

**Registered Replication Report:
Mazar, Amir, & Ariely (2008)**

Multilab direct replication of: Experiment 1 from Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45, 633-644.

Lead Authors: Verschuere, Bruno; Meijer, Ewout; Ariane, Jim; Hoogesteyn, Katherine; Orthey, Robin; McCarthy, Randy; Skowronski, John

Contributing Authors: Acar, Oguz A.; Aczel, Balazs; Bakos, Bence E.; Barbosa, Fernando; Baskin, Ernest; Bègue, Laurent; Ben-Shakhar, Gershon; Birt, Angie R.; Blatz, Lisa; Charman, Steve D.; Claesen, Aline; Clay, Samuel L.; Coary, Sean P.; Crusius, Jan; Evans, Jacqueline R.; Feldman, Noa; Ferreira-Santos, Fernando; Gamer, Matthias.; Gomes, Sara; González-Iraizoz, Marta; Holzmeister, Felix; Huber, Juergen; Isoni, Andrea; Jessup, Ryan K.; Kirchler, Michael; Klein Selle, Nathalie; Koppel, Lina; Kovacs, Marton; Laine, Tei; Lentz, Frank; Loschelder, David D.; Ludvig, Elliot A.; Lynn, Monty L.; Martin, Scott D.; McLatchie, Neil M.; Mechtel, Mario; Nahari, Galit; Özdoğru, Asil A.; Pasion, Rita; Pennington, Charlotte R.; Roets, Arne; Rozmann, Nir; Scopelliti, Irene; Spiegelman, Eli; Suchotzki, Kristina; Sutan, Angela; Szecsi, Peter; Tinghög, Gustav; Tisserand, Jean-Christian; Tran, Ulrich S.; Van Hiel, Alain; Vanpaemel, Wolf; Västfjäll, Daniel; Verlie, Thomas.; Vezirian, Kévin; Voracek, Martin; Warmelink, Lara; Wick, Katherine; Wiggins, Bradford J.; Wylie, Keith; Yıldız, Ezgi

Proposing Researchers: Bruno Verschuere & Ewout Meijer

Protocol vetted by: Nina Mazar, Dan Ariely, & On Amir

Protocol edited by: Daniel J. Simons

Citation: Verschuere, B., Meijer, E. H., Jim, A., McCarthy, R., Hoogesteyn, K., Skowronski, J., Orthey, R., Acar, O. A., (...), Yıldız, E. (2018). Registered Replication Report: Mazar, N., Amir, O., & Ariely, D. (2008). *Advances in Methods and Practices in Psychological Science*. Manuscript submitted for publication.

Data and registered protocols: <https://osf.io/vxz7q/>

Address Correspondence to: Bruno Verschuere, Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 VZ Amsterdam, The Netherlands. b.j.verschuere@uva.nl or Ewout Meijer, Department of Clinical Psychological Science, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands. eh.meijer@maastrichtuniversity.nl

Note: The first two authors share first authorship.

Keywords: cheating, morality, honesty, replication, many labs, preregistration

Acknowledgments: This research was funded by NWO grant number 401.16.001/3873. We would also like to thank the Association for Psychological Science (APS) and the Arnold Foundation who provided funding to participating laboratories to defray the costs of running the study. Thanks to Nina Mazar, On Amir, and Dan Ariely for providing materials for the study and for providing guidance about other tasks to include in the task battery. Thanks to Chris Chabris for providing the matrix reasoning task included as part of the battery. Thanks to Katherine Wood for assistance with the R scripts.

Running Head: DO MORAL REMINDERS REDUCE CHEATING?

Full list of authors with affiliations

Bruno Verschuere, University of Amsterdam
Ewout H. Meijer, Maastricht University
Ariane Jim, Ghent University
Randy McCarthy, Northern Illinois University
Katherine Hoogesteyn, Maastricht University
John Skowronski, Northern Illinois University
Robin Orthey, Maastricht University & University of Portsmouth
Oguz A. Acar, City, University of London
Balazs Aczel, Institute of Psychology, ELTE Eötvös Loránd University
Bence E. Bakos, ELTE Eötvös Loránd University
Fernando Barbosa, University of Porto
Ernest Baskin, Saint Joseph's University
Laurent Bègue, Université Grenoble Alpes
Gershon Ben-Shakhar, Hebrew University of Jerusalem
Angie R. Birt, Mount Saint Vincent University
Lisa Blatz, University of Cologne
Steve D. Charman, Florida International University
Aline Claesen, University of Leuven
Samuel L. Clay, Brigham Young University - Idaho
Sean P. Coary, Saint Joseph's University
Jan Crusius, University of Cologne
Jacqueline R. Evans, Florida International University
Noa Feldman, Hebrew University of Jerusalem
Fernando Ferreira-Santos, University of Porto
Matthias Gamer, University of Würzburg
Sara Gomes, University of Leuven
Marta González-Iraizoz, University of Warwick
Felix Holzmeister, University of Innsbruck
Juergen Huber, University of Innsbruck
Andrea Isoni, University of Warwick
Ryan K. Jessup, Abilene Christian University
Michael Kirchler, University of Innsbruck
Nathalie klein Selle, Hebrew University of Jerusalem
Lina Koppel, Linköping University
Marton Kovacs, ELTE Eötvös Loránd University
Tei Laine, Université Grenoble Alpes
Frank Lentz, Univ. Bourgogne Franche-Comté, Burgundy School of Business - Ceren
David D. Loschelder, Leuphana University of Lüneburg
Elliot A. Ludvig, University of Warwick
Monty L. Lynn, Abilene Christian University
Scott D. Martin, Brigham Young University - Idaho
Neil McLatchie, Lancaster University
Mario Mechtel, Leuphana University of Lüneburg
Galit Nahari, Bar-Ilan University

Asil Ali Özdoğru, Üsküdar University
Rita Pasion, University of Porto
Charlotte Pennington, University of the West of England
Arne Roets, Ghent University
Nir Rozmann, Bar-Ilan University
Irene Scopelliti, City, University of London
Eli Spiegelman, Univ. Bourgogne Franche-Comté, Burgundy School of Business - CEREN
Kristina Suchotzki, University of Würzburg
Angela Sutan, Univ. Bourgogne Franche-Comté, Burgundy School of Business - CEREN
Peter Szecsi, ELTE Eötvös Loránd University
Gustav Tinghög, Linköping University
Jean-Christian Tisserand, Univ. Bourgogne Franche-Comté, Burgundy School of Business - CEREN
Ulrich S. Tran, University of Vienna
Alain Van Hiel, Ghent University
Wolf Vanpaemel, University of Leuven
Daniel Västfjäll, Linköping University and Decision Research
Thomas Verliefe, University of Leuven
Kévin Vezirian, Université Grenoble Alpes
Martin Voracek, University of Vienna
Lara Warmelink, Lancaster University
Katherine Wick, Abilene Christian University
Bradford J. Wiggins, Brigham Young University - Idaho
Keith Wylie, Florida International University
Ezgi Yıldız, Üsküdar University

Abstract (current: ca 220 words)

The self-concept maintenance theory holds that many people will cheat in order to maximize self-profit, but only to the extent that they can do so while maintaining a positive self-concept. Mazar, Amir, and Ariely (2008; Experiment 1) gave participants an opportunity and incentive to cheat on a problem-solving task. Prior to that task, participants either recalled the 10 Commandments (a moral reminder) or recalled 10 books they had read in high school (a neutral task). Consistent with the self-concept maintenance theory, when given the opportunity to cheat, participants given the moral reminder priming task reported solving 1.45 fewer matrices than those given a neutral prime (Cohen's $d = 0.48$); moral reminders reduced cheating. The Mazar et al. (2008) paper is among the most cited papers in deception research, but it has not been replicated directly. This Registered Replication Report describes the aggregated result of 25 direct replications (total $n = 5786$), all of which followed the same pre-registered protocol. In the primary meta-analysis (19 replications, total $n = 4674$), participants who were given an opportunity to cheat reported solving 0.11 more matrices if they were given a moral reminder than if they were given a neutral reminder (95% CI: -0.09 ; 0.31). This small effect was numerically in the opposite direction of the original study (Cohen's $d = -0.04$).

Cheating is widespread and associated with substantial costs to society. As many as 60% of taxpayers evade taxes (Slemrod, 2007), and, in 2011, global tax evasion has been estimated to exceed 3.1 trillion USD (The Tax Justice Network, 2011). The self-concept maintenance theory (Mazar, Amir, & Ariely, 2008) holds that people try to maximize self-profit while maintaining a positive self-concept of honesty. This theory predicts that many people will cheat to benefit themselves, provided that they can preserve their positive self-concept. Often, that means people will justify small amounts of cheating. For example, participants asked to privately roll a die and to report the outcome, with higher rolls leading to higher financial gain, reported an average roll of 3.63. That amount is above what a random roll of a die should produce on average (3.50), but far from the maximum possible value of 6 (Halevy, Shalvi, & Verschuere, 2014).

According to the self-concept maintenance theory, people should be less likely to cheat when they think about their own honesty. In a well-known test of this prediction, Mazar, Amir, and Ariely (2008; Experiment 1) gave participants ($n = 229$) an opportunity and incentive to cheat on a problem-solving task. Prior to that task, participants either recalled the 10 Commandments (a moral reminder) or recalled 10 books they had read in high school (a neutral task). The problem-solving task was embedded amongst other filler tasks in a large booklet, and it required participants to find numbers in a matrix that add up exactly to 10 (e.g., 3.81 and 6.19; see Figure 1). After completing the task, half of the participants ripped the matrices sheet out of the booklet and wrote down the number of solved matrices on a separate scoring sheet in the booklet, giving them the opportunity to cheat. As a financial incentive for cheating, the task instructions stated that two randomly selected participants would receive \$10 for each matrix they reported solving. Thus, those participants had both an incentive and an opportunity to cheat;

they could receive payment and the only record of the number of matrices they had solved was their own self-report. Participants who tried to recall books they had read prior to the matrix task claimed to have solved 4.22 matrices, whereas participants who listed the 10 Commandments claimed to have solved 2.77 matrices. The other half of the participants were allocated to a control condition and did not tear out the matrices page, meaning there was no opportunity to cheat without being caught. In the control conditions, participants primed with recalling books or recalling the 10 Commandments solved similar numbers of matrices (3.06 and 3.12, respectively). Together, this pattern of results suggests that people who had the opportunity to cheat after recalling the 10 books cheated, while those who had the opportunity to cheat after recalling the 10 Commandments did not cheat (their performance was lower by 1.45 matrices; Cohen's $d = 0.48$).

Example		
1.69	1.82	2.91
4.67	3.81	3.05
5.82	5.06	4.28
6.36	6.19	4.57
Got it <input checked="" type="checkbox"/>		

Figure 1. Example of one matrix. The participant searches for the numbers that add up exactly to 10, here: 3.81 and 6.19.

The Mazar et al. (2008) paper has been cited over 1600 times on Google Scholar as of April 2018, and helped inspire research on how religious and other moral primes affect honest behavior (for a review, see Rosenbaum, Billinger, & Stieglitz, 2014). The 10 Commandments study also has political implications: It was cited as a critical building block of self-concept maintenance theory in a set of policy recommendations made to President Obama as part of the REVISE model (Shahar, Gino, Barkan, & Ariely, 2015). Knowing the size and reliability of this effect is critical both for policy makers and for self-concept maintenance theory. Yet, the literature includes no direct replication attempts for this study.

This Registered Replication Report (RRR) was designed to provide an accurate and precise estimate of the effect of “10 Commandment” priming on cheating in the matrix task. The focus of the RRR is on estimating the difference between the 10 Commandments priming condition and the 10 books priming condition in the number of matrices people reported solving. Given that the original study was conducted in the United States, and the vast majority of US inhabitants identify themselves as Christians (Pew Research Centre, 2015), we might expect a different outcome when testing in other cultures (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). Consequently, the RRR examines the heterogeneity of the effect across laboratories and also includes measures of religiosity to test the prediction that the effect of moral priming will be larger in laboratories whose participants hold stronger religious beliefs. These measures were collected after the primary tasks so that they would not influence the main outcome measure.

The RRR was announced by APS and on social media, and laboratories were invited to apply to contribute. By the deadline of Nov, 30, 2016 the editor had received and approved 29 applications to join. Twenty-five out of the 29 of the contributing laboratories completed the

study and provided data for the meta-analysis (see Footnote 1). The study was completed as part of a larger task battery that included tasks for another RRR project (McCarthy et al., 2018), and all contributing researchers are authors on both papers.

Methods

Protocol

The protocol for this study was developed in consultation with the original authors, Nina Mazar, On Amir, and Dan Ariely, who provided their materials and feedback on key aspects of the design. The final protocol and materials were approved by the original authors and are publicly available at <https://osf.io/vxz7q>. Note that participating labs were responsible for their own informed consent forms and for translation of the material (if not testing in American English or Dutch). When translations were required, laboratories were asked to translate the materials and then independently back-translate them to ensure accuracy of the translations. The editor helped coordinate translations so that all laboratories testing in a given language used the same materials.

Testing took place in large classrooms with at least 50 participants present simultaneously in any session, to ensure adequate anonymity as in the original study, which was run in one single session (labs were asked to aim for 100+ participants in each session). Each lab was required to collect usable data from at least 200 participants, with the sample being between 20-80% female. The tasks for this study were embedded in a larger battery of tasks (stapled into a packet), all of which were completed using paper-and-pencil. In total, there were 8 different versions of the packet (the four conditions for this study crossed with 2 conditions for the other RRR completed as part of the battery). Experimenters randomly shuffled the printed packets prior to each session to ensure random assignment of participants to conditions, and the packets included a cover page to mask the condition. The order of tasks in the packet was the same for all

participants and is listed in Table 1. Completion of the whole test battery took approximately 45 minutes. Participants were informed that some of the tasks would require them to time themselves, and a stopwatch was projected on a screen at the front of the room for that purpose. After providing written informed consent, participants worked through the tasks in the booklet.

Design and Procedures. Participants were required to be 18-25 year-old students who were compensated with extra-credit, module credit, course credit, or other non-monetary rewards (e.g., free workshop attendance, movie tickets).

The design manipulated two between-subjects variables: Priming task (recall 10 Commandments versus recall 10 books from high school) and whether participants had the opportunity to cheat or not (i.e., whether or not their self-reported number of matrices solved could be verified). For the Commandments prime, participants read the following instructions: “For this next task, please write down as many of the 10 Commandments from the Bible as you remember. Please time yourself and spend no more than 2 minutes on this task.” For the Books prime, they read: “For this next task, please write down the names of 10 books that you read in high school. Please time yourself and spend no more than 2 minutes on this task.” The combinations of the Priming factor and the Cheating factor yielded four conditions that will be referred to hereafter as Commandments-Cheat, Books-Cheat, Commandments-Control, and Books-Control.

The problem-solving task consisted of 20 matrices (half of which were unsolvable). Participants were told to allot 4 minutes to complete as many matrices as possible. The instructions on this page also noted that 2 participants, chosen at random from all participants in the study, would be paid \$10 for each matrix they solved (or the equivalent in another currency for labs testing outside of the United States). Participants in the Cheat conditions were asked to

tear out the page with the matrices and to keep it for themselves, handing in the remainder of the package that now contained only the page on which they reported the number of matrices they had solved. Participants in the Control conditions were asked to tear out a blank page (to mask the presence of other conditions in the testing session). Those participants submitted in their package both the page on which they reported the number of correctly solved matrices and the matrices sheet.

During protocol development and in consultation with the original authors, we identified several aspects of the original design that were not mentioned in the original paper but that potentially were important. In all cases, we used the same procedures as in the original study. First, unannounced to participants, half of the 20 matrices were actually unsolvable. Second, the example matrix (Figure 1) accompanying the written instructions showed two circled numbers adding to 10. However, the instructions did not specify that participants must circle no more than 2 numbers that added to 10, and it was left to participants whether they tried to solve the matrices with a set of more than two numbers. Third, as the matrix task was self-timed, participants could cheat either by over-reporting the number of correctly solved matrices or by taking more than the allotted 4 minutes and actually solving more matrices. In the latter case, participants would be cheating by violating the instructions rather than by inflating their performance report.

Differences from the original study. In the original study, participants in the Cheat conditions, but not those in the Control conditions, tore out a page from the booklet (the matrices). If some participants in a session tore out pages and others did not, that difference could suggest the presence of multiple conditions to the participants and it could reveal that participant's condition to the experimenters. To avoid this possible unblinding, participants in the RRR Control conditions tore out a blank page from the booklet.

The original study defined the number of matrices solved differently for the Cheat and Control conditions. For the Cheat conditions, for which participants kept the matrix page, the dependent measure was the self-reported number of matrices solved. For the control conditions, the experimenters coded the submitted matrix page to determine whether or not participants correctly circled the two numbers adding to 10, and the dependent measure used for analysis was the total number of correctly solved matrices (N. Mazar, personal communication, April, 2018). The RRR analysis used the self-reported total number of solved matrices for both the Cheat and the Control condition to ensure that differences between conditions could not be attributed to differences in the measure. We analyzed the results using the number of correctly solved matrices (coded by the experimenters from the matrix pages) for the Control conditions in an exploratory analysis.

In the original study, participants were not explicitly instructed to circle the numbers, but only to mark the ‘I got it’ box below each matrix they solved. The original authors noted that most participants in the control conditions followed the example set by the sample problem and spontaneously circled the two numbers adding to 10. To ensure the ability to verify the accuracy of self-reports, the RRR added explicit instructions in the Control conditions to circle the numbers adding up to 10 (see Footnote 2).

The RRR added “from the Bible” to the Commandments instructions because pilot testing showed that some participants (e.g., non-religious people) might not know what was meant by the “10 Commandments.” The RRR protocol also added text to the example matrix problem to ensure that the task was clear to participants (i.e., “In the example to the right, 3.81 and 6.19 add up exactly to 10”). For the RRR studies, we projected a stopwatch on screen at the front of the room rather than asking participants to use their own devices to time themselves. We

did so both to standardize procedures and because fewer students now carry watches, not all testing rooms have clocks, and smartphones might allow for cheating. Finally, the original authors have no record of the other tasks included in the testing battery. We selected a set of tasks, vetted by the original authors, that were not expected to influence performance on the primary task.

Inclusion criteria. To be included in the analyses, participants had to originate from a 20-80% female sample, and be 18-25 year-old students at the time of testing. They also had to follow task instructions and to complete all the tasks necessary for this replication study. These last two criteria were underspecified in the preregistered protocol, and the lead lab and editor clarified these criteria prior to examining the data or results. Not following task instructions included not having torn out the matrix/blank page or reporting having solved more than 20 matrices. Participants who wrote down no books/commandments and also reported spending no time recalling books/commandments were excluded for not having completed all the tasks. Note that participants were not excluded for not having completed other tasks in the packet. Finally, data were excluded when the experimenter did not administer the tasks correctly or when a testing session included fewer than 50 participants (to ensure adequate anonymity as in the original study).

Data-blind exceptions were allowed to the protocol (e.g., some labs were likely to recruit samples that had less than 20% males). All exceptions, the number of participants tested, and the number included in each lab's study can be found in Table 2. Note that all data, both excluded and included, are available on the OSF project page (<https://osf.io/vxz7q/>). Labs indicated in their data file whether or not data were included and, if not, the reason for the exclusion.

Results

The data analysis R scripts (R Core Team, 2013; R version 3.4.3) were written during the data collection phase, and registered before viewing data from the RRR studies (see <https://osf.io/vxz7q/>). Prior to conducting the primary data analysis, a data integrity script checked for potential errors in data entry or coding, and individual labs were asked to clarify or resolve potential errors. As is standard for RRR projects, the primary data analysis consisted of a random-effects meta-analysis (Simons, Holcombe, & Spellman, 2014). The effect size from each lab used in the meta-analysis was the difference in the mean number of solved matrices in the Books-Cheat condition and the Commandments-Cheat condition. To examine whether differences in the religiosity of participants across labs moderated the size of the effect observed in those labs, we conducted a preregistered exploratory meta-regression using the random-effects model (Thompson & Higgins, 2002). The analysis was based on the average across three single-item religiosity measures (see Table 1; separate analyses for each of the three religiousness measures can be found on the OSF project page (<https://osf.io/vxz7q/>)).

Primary Analyses

The primary analyses include data from 4674 participants from 19 laboratories that met all inclusion criteria or that were granted an a-priori or a results-blind exception. [The exceptions included: a sample with more than 80% females (7 labs) and allowing inclusion of participants up to 27 years of age (1 lab; exception granted but not needed)].

Our primary analysis concerned the meta-analytic difference between the Commandments-Cheat and the Books-Cheat conditions—do people given a moral prime report solving fewer matrices than those given a neutral prime when they are given the opportunity to cheat? In the original study, participants in the Commandments-Cheat condition reported solving 1.45 fewer matrices than in the Books-Cheat condition. In the RRR, participants reported solving

0.11 more matrices in the Commandments-Cheat condition than in the Books-Cheat condition (95% CI: -0.09 to 0.31; Figure 2). This corresponds to a Cohen's d of -0.04 (95% CI: -0.12 to 0.04; the negative sign reflects that the effect is numerically in the opposite direction of the original study). Seven out of the 19 studies showed an effect numerically in the same direction as the original study. Of those in the same direction, none had a 95% CI that excluded zero.

The sample of studies had no heterogeneity ($\tau^2 = 0$, $Q(18) = 13.16$, $p = .78$; Borenstein, Hedges, Higgins, & Rothstein, 2009), with 0% of the observed variance among the effect sizes attributable to systematic differences between studies (I^2). Together, these indices suggest that further analyses of moderation by religiousness are not warranted. For completeness, Figure 3 plots the moderation of the 10 Commandments effect by religiousness. The meta-regression showed no significant effect for religiousness, with the point estimate of the slope being 0.18, 95% CI [-0.46; 0.82], $p = .58$.

Ancillary Analyses: Other comparisons of interest

Mazar et al. (2008) predicted that a moral reminder would reduce cheating and they found that it completely eliminated cheating. Consequently, the reported number of matrices solved in the Commandments-Cheat condition should be comparable to that in the Commandments-Control condition. In the original study, this difference (Commandments-Cheat minus Commandments-Control) was -0.35 matrices. In the RRR, the meta-analytic effect was 0.24 matrices (95% CI: 0.03 to 0.44), with no significant heterogeneity across labs ($\tau^2 = 0.01$, $I^2 = 4.48$, $Q(18) = 19.23$, $p = .38$; Figure 4).

Second, the prime should not have an effect for those without an opportunity to cheat. Consequently, the reported number of matrices solved in the Commandments-Control condition should be comparable to that in the Books-Control condition. In the original study, this

difference (Commandments-Control minus Books-Control) was 0.05 matrices. In the RRR, the meta-analytic effect was 0.01 matrices (95% CI: -0.19 to 0.20), with no heterogeneity across labs ($\tau^2 = 0$, $I^2 = 0$, $Q(18) = 15.30$, $p = 0.64$; Figure 5).

Third, priming with books should not reduce the tendency to cheat. Consequently, the reported number of matrices solved in the Books-Cheat condition should be higher than that in the Books-Control condition. In the original study, this difference (Books-Cheat minus Books-Control) was 1.16 matrices. In the RRR, the meta-analytic effect was 0.15 matrices (95% CI: -0.03 to 0.34), with no heterogeneity across labs ($\tau^2 = 0$, $I^2 = 0$, $Q(18) = 14.00$, $p = .73$; Figure 6).

Fourth, the Cheat-Control difference should be greater for the Book conditions than for the Commandments conditions. In the original study, the Cheat-Control difference was 1.16 matrices for the Book conditions and -0.35 matrices for the Commandments conditions (difference = 1.51). In the RRR, the Cheat-Control difference was 0.11 matrices for the Book conditions and 0.35 matrices for the Commandments conditions (difference = -0.11; 95% CI: -0.39 to 0.17), with no heterogeneity across labs ($\tau^2 = 0$, $I^2 = 0$, $Q(18) = 16.21$, $p = .58$; Figure 7).

Ancillary Analyses: Primary outcome measure across all labs

We repeated the main analysis including the data of all 25 laboratories (total $n = 5786$) that submitted data for this RRR study. Participants reported solving 0.17 more matrices in the Commandments-Cheat condition than in the Books-Cheat condition (95% CI: -0.00 to 0.35; Figure 8). Seven out of the 25 studies showed an effect in the same direction as the original study. Of those in the same direction, none had a 95% CI that excluded zero. Cochran's Q revealed no heterogeneity in the sample of studies ($\tau^2 = 0$, $Q(24) = 17.46$, $p = .83$), and I^2 indicated that about 0% of the observed variance between effect sizes was caused by systematic

differences between studies. A meta-regression showed no significant effect for religiousness, with the point estimate of the slope being -0.26, 95% CI [-1.02; 0.50], $p = .50$ (see Figure 9).

Ancillary Analyses: Primary outcome measure across labs that strictly met all inclusion criteria

Finally, we repeated the main analysis including only the data of the 10 laboratories (total $n = 2645$) that strictly met all a priori inclusion criteria (no exceptions allowed). Participants reported solving 0.07 more matrices in the Commandments-Cheat condition than in the Books-Cheat condition (95% CI: -0.18 to 0.33; see Figure 10 on the OSF project page at <https://osf.io/vxz7q/>). Four out of the 10 studies showed an effect in the same direction as the original study. Of those in the same direction, none had a 95% CI that excluded zero. Cochran's Q revealed no heterogeneity in the sample of studies ($\tau^2 = 0$, $Q(9) = 4.71$, $p = .86$), and I^2 indicated that 0% of the observed variance between effect sizes was caused by systematic differences between studies. A meta-regression showed no significant effect for religiousness, with the point estimate of the slope being 0.26, 95% CI [-0.50; 1.02], $p = .94$ (see Figure 11 and further details the OSF project page at <https://osf.io/vxz7q/>).

Exploratory Analyses

To maximize power, we conducted all exploratory analyses on data from all 25 labs.

The original study found a higher number of matrices solved in the Books-Cheat condition than in the Books-Control condition, an effect that was attributed to cheating in the absence of a moral reminder when participants could cheat without risk of being caught (i.e., they submitted the matrix page with their packet). This RRR did not find this difference in the primary analysis, but the two analyses used different dependent measures for the control condition: The RRR used the self-reported total correct and the original study used the actual

number correct as verified by the experimenter. When we analyzed the data in the same way as in the original study, comparing self-reports for the Cheat condition to the verified number of correctly solved matrices for the Control condition, we found a similar (but smaller) effect: The reported number of matrices solved in Books-Cheat condition in the RRR was 0.56 higher than the actual number of matrices solved in the Books-Control condition (95% CI: 0.35 to 0.77; see Figure 12). Priming with the 10 Commandments did not result in reduced cheating when using this dependent measure, though: The reported number of matrices solved in Commandments-Cheat condition in the RRR was 0.83 higher than the actual number of matrices solved in the Commandments-Control condition (95% CI: 0.59 to 1.06; see Figure 13).

Discussion

Mazar, Amir, and Ariely (2008; Experiment 1) reported that recalling the 10 Commandments—a moral reminder—reduced cheating more than did recalling 10 books from high school. This project replicated their procedures but did not find evidence of reduced cheating following a moral reminder. The results from the primary analysis (19 labs, $N = 4674$) and two ancillary analyses (lenient inclusion criterion: 25 labs, $N = 5786$; strict inclusion criterion: 10 labs, $N = 2645$) were consistent in observing a “10 Commandments effect” close to zero. The effect was comparably small across labs, with no heterogeneity, suggesting that differences among labs are consistent with sampling error rather than unexplained moderation, see Footnote 3. For 24 laboratories, the confidence interval for this primary effect included zero, and the remaining lab found an effect in the opposite direction.

Given the discrepancy in findings, the differences between the RRR and the original study require consideration. A first difference is that the original study was run more than 10 years ago, at an elite university. The perceived rewards, perceived probability of getting caught,

and the perceived consequences of getting caught with cheating may have been different for the participants of the RRR. A second difference is the composition of the task battery that preceded the tasks for this RRR. Given that no record was kept of the tasks in the original battery, we selected new tasks for the RRR, and it is possible that the different tasks used in the two studies affected the extent of cheating observed. However, the tasks we selected were unrelated to the manipulation, and the lead authors and original authors agreed that there was no a-priori reason to predict that the chosen tasks would interfere with the manipulation or outcome measure.

The original study reported a difference of 1.16 matrices solved between the Books-Cheat condition and the Books-Control condition, a difference that was attributed to cheating when participants were given the opportunity to do so with impunity and in the absence of a moral reminder. This RRR found no difference between these conditions when using self-reported matrices solved, but found a similar difference in an exploratory analysis when using self-reports for the Cheat condition and verified correct responses for the Control condition (as in the original study). However, this difference might result from differences in those measures rather than differences in cheating. For instance, if participants did not reliably circle the two numbers for each matrix, their verified correct responses would be lower than their self-reported total, resulting in a difference between the Books-Cheat condition and the Books-Control conditions. Moreover, the difference using that dependent measure was greater rather than smaller (as it had been in the original study) when participants were primed with the 10 Commandments recall task, meaning that the RRR result was inconsistent with reduced cheating following a moral prime.

Future studies of the impact of moral reminders would benefit from using tasks that provide unambiguous evidence of cheating. Examples of such tasks include Gneezy's (2005)

deception game where participants can maximize self-profit by duping another player and variants of the coin tossing task that track participants prediction before the participant can maximize profit by falsely claiming having correctly predicted the result of the coin toss task (Peer, Acquisti, & Shalvi, 2014). Given skepticism about the effectiveness of religious priming (van Elk et al., 2015), future tests of the self-concept maintenance theory might further benefit from exploring the effectiveness of non-religious moral primes (e.g., an honor pledge; Mazar et al., 2008) to evaluate whether they have a stronger influence on the proposed balance between maximizing self-profit and feeling moral.

In sum, we did not observe the predicted reduction in cheating following priming with the 10 Commandments. These results question the effectiveness of using the 10 Commandments as a moral prime to reduce cheating.

References

- Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91, 340-345.
doi:10.1080/00223890902935878
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.
- Chabris, C.F., Engel, D., Kim, Y.J., Loken, E., Woolley, A.W., Malone, T.W., et al. (2018). *Using collective intelligence to develop a new test of individual intelligence*. Manuscript in preparation.
- Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95, 384-394.
- Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.
- Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research*, 40, 54-72.
doi:10.1111/hcre.12019
- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45, 633-644.
doi:10.1509/jmkr.45.6.633
- McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Profile of Mood States manual*. San Diego, CA: Educational and Industrial Testing Service.
- McCarthy, R.J., Skowronski, J.J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., Yildiz, E. (2017). Registered Replication Report: Srull & Wyer (1979).

Peer, E., Acquisti, A., & Shalvi, S. (2014). "I cheated, but only a little": Partial confessions to unethical behavior. *Journal of Personality and Social Psychology*, 106, 202-217.

<http://dx.doi.org/10.1037/a0035392>

Pew Research Centre (2015). America's Changing Religious Landscape. Retrieved from <http://www.pewforum.org/2015/05/12/americas-changing-religious-landscape/>

Advances in Methods and Practices in Psychological Science. Manuscript submitted for publication.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: <http://www.R-project.org/>.

Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45, 181-196. doi:10.1016/j.joep.2014.10.002

Shahar, A., Gino, F., Barkan, R., & Ariely, D. (2015). Three principles to REVISE people's unethical behavior. *Perspectives on Psychological Science*, 10, 738-741. doi:10.1177/1745691615598512

Simons, D., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, 9, 552-555. doi:10.1177/1745691614543974

Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion. *The Journal of Economic Perspectives*, 21, 25-48. doi:10.1257/jep.21.1.25

- Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, 37, 1660-1672. doi:10.1037/0022-3514.37.10.1660
- The Tax Justice Network (2011). The cost of tax abuse. A briefing paper on the cost of tax evasion worldwide. Retrieved from <https://www.taxjustice.net/wp-content/uploads/2014/04/Cost-of-Tax-Abuse-TJN-2011.pdf>
- Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21, 1559-1573. doi:10.1002/sim.1187
- Van Bavel, J. J., Mende-Siedlecki, P. M., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113, 6454-6459. doi:10.1073/pnas.1521897113
- van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, 6, Article ID 1365.

Footnotes

Footnote 1. Four labs were approved but did not complete the study according to the requirements of the protocol. Two labs changed aspects of the procedure that were important to the design (Huntjens, Sumampouw) and two labs (Batra, Willis) recruited fewer than 200 participants. Data from those four labs, if any, are available from their individual lab project pages that are linked from the Contributing Labs component of the OSF page. The exclusion criteria specified in the protocol turned out to be vaguely worded, a problem we discovered as labs began to code their data. Consequently, we made a results-blind decision to include labs that tested at least 200 participants before exclusions but fewer than 200 after exclusions in the ancillary analyses of all labs contributing data. Data from these labs were excluded from the primary analysis and from the ancillary analyses of data from labs that strictly adhered to all protocol requirements.

Footnote 2. This additional instruction went unnoticed by 9 labs during the translation process (see Table 2). These 9 labs were excluded from the ancillary analyses on labs that strictly met all inclusion criteria.

Footnote 3. Although there was no heterogeneity at the lab level, there might still be moderation of the effect at the individual level. Further analysis including moderators coded at the individual level might yield meaningful information about who cheats and who does not.

Table 1. *List of tasks in combined SW RRR and MAA RRR*

<i>Task</i>	<i>Description</i>	<i>RRR</i>
<i>Demographics and informed consent</i>	<i>Provided their age, sex and major and written informed consent</i>	<i>[Both]</i>
<i>Scrambled sentence</i> <i>(hostility priming)</i> <i>(Srull and Wyer, 1979, Exp. 1)</i>	<i>Mark for 30 groups of 4 words the 3 words that make a complete sentence (e.g., <u>child</u> <u>the</u> question <u>watch</u>). The correct solution was either 80% hostile OR 20% hostile</i>	<i>SW</i>
<i>Vignette</i> <i>(Srull and Wyer, 1979, Exp. 1)</i>	<i>Read short story about a man named Ronald who behaved in manner that could be seen as hostile (e.g. told a beggar to find a job)</i>	<i>SW</i>
<i>Judgement Ronald</i> <i>(Srull and Wyer, 1979, Exp. 1)</i>	<i>Judge man from Vignette on 12 characteristics (e.g., Unfriendly)</i>	<i>SW</i>
<i>Judgement Situations</i> <i>(Srull and Wyer, 1979, Exp. 1)</i>	<i>Judge 15 situations on hostility (e.g., Refusing to let a salesperson into their house)</i>	<i>SW</i>
<i>Abstract Reasoning</i> <i>(Chabris et al., 2018)</i>	<i>Solve the 10-item version of non-verbal intelligence task</i>	<i>[Filler]</i>

<p><i>Recall 10 commandments or 10 books</i></p> <p><i>(moral reminder)</i></p>	<p><i>Recall the 10 commandments.</i></p> <p><i>OR</i></p> <p><i>Recall 10 books from high school</i></p>	<p><i>MAA</i></p>
<p><i>Matrix</i></p> <p><i>(cheating opportunity)</i></p> <p><i>(Mazar et al., 2008; Exp 1)</i></p>	<p><i>In each of the 20 matrices, find the numbers that add up exactly to 10 (e.g., 3.18 and 6.82).</i></p> <p><i>Tear out blank page</i></p> <p><i>OR</i></p> <p><i>Tear out matrix page</i></p>	<p><i>MAA</i></p>
<p><i>Collection slip</i></p> <p><i>(Mazar et al., 2008; Exp 1)</i></p>	<p><i>List how many matrices solved</i></p>	<p><i>MAA</i></p>
<p><i>Alternative Uses Test</i></p> <p><i>(Guilford, 1967)</i></p>	<p><i>List as many possible uses of a paper clip</i></p>	<p><i>[Filler]</i></p>
<p><i>Religiousness</i></p>	<p><i>Report religiousness. Specifically, participants were asked to rate, on a scale from 1 (not at all) to 5 (completely),</i></p> <p><i>(1) How religious are you? (2) To what extent do you believe in a God? (3) To what extent do you believe in a punishing God?</i></p>	<p><i>[Preregistered , exploratory moderator of MAA]</i></p>
<p><i>Fatigue</i></p> <p><i>(POMS; McNair et al., 1971) and sleep</i></p>	<p><i>Report fatigue and hours of sleep in last night</i></p>	<p><i>[Exploratory moderator of MAA]</i></p>

<i>Time estimation</i>	<i>Estimate time taken in timed tasks of this battery</i>	<i>[Exploratory moderator of MAA]</i>
<i>HEXACO (Ashton & Lee, 2009)</i>	<i>Complete 60-item personality scale</i>	<i>[Exploratory moderator of MAA]</i>

Note. This table lists the order of all of the tasks included in the combined Srull and Wyer (1979; SW) Registered Replication Report (RRR) and Mazar, Amir, and Ariely (2008; MAA) RRR. It appears both here and in McCarthy et al. (this issue).

Table 2. Descriptive statistics and general information per participating lab.

Lab	Country (language)	Total tested (total included)	<i>M (SD)</i> Matrices Commandment s-Cheat	<i>M (SD)</i> Matrices Books-Cheat	Data-blind exceptions to inclusion criteria
Acar	UK (British English)	237 (163)	3.97 (2.93)	3.75 (2.48)	None
Aczel	Hungary (Hungarian)	245 (215)	3.18 (2.28)	3.44 (2.77)	Omitted added instruction to circle the two numbers
Baskin	USA (American English)	207 (173)	3.31 (2.21)	3.05 (3.19)	None
Birt	Canada (American English)	234 (205)	3.04 (2.77)	2.63 (2.31)	Lower male to female ratio
Blatz	Germany (German)	320 (196)	4.31 (3.47)	3.24 (3.47)	Lower male to female ratio; Omitted added instruction to circle the two numbers

Evans	USA (American English)	332 (231)	3.08 (2.88)	2.23 (2.31)	None
Ferreira-Santos	Portugal (Portuguese)	291 (211)	2.67 (2.56)	2.85 (2.41)	None
González-Iraizoz	UK (British English)	235 (214)	3.67 (3.05)	3.41 (2.46)	Lower male to female ratio
Holzmeister	Austria (German)	274 (246)	7.02 (4.33)	5.91 (3.63)	Omitted added instruction to circle the two numbers
klein Selle & Rozmann	Israel (Hebrew)	337 (283)	3.31 (2.40)	3.58 (2.66)	None
Koppel	Sweden (Swedish)	263 (236)	3.11 (2.13)	2.73 (2.31)	Omitted added instruction to circle the two numbers
Laine	France (French)	313 (224)	2.19 (2.09)	2.56 (2.32)	Lower male to female ratio; Omitted added instruction to circle the two numbers
Loschelder	Germany (German)	248 (212)	3.98 (1.79)	4.09 (2.27)	Omitted added instruction to circle the two numbers

McCarthy	USA (American English)	318 (218)	3.40 (3.76)	2.83 (3.46)	None
Meijer	Netherlands (English)	377 (336)	3.02 (1.52)	3.18 (2.33)	None
Özdoğan	Turkey (Turkish and English)	365 (237)	4.45 (2.88)	3.26 (3.06)	Lower male to female ratio; Omitted added instruction to circle the two numbers
Pennington	UK (British English)	255 (197)	2.85 (2.01)	2.10 (1.59)	None
Roets	Belgium (Dutch)	253 (192)	3.76 (2.45)	3.40 (2.13)	Lower male to female ratio
Suchotzki	Germany (German)	256 (240)	3.83 (2.25)	3.83 (2.89)	Lower male to female ratio; Omitted added instruction to circle the two numbers
Sutan	France (French and English)	304 (300)	4.68 (1.89)	4.66 (2.38)	None

Tran	Austria (German)	277 (191)	3.83 (3.18)	3.73 (2.23)	Omitted added instruction to circle the two numbers
Vanpaemel	Belgium (Dutch)	288 (227)	3.61 (1.65)	3.44 (2.22)	None
Verschuere	Netherlands (Dutch)	302 (265)	3.73 (2.30)	3.55 (1.97)	None
Wick	USA (American English)	367 (334)	3.19 (2.50)	3.28 (3.60)	None
Wiggins	USA (American English)	259 (240)	2.34 (2.19)	2.15 (1.79)	None
Across all labs	-	7158 (5786)	3.54 (2.74)	3.36 (2.73)	-

Figure 2. Forest plot presenting the meta-analytic effect size of the 10 Commandments effect (primary analyses). Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means.

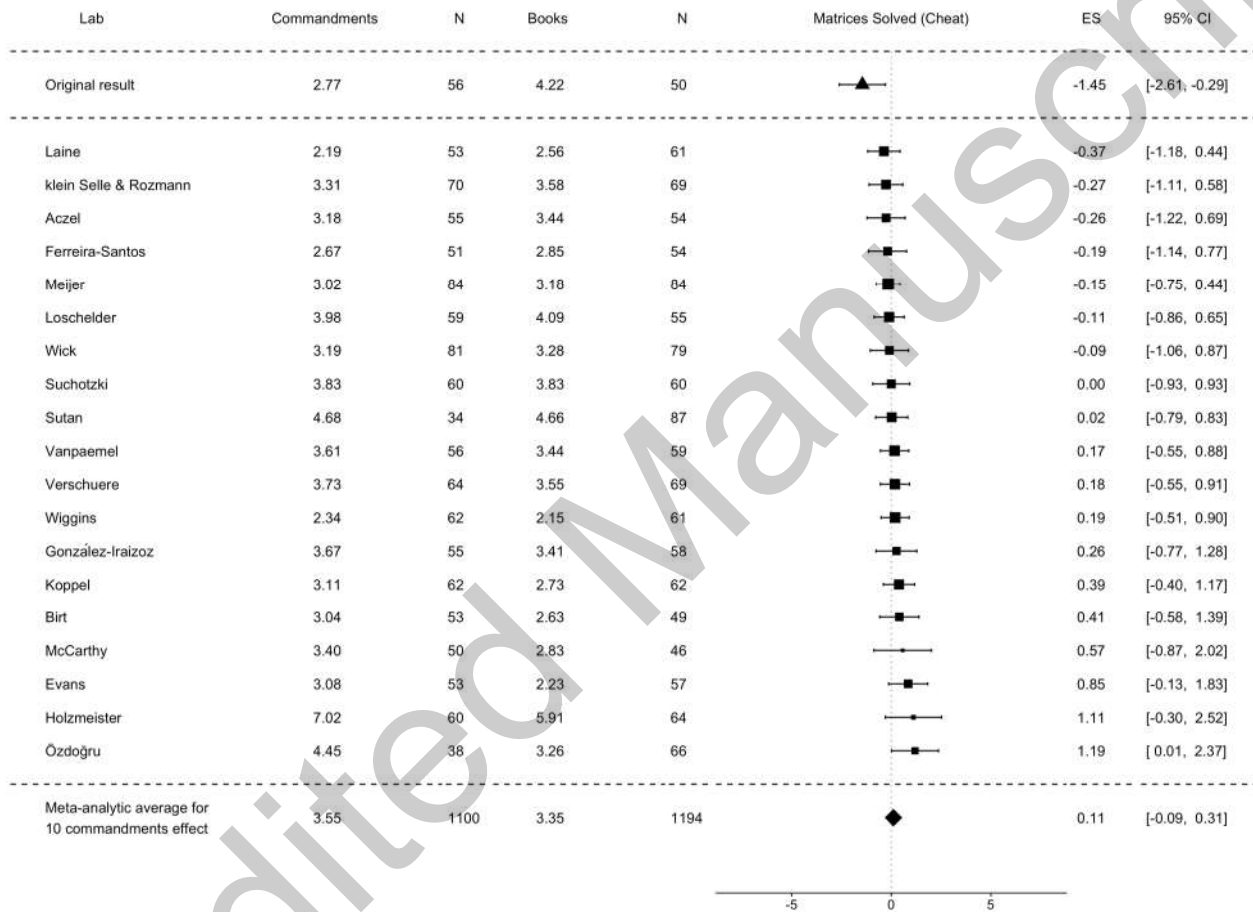


Figure 3. Moderation of the 10 Commandment effect by religiousness (primary analyses).

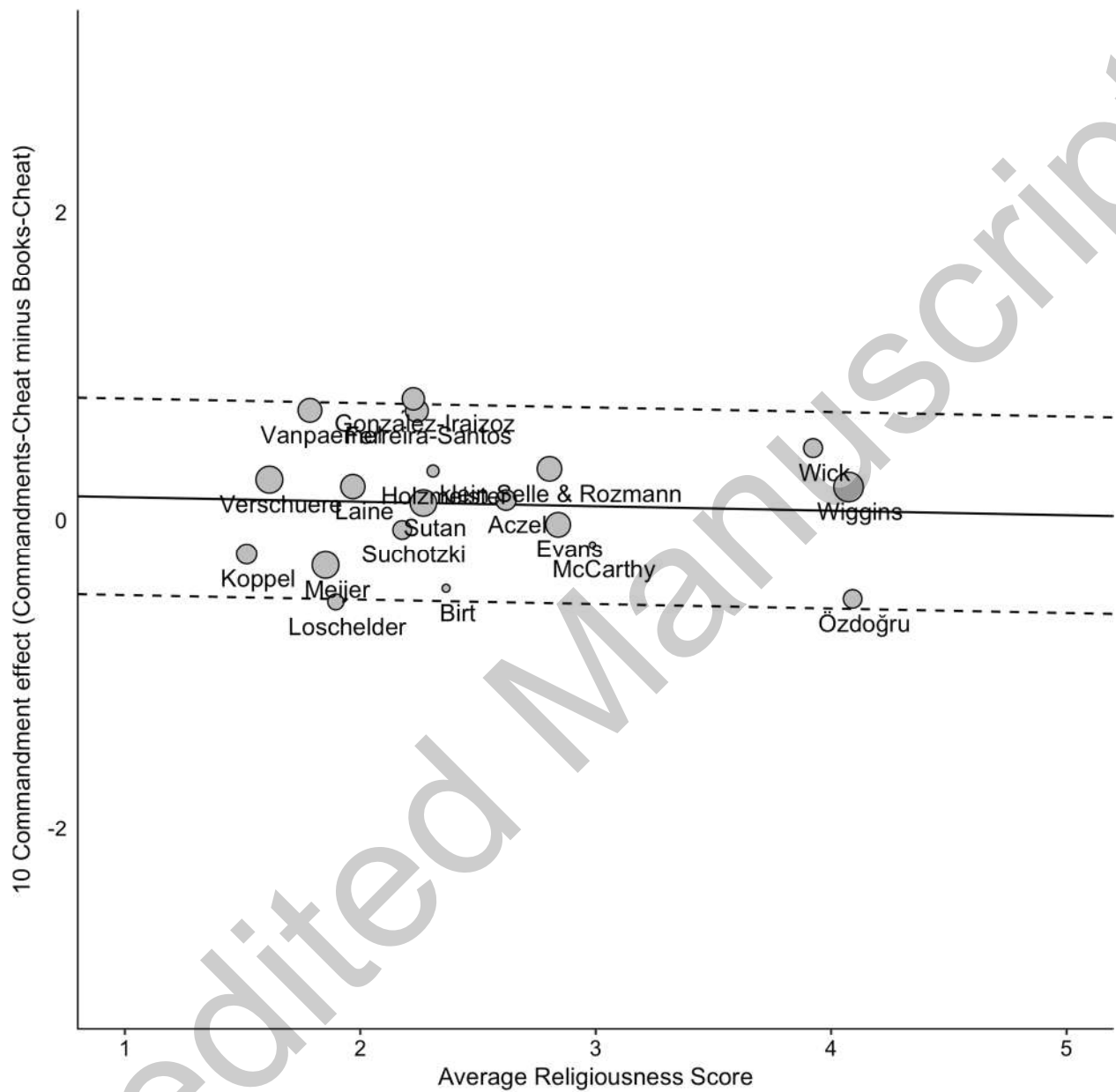


Figure 4. Forest plot presenting the meta-analytic difference between the Commandment-Cheat and the Commandments-Control conditions. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means

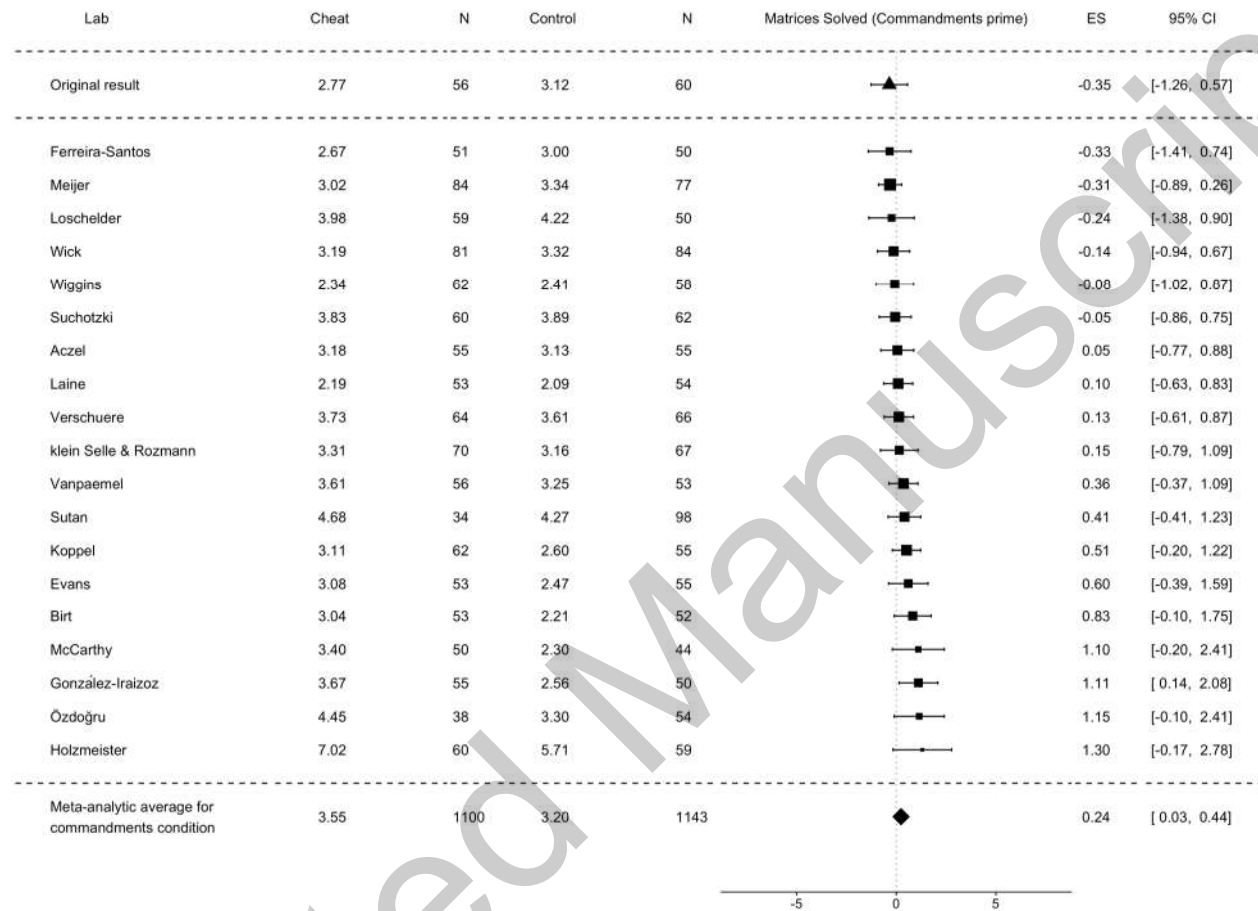


Figure 5. Forest plot presenting the meta-analytic difference between the Commandments-Control and the Books-Control conditions. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means

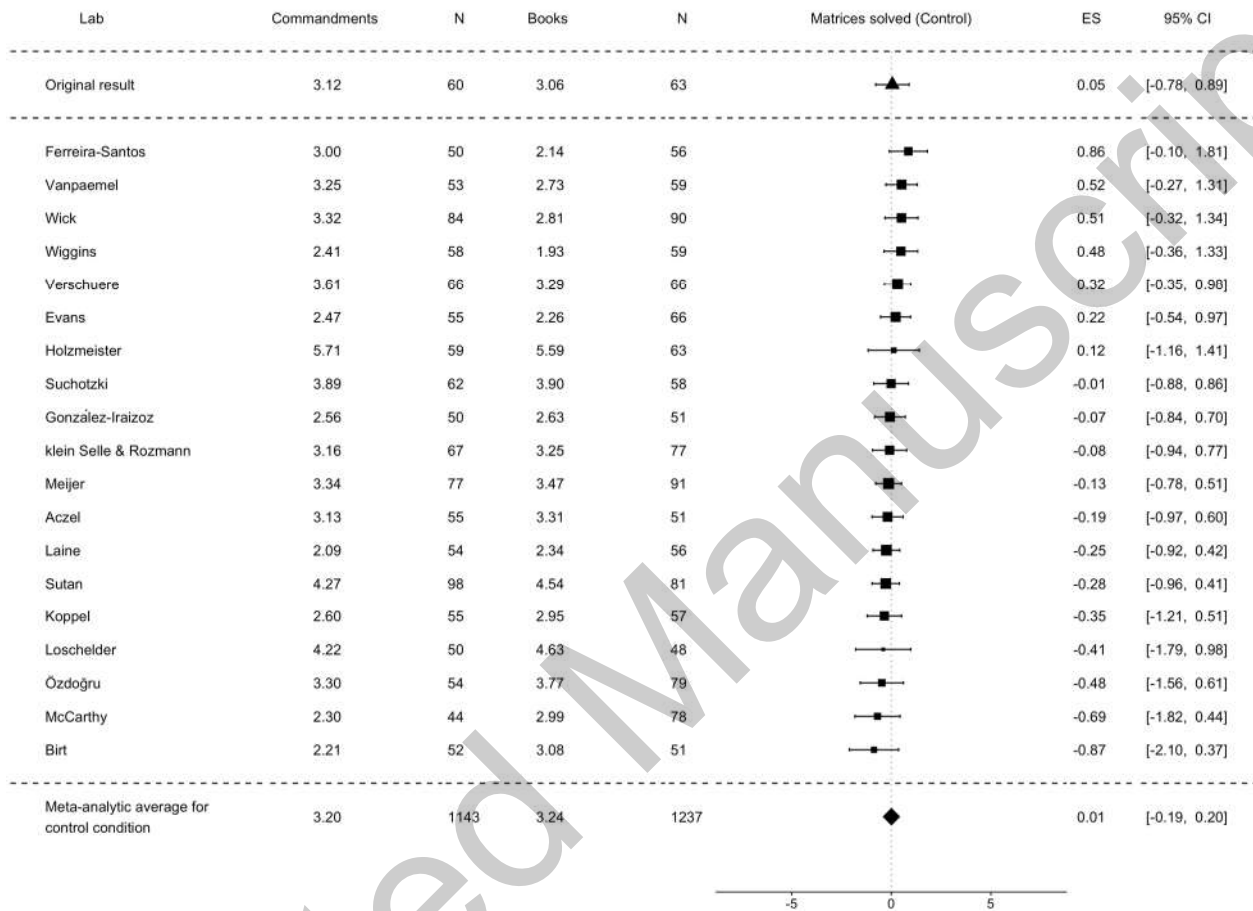


Figure 6. Forest plot presenting the meta-analytic difference between the Books-Cheat and the Books-Control conditions. Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means

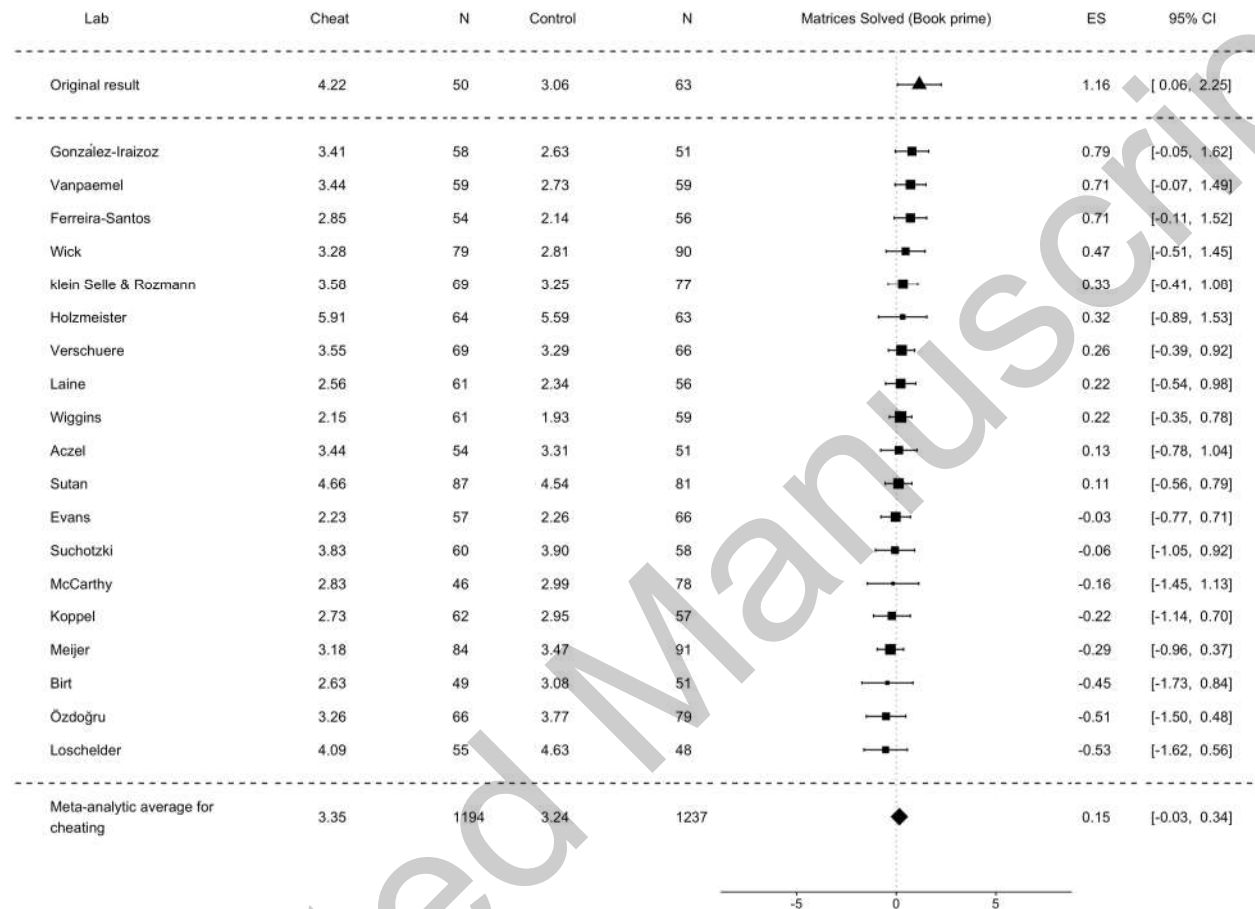


Figure 7. Forest plot presenting the meta-analytic difference between the Cheat-Control effect in the Books conditions (Books-Cheat minus Books-Control) versus the Commandments conditions (Commandments-Cheat minus Commandments-Control). Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means

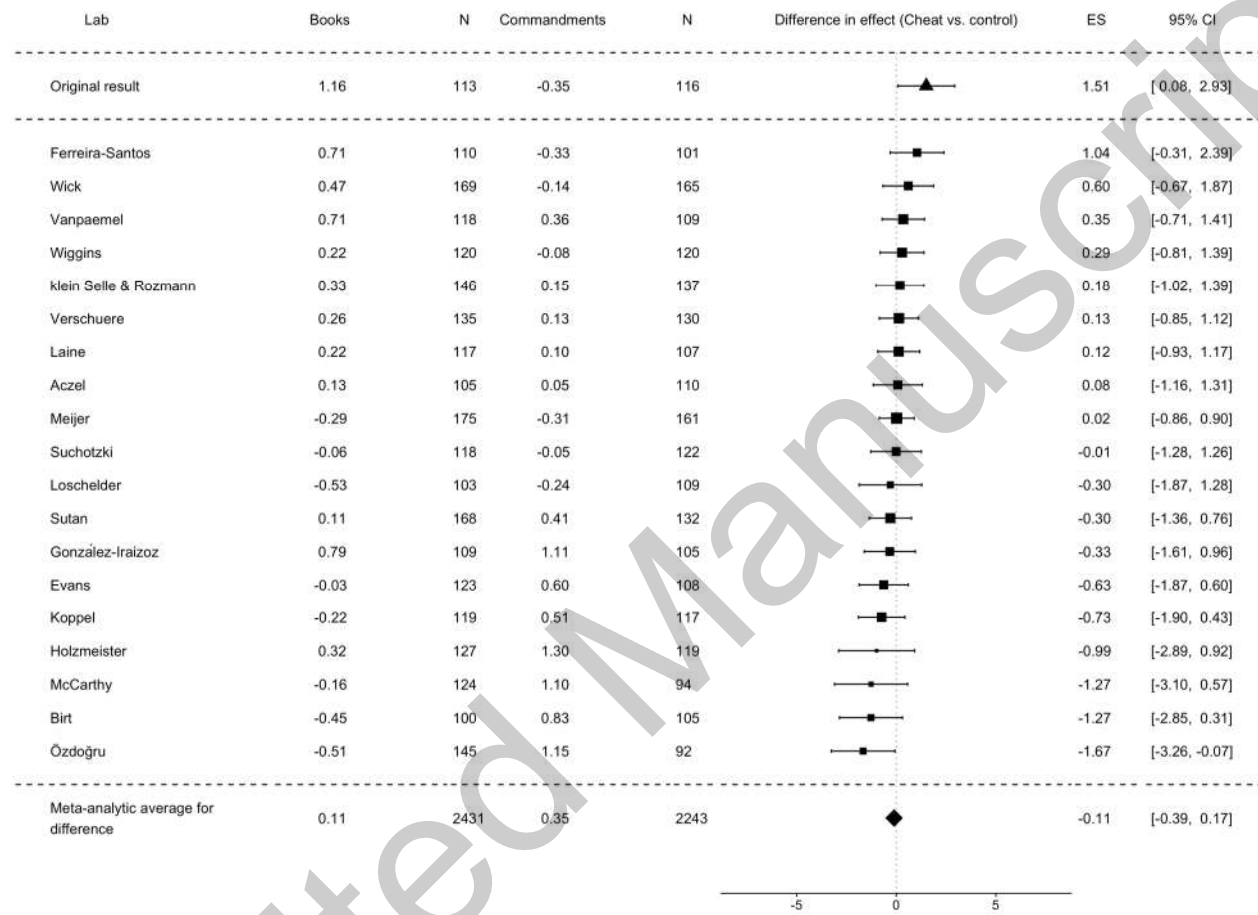


Figure 8. Forest plot presenting the meta-analytic effect size of the 10 Commandments effect (Ancillary analyses: All labs). Note that the meta-analytic estimate of the difference between conditions does not necessarily equal the difference between the means

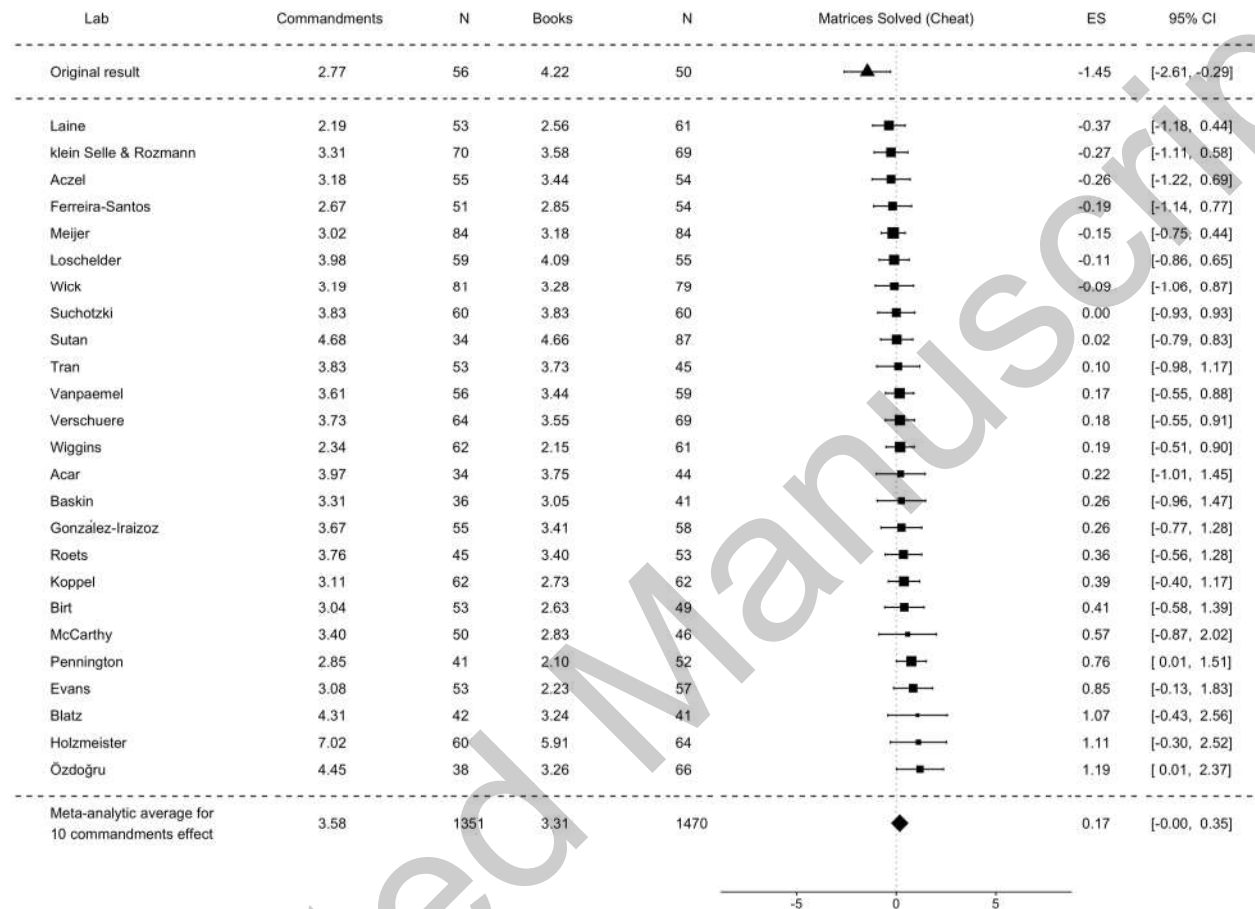


Figure 9. Moderation of the 10 Commandment effect by religiousness (Ancillary analyses: All labs).

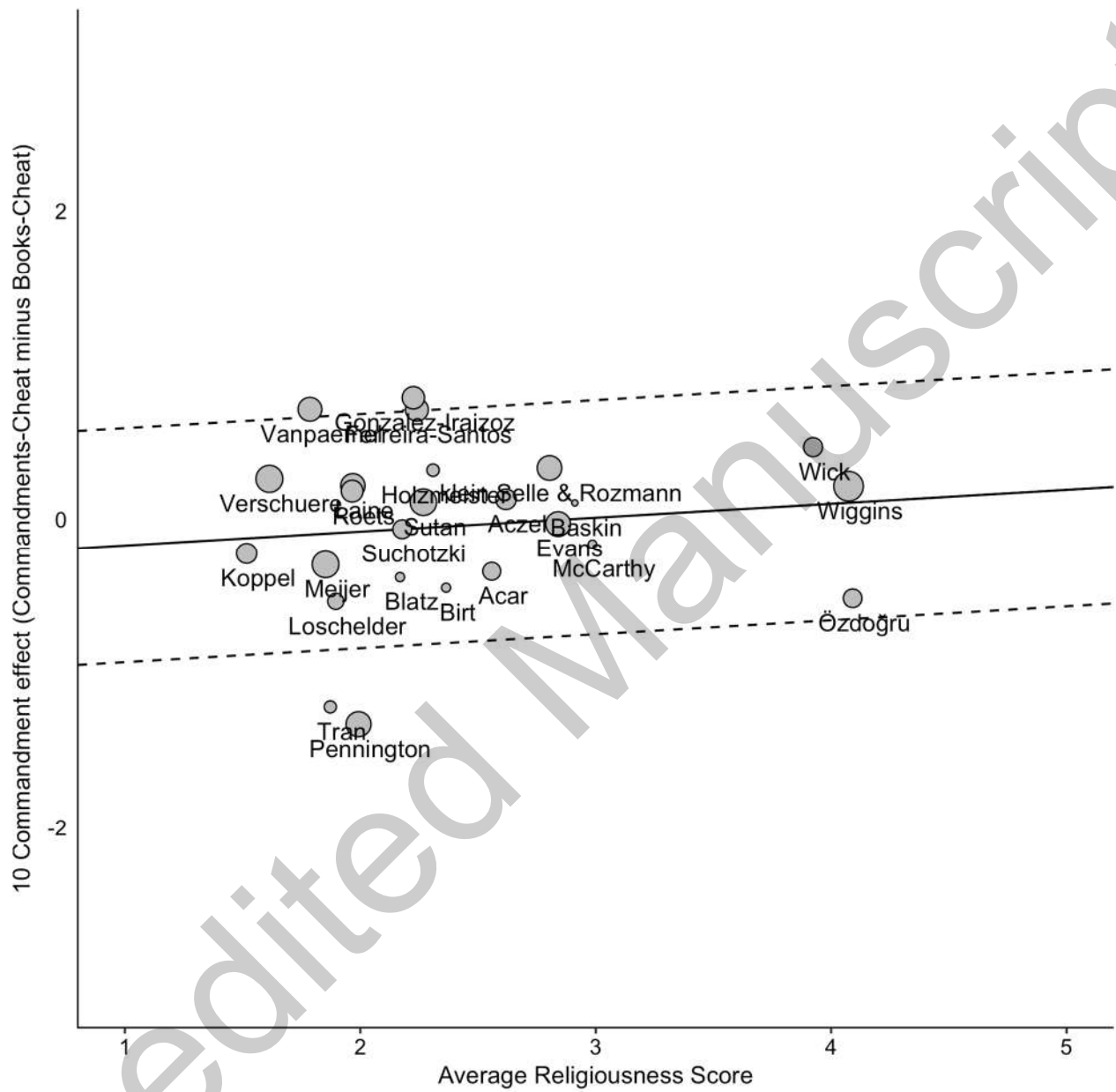
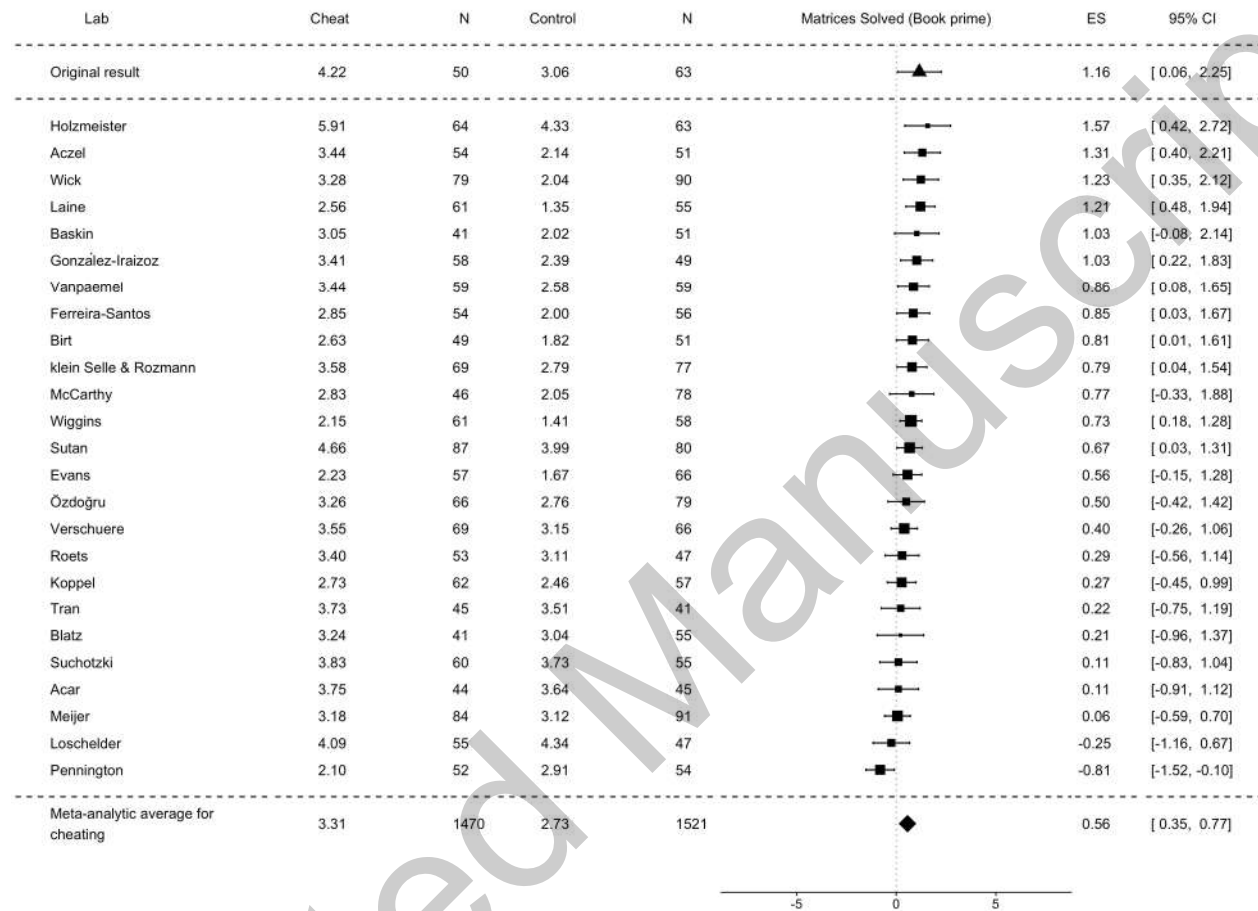


Figure 12. Forest plot presenting the meta-analytic effect size of the Reported solved in Book-Cheat versus Actual solved in Book-Control (Exploratory analysis on data from all labs).



Appendix

Individual Lab descriptions organized alphabetically by the last name of the first author of each lab

Posted on the OSF project page (<https://osf.io/vxz7q/>)