree with them. I feel a moral duty to obey the instructions of police officers, even when I do not agree with them.	
Coerced obligation	People like me have no choice but to obey the police. If you don't do what the police tell you they will treat you badly. I only obey the police because I am afraid of them.	
Social identity	I see myself as a member of the Scottish community. It is important to me that others see me as a member of the Scottish community.	
Sense of power	How much power do you think people like you have over the police?	1 – Very little power 2 – A little power 3 – Some power 4 – A lot of power
Power distance	How much power do you think the police have over people like yourself?	

Risk of sanction	How likely do you think is getting caught if… …breaking the speed limit while out driving …jumping a red light	1 – Not at all likely 2 – Not very likely 3 – Fairly likely 4 – Very likely
Personal morality	How wrong do you think is… …breaking the speed limit while out driving …jumping a red light	1 – Not wrong at all. 2 – Not very wrong. 3 – Fairly wrong. 4 – Very wrong.

Table 3/a List of constructs, measures, and response alternatives

Causal mediation analysis results without covariates

Procedural justice	*Type*	*Average effect*	*Mediate %*	*Mean ϱ*	*Residual R^2*	*Total R^2*
Moral alignment	NIE	-0.247* [-0.445, -0.067]	81.1%	0.6	0.36	0.21
	NDE	-0.047 [-0.292, 0.207]		~0.1	0.01	~0.01
Duty to obey	NIE	-0.179* [-0.325, -0.038]	44.2%	0.5	0.25	0.19
	NDE	-0.223 [-0.493, 0.052]		0.5	0.25	0.19
Coerced obligation	NIE	0.132* [0.033, 0.246]	36.5%	0.4	0.16	0.13
	NDE	-0.161 [-0.426, 0.103]		0.4	0.16	0.13
Social identity	NIE	-0.071* [-0.133, -0.012]	24.9%	0.3	0.09	0.07
	NDE	-0.209* [-0.384, -0.036]		0.8	0.64	0.55
Sense of power	NIE	-0.098* [-0.178, -0.027]	61.2%	0.5	0.25	0.18
	NDE	-0.062 [-0.194, 0.063]		0.3	0.09	0.06
Power distance	NIE	-0.001 [-0.027, 0.026]	~1%	0.1	0.01	0.01
	NDE	-0.218* [-0.392, -0.042]		0.8	0.64	0.61
Risk of sanction	NIE	-0.055* [-0.119, -0.010]	25.9%	0.2	0.04	0.04
	NDE	-0.119 [-0.373, 0.134]		0.3	0.09	0.09
Personal morality	NIE	-0.086* [-0.162, -0.020]	14.7%	0.3	0.09	0.08
	NDE	-0.485** [-0.733, -0.241]		0.9	0.81	0.69

Table 4/a Causal mediation analysis results without accounting for the pre-treatment covariates