ORIGINAL RESEARCH REPORT

When Cheating is an Honest Mistake: A Critical Evaluation of the Matrix Task as a Measure of Dishonesty

Tom Heyman*,†, Hendrik Vankrunkelsven*, Wouter Voorspoels*, Anne White*, Gert Storms* and Steven Verheyen*,‡

Dishonesty is an intriguing phenomenon, studied extensively across various disciplines due to its impact on people's lives as well as society in general. To examine dishonesty in a controlled setting, researchers have developed a number of experimental paradigms. One of the most popular approaches in this regard, is the matrix task, in which participants receive matrices wherein they have to find two numbers that sum to 10 (e.g., 4.81 and 5.19), under time pressure. In a next phase, participants need to report how many matrices they had solved correctly, allowing them the opportunity to cheat by exaggerating their performance in order to get a larger reward. Here, we argue, both on theoretical and empirical grounds, that the matrix task is ill-suited to study dishonest behavior, primarily because it conflates cheating with honest mistakes. We therefore recommend researchers to use different paradigms to examine dishonesty, and treat (previous) findings based on the matrix task with due caution.

Somewhere between the honest truth and the deceptive lie is the deceptive truth and the honest lie (~Robert Brault).

Until recently, purely utilitarian views dominated theories on dishonesty, crucially hinging on the notion that people deliberately cheat after weighing the benefits of such behavior against the risks and costs associated with (potentially) being exposed as a cheater (Becker, 1968). Little over a decade ago, Mazar, Amir, and Ariely (2008) proposed an alternative theory, building on principles from social psychology, stating that people are guided by two conflicting forces: On the one hand, people are indeed tempted to maximize their comfort and profits, but on the other hand, their behavior is restrained by cultural norms and values that have been internalized through socialization. While people indeed want to reap the benefits from being dishonest, at the same time, they also want to maintain a positive self-concept. Put differently, people cheat, but only to the extent that they can still look at themselves in the mirror (see also Ariely, 2012).

In a series of studies, Dan Ariely, Francesca Gino, and colleagues have provided evidence to back this view on dishonest behavior. Among other things, they have shown empirically that circumstances which bring to mind internalized moral standards suppress the tendency to cheat, whereas providing or leaving open the possibility to come up with rationalizations of selfish choices have been demonstrated to increase dishonest acts (Ayal & Gino, 2011; Gino, Ayal, & Ariely, 2009; Gino et al., 2013a; Gino, Schweitzer, Mead, & Ariely, 2011; Kouchaki & Gino, 2016; Shu, Gino, & Bazerman, 2011; Shu, Mazar, Gino, Ariely, & Bazerman, 2012).

Besides a theoretical framework, Mazar et al. (2008) also introduced a new paradigm to examine dishonest behavior, called the matrix task. Although there are different variants (see Gerlach, Teodorescu, & Hertwig, 2019, for a review), the basic principle is that participants receive several matrices, typically 20, in which they have to find two numbers that sum to 10 (see **Figure 1** for an example). Participants are told to find as many correct solutions as possible within a limited time frame

* KU Leuven, BE

† Leiden University, NL

‡ Erasmus University Rotterdam, NL

Corresponding author: Tom Heyman (t.d.p.heyman@fsw.leidenuniv.nl)

1.69	1.82	2.91
4.67	4.81	3.05
5.82	5.06	4.28
6.36	5.19	4.57

Figure 1: An example of the stimuli used in the matrix task (adapted from Mazar et al., 2008). Participants have to circle two numbers that sum to 10.

(four minutes in Mazar et al.), knowing that their performance will be rewarded somehow (e.g., $0.5 per matrix solved). The short duration of the task ensures that only exceptionally gifted people would be able to complete all matrices. Critically, in one condition (i.e., the cheat condition), participants are given the opportunity to cheat by overstating their true performance. In the first study by Mazar et al., this was accomplished by providing a *separate* answer sheet on which participants needed to indicate how many matrices they had solved correctly. Participants kept the work sheet featuring the matrices, and handed in the answer sheet, thus allowing them to overclaim without risking to get caught. In contrast, participants in the control condition have their work sheet verified by an experimenter. The difference between the claimed performance in the cheat condition and the actual performance in the control condition (i.e., number of correctly solved matrices) is attributed to dishonest behavior. In other studies using the matrix task, the work sheets are (supposedly) discarded or destroyed (e.g., Gino et al., 2009; Zhong, Bohns, & Gino, 2010), again providing participants the opportunity to cheat. However, unbeknownst to them, the answer sheet they handed in can be linked to their work sheet via a hidden identifier. Consequently, someone's claimed performance can be compared to his/her actual performance, and any overreporting is considered cheating.

The matrix task has been used in over 100 experiments since its introduction by Mazar and colleagues in 2008 (see Gerlach et al., 2019), resulting in a substantial body of empirical evidence supporting contemporary theories on dishonest behavior. However, in the current paper, we argue that the matrix task provides an invalid measurement of dishonesty. The crux of our argument is that overreporting in the matrix task often arises as a result of honest mistakes. Due to this confound, most of the conclusions drawn from the paradigm need to be revisited. To build our case, we briefly discuss the rationale behind the task and point to a number of theoretical arguments and previous findings that undermine the validity of the matrix task as a method to examine dishonesty. Then, we describe a new empirical study that supports our assertion.

People can add, right?

According to Mazar et al. (2008), the matrix task is essentially a type of search task: Although it may take some time to locate the complementary numbers in each matrix (i.e., 4.81 and 5.19 in **Figure 1**), once found, it should be easy for participants to "unambiguously evaluate whether they had solved the question correctly (assuming that they could add two numbers to 10 without error), without the need for a solution sheet and the possibility of a hindsight bias" (p. 636).

Can people indeed add two numbers to 10 without error? It turns out that answering this question is not as straightforward as it seems. As with most things, it depends. For example, 7 + 3 versus 7.4379 + 2.5621 are obviously not equally trivial. Unsurprisingly, addition accuracy and response times have been shown to depend on problem size as well as the involvement of carry operations (i.e., when a 1 is transferred from one digit

position to another as in, say, 27 + 35, where 7 + 5 gives 12 from which the 1 is carried to the left; see e.g., Klein et al., 2010). The additions in the typical matrix task involve three digits and multiple carry operations, hence one should by no means expect a perfect performance, especially from individuals with limited mathematical abilities. To further complicate the matter, some matrices feature distractors: The example matrix (see **Figure 1**, adapted from Mazar et al., 2008) not only contains the correct solution (i.e., 4.81 + 5.19), but also a pair that sums to 10.1 (i.e., 5.82 + 4.28). The inclusion of such distractors conceivably increases the number of mistakes. Indeed, in a control condition similar to Mazar et al.'s (2008), Kajackaite (2018) found that "wrong reporting seems to be caused by honest mistakes (e.g., the most common mistake was 9.41 + 0.49 = 10)" (p. 197). If we also factor in time pressure, which leads to errors even in the most trivial tasks, it should not come as a surprise that participants would indeed make honest mistakes.

Of course, one has to evaluate these possible objections in the context of the matrix task in its totality. The idea is that participants, if given ample time, should be able to *check* their answers and *correct* any potential mistakes. Put differently, participants might initially make mistakes when filling in the work sheet, but they will review their responses, thus reporting only the number of correctly solved matrices on their answer sheet. However, to what extent this is communicated to participants is not always clear. Moreover, seeing that some people even fail to check whether they filled in all questions on an exam, should we then expect participants to verify their answers in the matrix task? After all, they have nothing to gain by checking (as opposed to the exam example), and some might not bother anyway as they consider the task a nuisance, or because they overestimate their ability. Ironically, the latter thought process is the very justification provided by Mazar et al. (2008) of why honest mistakes should never happen: everyone can count to 10, after all (so why would one need to review one's answers). In other words, the matrix task lumps together potential cheating with math ability, negligence, laziness, annoyance, overconfidence, et cetera. The difference between reported performance and actual performance may reflect a deliberate act of dishonesty, or (a combination of) the other processes and motives listed above.

Small lies, honest mistakes?

To our knowledge, only one study, by Faravelli, Friesen, and Gangadharan (2015), explicitly acknowledged the possibility that many small lies in the matrix task are in reality honest mistakes. Faravelli and colleagues actually demonstrated the robustness of their findings (regarding the relation between competition and dishonesty) by also considering the possibility that small lies were in fact mistakes. Even more important for the present discussion is that they also provided three arguments against the notion that (most) small lies were really honest errors. First, Faravelli et al. presumed participants correct themselves (but see above). Second, they argued that honest mistakes should lead to some *underreporting* too, which only happened in two out of 119 participants. Critically, this

notion applies to the act of *counting* the number of solved matrices, not *solving* the matrices or *reviewing* the answers. Honest mistakes in the former process should indeed balance each other out, but, as already discussed, mistakes in the latter should bias the results in one direction. That is, participants who mistakenly identify 5.82 and 4.28 as the correct solution of the matrix in **Figure 1**, will overreport, not underreport. Third, Faravelli et al. reported observing only few errors in pilot sessions not further described (i.e., one out of 16 participants). This is a stronger argument, though the sample size was rather small, and the lack of a detailed description impedes further scrutiny. The finding does appear to be at odds with Kajackaite's (2018) observation that 29 out of 100 participants overreported as a result of honest mistakes (see the quote referenced above), versus only six participants who underreported.

Taken together, there are strong theoretical and empirical reasons to believe that the difference between reported performance and actual number of correctly solved matrices is not exclusively due to dishonesty, yet a critical experiment putting the two hypotheses to the test is lacking. The current study seeks to fill this void in order to empirically establish whether and to what extent the matrix task conflates cheating with honest mistakes. In our experiment, participants were randomly assigned to one of two conditions. In both conditions, participants received a work sheet and a separate answer sheet. They were initially told to hand in the answer sheet and keep the work sheet, but, at the end of the experiment, all participants had to return their work sheet as well, which we could link to their answer sheet via a hidden identifier. Critically, in one condition, we projected the correct solutions to each matrix just before participants filled in their answer sheet, without allowing them the opportunity to edit their work sheet. The question was whether and to what degree this would reduce the difference between reported performance and actual performance, compared to the regular condition where participants had the opportunity to self-correct, though without the solutions being displayed. Crucially, based on the rationale behind the matrix task, providing the solutions should not matter, as participants accurately evaluate their answers anyway (Mazar et al., 2008; we will revisit this assertion in the discussion section).

Method
Participants
A total of 268 participants took part in the experiment (131 male, 134 female, three non-identifiable), which was framed as a live demonstration in an introductory psychology course for undergraduate economy students at KU Leuven (Belgium). As such, sample size was not determined a priori, but based on attendance. The experiment was carried out according to the principles expressed in the Declaration of Helsinki.

Materials
Besides an informed consent form, participants received a work sheet and an answer sheet, stapled together. The two sheets were later separated (see Procedure), but could be linked to one another via an identifier written in invisible ink on both sheets. After the experiment, corresponding sheets were matched using UV/Black light (see Rigdon & D'Esterre, 2015 for a similar approach).

The work sheet contained 20 4-by-3 matrices, and one such matrix with the correct solution already highlighted as an example. Each matrix element was a number between 0 and 10 with two digits after the decimal point (as in **Figure 1**). All 20 matrices were printed on a single page in four orderly columns. The sheet was modeled after the publically available materials of Verschuere et al. (2018), who conducted a large-scale replication study of Mazar et al.'s (2008) first experiment. However, we only included matrices that were solvable, whereas the latter studies comprised 10 unsolvable items (i.e., no set of numbers added up to 10). Furthermore, all of our matrices featured at least one distractor pair, which was defined as a set of numbers that summed to 9.9, 10.1, or 11.0. As such, our materials had three overlapping matrices with Verschuere et al. (2018)/Mazar et al. (2008), including the very first matrix participants saw, assuming they worked from left to right and top to bottom (i.e., the matrix displayed in **Figure 1**).

The answer sheet contained three questions: it asked for the number of correctly solved matrices (open-ended), their student number (open-ended), and whether they had read or heard about this type of study before (yes/no question). Note that the experiment's purpose was not specified, nor was there any mention of dishonesty or cheating. So, it is possible that some participants who responded affirmatively to the latter question, actually mistook the task for a test of mathematical ability.

Procedure
Before testing took place, participants were divided into two groups based on their student number: odd numbers were instructed to come to one auditorium, even numbers were instructed to come to another auditorium at the same time. The procedure for both groups was identical, except for the critical manipulation mentioned at the end of this section.

The experimental materials (see above) were distributed evenly across the entire auditorium before the participants entered. Participants were spread out so they could not simply copy their neighbors' results. An informed consent form was lying face up; the work and answer sheets were stapled together lying face down underneath the informed consent form. When participants were led into the auditorium, they were told to pick a spot, but not to inspect the materials yet. Once everyone was seated, an experimenter collectively went over the content of the informed consent, after which two other experimenters collected the signed forms. Next, the matrix task was introduced. The exact instructions were as follows (see the Materials component on OSF, https://osf.io/hpq9w/):

> In each of the 20 boxes, you can find a set of numbers that sum up exactly to 10. For each box in which you found the set, circle the numbers that add up to 10. 2 students will be selected randomly and will receive €10 for each solution they found. You have 4 minutes to find as many sets as possible.

A timer was projected so that participants could track their progress. After the four minutes were over, participants were instructed to put their pens down, count the number of correct solutions, and tear apart the work and answer sheets. They were told to put away the work sheets, with the cover story that the following class would touch upon it. When that was done, participants could use their pens again to fill in the answer sheet. Then, the experimenters collected all answer sheets, after which participants got the, for them surprising, directive to hand in their work sheet as well. Compliance with all instructions was verified by the experimenters.

Importantly, in one of the auditoria (i.e., the *check* condition), the correct solutions were projected on a screen, after participants had put their pens down. All matrices were shown one by one with the correct set of numbers highlighted. Participants were asked to compare their responses with the solutions, and count how many they had found. In the *regular* condition, participants were given ample time to count how many matrices they solved correctly, but no additional information was provided. Note that participants in both conditions had an equal opportunity to cheat by overclaiming the amount of matrices they had solved correctly.

Data processing and analysis

Participants who indicated having prior knowledge about this type of study ($N = 22$) as well as the ones not answering that question ($N = 13$) were excluded from the analyses. The resulting group sizes were very similar (i.e., $N = 115$ in the regular condition, $N = 118$ in the check condition).

Two researchers independently evaluated the work sheets, assigning a code to each matrix indicating whether it was solved correctly (i.e., the right set of numbers was encircled), left open, solved incorrectly because a distractor pair was encircled, or solved incorrectly for another reason. A third researcher compared the results, and solved any discrepancies. This gave rise to three variables: the number of matrices claimed to be solved as reported by the participants on their answer sheet (henceforth *reported*), the number of correctly solved matrices based on the work sheet (henceforth *correct*), and the number of matrices for which participants had provided an answer by encircling two numbers on their work sheet (henceforth *circled*). If participants make mistakes (i.e., $N_{circled} > N_{correct}$), and correct them, as assumed by Mazar et al. (2008), then $N_{reported} = N_{correct}$, unless they are being dishonest (or miscount the number of correctly solved matrices). Consequently, providing the solutions, like in the check condition, should not matter, as participants should be able to unambiguously evaluate their answers anyway (Mazar et al., 2008). Put differently, one would expect the prevalence of overreporting (i.e., the percentage of participants for which $N_{reported} > N_{correct}$) to be constant across both conditions.[1]

Results

In the regular condition, 40.00% of the participants overreported (i.e., percentage of participants whose $N_{reported}$ was greater than their $N_{correct}$), which is consistent with the average rate reported in the meta-analysis of Gerlach et al. (2019) (i.e., 39% when correcting for publication bias via a trim and fill procedure). In the check condition, this figure dropped to 16.10% ($\chi^2(1, n = 233) = 16.54$, $p < .001$, Cohen's w = .27, $BF_{10} = 614.95$[2]). Crucially, the percentage of participants whose $N_{circled}$ was greater than their $N_{correct}$ did not differ between the conditions (i.e., 47.83% in the regular condition, 50.00% in the check condition, $\chi^2(1, n = 233) = 0.11$, $p = .740$, Cohen's w = .02, $BF_{01} = 5.82$). This is taken to mean that participants in both conditions made mistakes to a similar degree. However, when given the correct solutions, many more participants actually corrected their mistakes.

To further examine this, we subdivided our sample in two groups: participants who made no mistakes on their work sheet (i.e., their $N_{circled}$ was equal to $N_{correct}$, 119 participants in total) versus those who did (i.e., $N_{circled} > N_{correct}$, 114 participants in total). When crossed with condition, we see that the former rarely overreported in general: 13.56% in the check and 8.33% in the regular condition, respectively ($\chi^2(1, n = 119) = 0.83$, $p = .361$, Cohen's w = .08, $BF_{01} = 4.73$). In contrast, participants who did make one or more mistakes on their work sheet, were much more likely to overreport in the regular condition compared to the check condition (i.e., 74.55% and 18.64%, respectively; $\chi^2(1, n = 114) = 35.86$, $p < .001$, Cohen's w = .56, $BF_{10} > 1,000$).

Figure 2 shows the extent to which participants overreported in both conditions. In the regular condition, we observed a substantial number of overreporters, in addition to a few underreporting participants, and a large contingent of (presumably) honest participants. Consistent with results in other studies using the matrix task, most participants who overclaimed did so to a limited extent, typically only one or two matrices. According to Mazar et al. (2008), these participants are cheating, but only a little bit as to not impact their self-concept in a negative way. In the check condition, there were considerably fewer overreporters. It seems that most of the "small-time cheaters" actually made one or two honest mistakes, which got filtered out by providing the correct
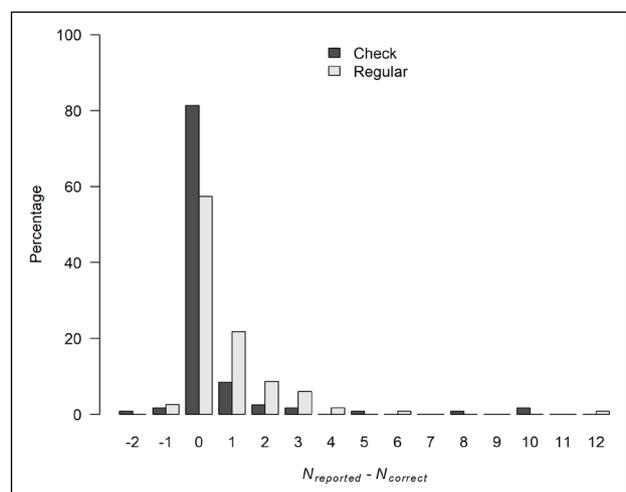
Figure 2: Distribution of $N_{reported} - N_{correct}$ for the regular condition (light grey) and the check condition (dark grey). As can be seen, many of the small lies disappear in the check condition.

solutions. In other words, the degree of overreporting was much more uniformly distributed in the check condition, though we still observed a disproportionate number of participants overclaiming by just one matrix. This pattern is qualitatively in line with Mazar and colleagues' theoretical framework, but the prevalence of overreporting turned out to be much smaller once honest mistakes were taken into account.

Discussion

The matrix task is one of the most popular paradigms to examine dishonesty. It requires participants to report how well they performed on a set of math puzzles. Participants claiming to have solved more items than they actually solved, are considered cheaters, because everyone should be able to unambiguously evaluate whether they made any mistakes, and should engage in such a self-corrective effort. The present study raises serious questions about the validity of this assumption. It shows that participants' tendency to overreport decreases substantially when given the solutions, suggesting that many of the supposedly small lies observed in the regular version of the matrix task, are in fact honest mistakes.

One might object that presenting participants with the solutions made them worry that they could not get away with cheating, or made it harder on them to justify cheating as people wish to maintain a positive self-concept. So, rather than removing honest mistakes, it just reduced the likelihood that participants would cheat. In this respect, it is worth pointing out that participants who made no mistakes (i.e., those with $N_{circled} = N_{correct}$) did overreport at a similar rate, regardless of whether they received the solutions. If anything, they were more likely to overreport in the condition *with* the correct solutions, though the data were more in line with a null effect. Indeed, presenting the solutions only had a robust influence on participants who made one or more mistakes on their work sheet (i.e., those with $N_{circled} > N_{correct}$). This would imply that worrying about getting caught or maintaining a positive self-concept solely manifests itself in the latter group of participants, which seems rather implausible.

Another point to consider is that all 20 matrices used in the current experiment contained distractors whereas the original materials of Mazar et al. (2008) featured only three such matrices (including the very first matrix participants see). Therefore, the number of participants misclassified as cheaters might differ across studies. Thus, the take-home message is that the matrix task is not well-equipped to separate honest mistakes from actual cheating, but we remain somewhat agnostic as to *how many* participants were wrongly considered liars in previous studies. Interestingly, Gerlach et al. (2019) briefly touch upon this issue in their review of the dishonesty literature, noting the following:

> [P]articipants might falsely believe that they found the solution to a matrix although they did not, or they might miscount the total number of "solved" matrices. False reporting in the matrix task should therefore not always be equated with dishonest behavior. (p. 3)

Finally, one might argue that honest mistakes should occur in all conditions, so it is basically a wash. However, this rationale is flawed, because researchers typically compare reported performance with actual performance, either across participants (if there is a cheat and a control condition) or within participants (if their work sheet can be linked to their answer sheet). As we have demonstrated, taking the difference between reported and actual performance conflates cheating with honest mistakes: part of it is due to deliberate deception, but a substantial portion stems from undetected miscalculations. One could mitigate this issue by comparing the reported performance in the control condition with the reported performance in the cheat condition across participants, as did Verschuere et al. (2018). However, this approach is only valid if the rate of honest mistakes is the same in both conditions.

Taken together, all dependent variables derived from the matrix task capture unrelated factors to some degree. Hence, when any given manipulation (e.g., priming with the Ten Commandments) yields an effect, it could be because it targets cheating *or* because it affects the prevalence of honest mistakes (or both). For example, listing the Ten Commandments could prime the idea that someone is watching over one's shoulder, causing participants to double check their calculations, which in turn would reduce (or eliminate) honest mistakes in that condition. The plausibility of the latter statement is irrelevant; it merely serves to illustrate that there are (many) alternative explanations for the same pattern of results, which have nothing to do with dishonesty whatsoever (the next section discusses a more realistic, theory-driven account). As internal validity is of critical importance in laboratory experiments, the matrix task is inherently flawed.

Implications

In light of the extended literature relying on the matrix task, our conclusion does beg the question of why various interventions proposed to foster or discourage ethical behavior showed an effect in previous studies (e.g., priming via the Ten Commandments, as in Mazar et al., 2008). Of course, any given manipulation *might* truly affect the cheating component that does get captured by the matrix task. However, it is impossible to unambiguously establish this, due to the lack of internal validity, unless a conceptual replication using a different paradigm provides converging evidence. Indeed, the coin-flip and die-roll task, two popular alternative paradigms to examine dishonesty, do not suffer from the same issues (Bucciol & Piovesan, 2011; Fischbacher & Föllmi-Heusi, 2013). In these tasks, participants privately flip a coin or roll a die (once or several times), and have to report the outcome. Critically, the payoff scheme is such that participants receive a greater reward for certain outcomes (e.g., heads, or higher dice numbers). One can compare the reported outcomes aggregated across participants against chance levels to gauge the amount of cheating in a certain condition (though it is not possible to determine whether an individual participant cheated). Using such paradigms to establish the effectiveness of honesty-inducing/reducing interventions can alleviate the concerns raised here.

Moreover, certain effects could also turn out to be Type 1 errors. For example, there are serious question marks about the replicability of many so-called behavioral priming effects (Yong, 2012). In fact, the notion that listing the Ten Commandments would reduce cheating, as measured through the matrix task, did not receive any empirical support in a multi-lab, pre-registered replication study (Verschuere et al., 2018).

A third option is that (some) honesty-inducing/reducing interventions actually affect the speed-accuracy trade-off. By design, the matrix task puts participants under pressure, as most won't be able to solve all matrices within the allotted time frame. Consequently, participants might adjust their decision criteria (just as in any other perceptual or cognitive task) meaning that less evidence is required to initiate a response. The latter is a key notion in sequential sampling models of decision making backed by much empirical evidence (see Heitz, 2014, for a review). In particular, lowering the response criteria results in faster, yet more error-prone decisions. So, if a certain manipulation puts relatively more emphasis on speed, one would expect more incorrect responses, thus increasing the number of reported matrices, without necessarily boosting the number of correct responses. In other words, one needs to carefully evaluate the instructions, because subtle cues might affect how participants weigh speed against accuracy.

This rationale does not only apply to honesty-inducing/reducing interventions, though. Indeed, the very manipulation of giving participants the opportunity to cheat might influence how much they value speed versus accuracy. For instance, knowing that their solutions will be thrown away/recycled, might lead participants to believe that accuracy is not that important. Conversely, telling participants that their performance will be evaluated, suggests that accuracy is important, which might prompt them to adjust their response criteria accordingly, or increase the likelihood that they will check their responses retrospectively. Unfortunately, the precise instructions are not always available, so it is difficult to evaluate how many honesty-inducing/reducing manipulations actually involved a straightforward shift in the speed-accuracy trade-off.

One might argue, though, that favoring speed over accuracy should be viewed as cheating in its own right. However, it is important to note that such lowering of response criteria is not necessarily a deliberate decision. Moreover, it would become a slippery slope as to which phenomena should be considered acts of cheating. What about errors in a Stroop task, or wrong answers to trivial questions under the (time) pressure of a game show (e.g., *frog* in response to "name an animal with three letters in its name")? These behaviors are by no means comparable to the illustrious examples often listed to demonstrate the importance of research on dishonesty (e.g., taking prohibited substances to gain an edge in sports, or tax fraud).

More than anything, this discussion exemplifies the need for clear, testable theoretical frameworks. If we assume, for the sake of argument, that the matrix task

is a valid instrument to measure cheating, then there is still the question of why participants are dishonest: for the (potential) monetary gain, or simply due to demand effects? With regard to the latter possibility, Gino et al. (2013b) explicitly mentioned that in the original matrix paradigm "participants might have interpreted the research context as one in which the researcher wanted some participants to misreport, thus providing the plausible impression that the researcher's purpose might have benefitted from misreporting" (p. 2192–2193). Including ten unsolvable matrices, as did Mazar et al. (2008) as well as many follow-up studies, might have added fuel to the idea that participants were encouraged to cheat; after all, the experimenters are themselves being deceptive (or they at least violated the cooperative principle of communication, see Grice, 1975). Similarly, the procedure of solving matrices on one piece of paper, writing down the number of correctly solved items on another piece of paper, and then possibly having to throw away the original work sheet, might be considered bizarre to the extent that participants view it as a solicitation to cheat from the part of the experimenter.

In sum, the present study suggests, both on theoretical and empirical grounds, that the matrix task as a measure of dishonesty lacks internal validity. As such, we recommend researchers to use or create different paradigms, and treat outcomes based on the matrix task with due caution.

Data Accessibility Statement

The manuscript was written in R (R Core Team, 2016) using the packages papaja (Aust & Barth, 2017) and rmarkdown (Allaire et al., 2016). On the project's page (https://osf.io/sepkd/), one can find the .Rmd file, the materials, and the data.

Notes

[1] We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012). A detailed explanation of all variables can be found on the OSF project page (see the Data component, https://osf.io/576jw/).

[2] The statistical tests reported here and throughout this paper are a Pearson's chi-squared test, and a Bayesian alternative, both evaluating the independence assumption. For the latter, we used the contingencyTableBF function from the BayesFactor package (Morey & Rouder, 2015) with default priors. The abbreviation BF, short for Bayes factor, has a subscript 10 indicating the relative plausibility of the data under the alternative hypothesis versus under the null hypothesis assuming independence. So, in this case, we have strong reasons to believe that the underlying probabilities are different across the two conditions (presuming that both hypotheses are equally likely a priori). Throughout the text, we will use the subscripts 10 or 01 to indicate which hypothesis is preferred over the other, as to facilitate the interpretation of the results. In addition, we also report Cohen's w as a measure of effect size. Note that

a sensitivity power analysis for the chi-square test with the current sample size, an alpha of .05 and a power of .80 yields a Cohen's w of .18.

Competing Interests
The authors have no competing interests to declare.

Author Contributions
All authors developed the design of the study. HV processed the data. TH analyzed the data and drafted the manuscript. All authors provided critical revisions and approved the final version for submission.

References
Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., ... Hyndman, R. (2016). *rmarkdown: Dynamic Documents for R*. Retrieved from https://CRAN.R-project.org/package=rmarkdown

Ariely, D. (2012). *The (honest) truth about dishonesty: How we lie to everyone − Especially ourselves*. New York, NY: Harper Collins Publishers.

Aust, F., & **Barth, M.** (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Ayal, S., & **Gino, F.** (2011). Honest rationales for dishonest behavior. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 149–166). Washington, DC: American Psychological Association. DOI: https://doi.org/10.1037/13091-008

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76*(2), 169–217. DOI: https://doi.org/10.1086/259394

Bucciol, A., & **Piovesan, M.** (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology, 32*(1), 73–78. DOI: https://doi.org/10.1016/j.joep.2010.12.001

Faravelli, M., Friesen, L., & **Gangadharan, L.** (2015). Selection, tournaments, and dishonesty. *Journal of Economic Behavior & Organization, 110*, 160–175. DOI: https://doi.org/10.1016/j.jebo.2014.10.019

Fischbacher, U., & **Föllmi-Heusi, F.** (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association, 11*(3), 525–547. DOI: https://doi.org/10.1111/jeea.12014

Gerlach, P., Teodorescu, K., & **Hertwig, R.** (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin, 145*(1), 1–44. DOI: https://doi.org/10.1037/bul0000174

Gino, F., Ayal, S., & **Ariely, D.** (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science, 20*(3), 393–398. DOI: https://doi.org/10.1111/j.1467-9280.2009.02306.x

Gino, F., Ayal, S., & **Ariely, D.** (2013a). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization, 93*, 285–292. DOI: https://doi.org/10.1016/j.jebo.2013.04.005

Gino, F., Krupka, E. L., & **Weber, R. A.** (2013b). License to cheat: Voluntary regulation and ethical behavior. *Management Science, 59*(10), 2187–2203. DOI: https://doi.org/10.1287/mnsc.1120.1699

Gino, F., Schweitzer, M. E., Mead, N. L., & **Ariely, D.** (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes, 115*(2), 191–203. DOI: https://doi.org/10.1016/j.obhdp.2011.03.001

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3: Speech acts, pp. 41–58). New York, NY: Academic Press.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*, 150. DOI: https://doi.org/10.3389/fnins.2014.00150

Kajackaite, A. (2018). Lying about luck versus lying about performance. *Journal of Economic Behavior & Organization, 153*, 194–199. DOI: https://doi.org/10.1016/j.jebo.2018.07.010

Klein, E., Moeller, K., Dressel, K., Domahs, F., Wood, G., Willmes, K., & **Nuerk, H.-C.** (2010). To carry or not to carry—Is this the question? Disentangling the carry effect in multi-digit addition. *Acta Psychologica, 135*(1), 67–76. DOI: https://doi.org/10.1016/j.actpsy.2010.06.002

Kouchaki, M., & **Gino, F.** (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences, 113*(22), 6166–6171. DOI: https://doi.org/10.1073/pnas.1523586113

Mazar, N., Amir, O., & **Ariely, D.** (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633–644. DOI: https://doi.org/10.1509/jmkr.45.6.633

Morey, R. D., & **Rouder, J. N.** (2015). *BayesFactor: Computation of Bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rigdon, M. L., & **D'Esterre, A. P.** (2015). The effects of competition on the nature of cheating behavior.

Southern Economic Journal, *81*(4), 1012–1024. DOI: https://doi.org/10.4284/0038-4038-2012.301

Shu, L. L., Gino, F., & **Bazerman, M. H.** (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, *37*(3), 330–349. DOI: https://doi.org/10.1177/0146167211398138

Shu, L. L., Mazar, N., Gino, F., Ariely, D., & **Bazerman, M. H.** (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences*, *109*(38), 15197–15200. DOI: https://doi.org/10.1073/pnas.1209746109

Simmons, J. P., Nelson, L. D., & **Simonsohn, U.** (2012). A 21 word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social*

Psychology, *26*(2), 4–7. DOI: https://doi.org/10.2139/ssrn.2160588

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., ... Yıldız, E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317. DOI: https://doi.org/10.1177/2515245918781032

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature*, *490*, 7418. DOI: https://doi.org/10.1038/nature.2012.11535

Zhong, C.-B., Bohns, V. K., & **Gino, F.** (2010). Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. *Psychological Science*, *21*(3), 311–314. DOI: https://doi.org/10.1177/0956797609360754