Exp Econ
DOI 10.1007/s10683-016-9479-y

ORIGINAL PAPER

Clever enough to tell the truth

Bradley J. Ruffle[1] · **Yossef Tobol**[2,3]

Received: 4 July 2014 / Revised: 29 February 2016 / Accepted: 2 March 2016

Abstract We conduct a field experiment on 427 Israeli soldiers who each rolled a six-sided die in private and reported the outcome. For every point reported, the soldier received an additional half-hour early release from the army base on Thursday afternoon. We find that the higher a soldier's military entrance score, the more honest he is on average. We replicate this finding on a sample of 156 civilians paid in cash for their die reports. Furthermore, the civilian experiments reveal that two measures of cognitive ability predict honesty, whereas general self-report honesty questions and a consistency check among them are of no value. We provide a rationale for the relationship between cognitive ability and honesty and discuss its generalizability.

Keywords Honesty · Cognitive ability · Soldiers · High non-monetary stakes

JEL Codes C93 · M51

Electronic supplementary material The online version of this article (doi:10.1007/s10683-016-9479-y) contains supplementary material, which is available to authorized users.

✉ Bradley J. Ruffle
bradleyruffle@gmail.com

[1] Department of Economics, Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada

[2] School of Management, Jerusalem College of Technology, Jerusalem 91160, Israel

[3] IZA, Bonn, Germany

Published online: 09 March 2016

1 Introduction

Is it possible to screen effectively for honesty? This is the million-dollar question for employers interviewing job candidates, investors vetting corporate conference calls, tax auditors, voters watching a political candidates' debate, potential business partners and courting couples, to name a few examples. Continuing with the workplace example, the U.S. retail industry alone loses $53.6 billion a year to employee theft. Moreover, employee theft is on the rise due to poor pre-employment screening and a decline in supervision (Brooks and Chad 2013). Can employers do better in screening for honest employees? Polygraph tests have been shown to be unreliable and their use by employers is unlawful in North America and Europe. Neuroscience-based lie-detection technologies remain unproven and in their infancy. Consequently, many employers continue to rely on written personality tests consisting of self-report questions and job interviews led by the company's human resource personnel to evaluate job candidates' honesty. However, a considerable body of research casts doubt on the usefulness of these methods (see, e.g., Morgeson et al. 2007 for a survey and Ones et al. 2007 for a rejoinder).

In this paper, we use the methods of a laboratory experiment on both soldier and civilian populations in the field to determine the effectiveness of the compulsory military entrance exam employed by the Israeli Defense Forces (IDF) to categorize soldiers on the basis of their cognitive ability and honesty, among other traits. We extend Fischbacher and Föllmi-Heusi's (2013) innovative die-rolling paradigm to the field where 427 soldiers and 156 civilians each rolls a six-sided die in private and reports the outcome. For each additional pip reported on the die, soldiers are rewarded with a half-hour early release from the army base on Thursday and civilians with extra cash payment. We find soldiers with higher exam scores are more honest, that is, report lower die outcomes. On a civilian population that has completed its mandatory military service we again find that a higher entrance exam score predicts a lower die report. The robustness of the relationship between exam score and honesty to a civilian sample eliminates a set of strategic and reputational hypotheses for our result and points to the content of the entrance exam as an invaluable predictor of honesty. However, pinpointing which features of the exam account for honesty is complicated by the diverse number of items that determine a soldier's score. Our experiments on civilians reveal that self-report honesty questions and a consistency check among them cannot explain subjects' variation in reported die outcomes.[1] Instead, higher scores on two cognitive-ability tests (namely, the cognitive reflection task (CRT) (Frederick 2005) and an abbreviated version of the Raven advanced progressive matrices test (Arthur and Day 1994)) predict increased honesty.

In the developmental psychology literature, a wealth of correlational studies dating back to the late 1920s explores the observed relationship between intelligence and socially approved or disapproved behaviors. Among the first

[1] Interestingly, a large literature in personnel psychology debates the usefulness of personality tests as an aid in hiring decisions, job placement and worker evaluation. Most recently, Morgeson et al. (2007) review over 7000 manuscripts on the usefulness of these tests and conclude that they have low predictability of job performance and that alternatives to these self-report measures should be sought. Our findings support this conclusion.

studies, Hartshorne et al. (1928) collect observational and self-report data from over 10,000 children on three forms of deceptive behavior (cheating, lying and stealing) and attempt to correlate it with numerous traits including intelligence. They provide as their "best estimate of the relation between intelligence and a theoretical combination of all of our deception tests, a correlation of −.50 to −.60" (p. 189) (quoted from Unger 1964, p. 300). In summarizing several reviews of this literature, Unger writes that "brains and character actually tend to go together", the existence of a positive relation between intellect and morality "is practically unanimous and unequivocal", "the most helpful and cooperative children [a]re nearly always also among the brightest" and "sixth-grade boys rated to be low in delinquency potential [are found] to have markedly superior IQ scores". A number of studies have found that cheating on exams is most frequent among students with low grades or low GPA scores (see, e.g., Hetherington and Feldman 1964 and the references therein). Gottfredson and Hirschi (1990) review the cross-cultural determinants of crime and conclude that "the individual-level correlates of delinquency that appear everywhere include sexual precocity, limited scholastic aptitude and drug use" (p. 178).

Methodologically closer to our study (i.e., incentivized experiments on dishonest reporting), two recent studies include measures of intelligence as a moderator variable in explaining the relationship between honesty and the variable of interest. Gino and Ariely (2012) find that creativity predicts dishonest reporting across a series of five studies. In one of the studies, the authors also measure intelligence through the CRT and a vocabulary test (verbal intelligence) and show that neither intelligence measure is significantly correlated with dishonesty. Fosgaard et al. (2013) also control for CRT scores in an experiment designed to separate the effect that cheating is an option (i.e., cheating awareness) from the effect that cheating is the norm (i.e., cheating conformity). The authors find that CRT scores are positively associated with the probability of cheating. After finding that *kaba* scores predict honesty in our soldier experiments, our study shifts to focus on the relationship between intelligence and honesty.

Our study differs from a rapidly expanding experimental literature on honesty (see Rosenbaum et al. (2014) for a survey) in two further respects. First, our subject pool: unlike student subject pools or even most field experiments targeted at a particular population, soldiers completing their mandatory military service constitute a representative cross-section of society as a whole.[2] What is more, this population is particularly well suited to examine the relevance of screening criteria for honesty because military units are highly susceptible to dishonesty. The

[2] Abeler et al. (2014) conduct an experiment on honesty by telephone on a representative sample of German respondents. Other experiments conducted on soldiers are: Goette et al. (2012) who compare the in-group cooperativeness and willingness to punish of extant groups of Swiss soldiers with those of randomly formed groups of soldiers. Lahav et al. (2011) distribute questionnaires on trains traveling between major Israeli cities to soldiers, teenagers and university students and show that soldiers have higher subjective discount rates than non-soldiers. In a companion paper based on the same soldier experiments, Ruffle and Tobol (2014) show that temporally distancing decisions from the receipt of payment increases honest reporting. Specifically, soldiers who participated in the die-rolling experiment on earlier days of the week reported low outcomes on average than those who participated closer to the end of the week. Soldiers' military entrance scores served merely as a control variable in the analysis and was not explored in any depth. In the current paper, we focus on the relationship between military entrance scores and honesty and test the robustness of our findings on a civilian population.

hierarchical organizational structure inherent in a nation's military inevitably means that a commanding officer who assigns a duty or issues an order to a soldier has no opportunity to verify whether the soldier has completed the assigned task. Yet any well-functioning military relies on honesty between its troops and even seeks this trait when recruiting and promoting soldiers. Indeed, honesty is among the highest declared values for Israeli soldiers and part of the creed of the Israel Defense Forces.

Another source of novelty of our experiment is that, to the best of our knowledge, it is the first to examine honesty toward one's employer. Soldier subjects in our experiment cheat their boss with whom they interact on a daily basis, rather than an anonymous firm (e.g., Levitt 2006; Pruckner and Sausgruber 2013), anonymous subjects (e.g., Gneezy 2005), wait staff at a restaurant (Azar et al. 2013), or the experimenter (e.g., Fischbacher and Föllmi-Heusi 2013).[3]

2 Experiment design, procedures and sample

We approached all of the combat and non-combat military units for which we might possibly receive permission—the elite air force units were omitted. Our sample consists of 15 distinct army units that granted us permission to conduct the experiment. These units were distributed across 27 different permanent and provisional military bases throughout Israel.[4] All 427 soldiers in our sample were serving in their first of three years of required military service during which they participate in basic training through completing courses. The sample of courses represents the broad spectrum of preparatory courses. For non-combat soldiers, these courses include machinery, supplies, programming, diving, armaments, meteorology, intelligence, medics, cooks, while for combat soldiers the courses include infantry, engineering corps, armored corps, artillery corps, and the navy. The experiments were conducted between December 28, 2010 and June 19, 2011, a period of relative quiet in Israel as it was not involved in any wars or military confrontations.

All of the experiments were conducted just prior to the soldiers' breakfast hour in the dining hall. The cadet coordinator (CC) of the participating army unit called each soldier by name one-at-a-time to a room or large tent with two entrances/exits located on the army base and used for the purpose of the experiment. Each participating soldier entered through one designated entrance. The CC then read the rules of the experiment from a script to the soldier. Namely, the soldier was told that s/he would be asked to roll a 6-sided die in private and then to report the outcome to the CC. For each point on the die, the soldier would be released on Thursday half an

[3] To a lesser extent they also cheat their colleagues because a soldier who leaves the army base early necessitates that his uncompleted duties are distributed among those soldiers who remain behind.

[4] With the exception of our purposeful oversampling of religious companies, we view our sample of soldiers as representative of the overall population of Israeli soldiers. In fact, in Sect. 3.2 we will see that the distribution of military entrance scores of soldiers in our sample mirrors the overall distribution. What is more, because military service is mandatory for all Israelis (except for the Arab-speaking and ultra-orthodox Jewish populations for whom it is optional), our sample constitutes a representative cross-section of society as a whole.

hour ahead of the scheduled time. To avoid any possible confusion, the exact payment in the form of hours of early release for each of the six possible outcomes was enumerated. The soldier was told that, after all soldiers in the unit had completed the experiment, the CC would submit the list of early release times to the unit commander who had approved the experiment and the terms of early release.

The soldier was handed a 6-sided die and proceeded to a table at the other side of the room or tent, out of sight of the CC. After rolling the die in private, the soldier returned to the CC to report the outcome. Finally, the soldier completed a brief post-experiment questionnaire (included in Online Appendix A), submitted it to the CC and was directed to proceed to the dining hall through the door or tent opening designated as the exit and through which he had not entered. The distinction between the two doors or tent openings as entrance and exit was maintained to prevent soldiers from having contact with others who had not yet participated in the experiment. The CC called in the next soldier according to the list and so on until all soldiers in the unit had completed the experiment.

The entire experiment including the questionnaire took about seven minutes for each soldier. In view of the value soldiers attribute to an early release of half an hour (median = 30 NIS, see rows 2 and 3 of the left panel of Table 1) and of three hours (median = 100 NIS, row 4 of Table 1), the experimental payment can be deemed salient.[5]

3 Results

3.1 Overall distribution

The distribution of reported die outcomes for our entire sample of soldiers (N = 427) is displayed in Fig. 1. If all soldiers reported the truth, we would expect a uniform distribution of reported die outcomes. This hypothesis is soundly rejected (Pearson Chi square test χ^2 (5) = 16.2, p = .001).[6] Soldiers clearly inflate their reported outcomes, but do not profit maximize.[7] What is more striking is the observed decline in frequency from 5 to 6. While Fischbacher and Föllmi-Heusi

[5] Consider the following back-of-the-envelope calculation. The average soldier reported a die outcome of 3.87 (see row 1 of left panel of Table 1), equivalent to 1.94 h early release. If we assume, for simplicity, that the median willingness to pay increases linearly with each additional half hour of early release, then the median willingness to pay for 1.94 hours equals 58.1 NIS for seven minutes of work. Contrast this with combat soldiers' monthly wage of 700 NIS and non-combat soldiers' monthly salary of between 300 and 500 NIS, depending on their job. At the time of the experiments, 3.5 NIS equaled $1 USD.

[6] Further evidence against uniformly distributed die outcomes comes from the frequencies of reported 1 and 2 s, both significantly less than the percentage of 16.67 % expected from a uniform distribution (p < .001 from one-sided Binomial tests in both cases). At the same time, the frequencies of 4 and 5 s, are significantly greater than 16.67 % (p = .04 and p < .001, respectively). Only the frequencies of reported 3 s and 6 s cannot be rejected as significantly different from 16.67 % (p = .13 and p = .38, respectively).

[7] Incomplete cheating appears to be a robust finding in the emerging literature on cheating regardless of whether the die-rolling paradigm (e.g., Shalvi et al. 2011; Fischbacher and Föllmi-Heusi 2013; Hao et al. 2013) or some other experimental method is used (e.g., Gneezy 2005; Charness and Dufwenberg 2006; Erat and Gneezy 2012).

Table 1 Descriptive statistics for soldier and civilian samples

Variable (possible values)	Soldiers Mean (SD)	Civilians Mean (SD)
Reported die outcome	3.87	3.95
(1,2,3,4,5,6)	(1.61)	(1.58)
WTP for half hour early	42.7, 30	–
(in Israeli NIS)	(67.2)	
WTP for half hour early—outliers excluded	33.9, 30	–
(in Israeli NIS)	(28.1)	
WTP for 3 h early[§]	194.1, 100	–
(in Israeli NIS)	(225.5)	
Military entrance test score (kaba)	51.1	51.5
(45–56)	(2.5)	(2.8)
Claim know test score[§]	.839	.776
(0,1)	(.369)	(.419)
Actual know test score[§]	.406	.526
(0,1)	(.492)	(.501)
Female	.424	.481
(0,1)	(.495)	(.501)
Religious	.337	.244
(0,1)	(.473)	(.431)
City resident	.724	.872
(0,1)	(.458)	(.335)
Self-reported honesty	2.12	1.30
(1 = always tell truth − 4 = truth when convenient)	(.99)	(.95)
Others think	3.68[§]	3.52
(1 = very important − 7 = not important at all)	(1.96)	(1.71)

The die outcome refers to the roll of the die reported; the WTP variables (soldiers only) indicate the willingness to pay for a half-hour and a three-hour early release from the base (median values are displayed to the right of the means); the military test score outcomes in our sample range from 45 to 56; the indicator variables "*claim know test score*", "*actual know test score*", "*female*", "*religious*" and "*city resident*" indicate the fraction of participants that claims to know their army test score, the fraction that knows their true test score; the fractions of female soldiers, religious soldiers and soldiers from cities, respectively; participants also answered questions on the extent to which they generally tell the truth ("*self-reported honesty*") and how important it is to them what others think of them ("*others think*"). See Online Appendix A for the precise wording of the questions

[§] Indicates that the question appeared for the last 217 soldiers only. All other statistics are based on the full sample of 427 soldiers and 156 civilians

(2013) also report incomplete cheating, they still register a higher percentage of subjects who report the highest outcome, whereas we witness a sharp decline. One explanation for our observed decline in reported 6 s is that payments are publicly observable. A soldier seen leaving the base three hours early on Thursday may be concerned that his peers will view him as dishonest.

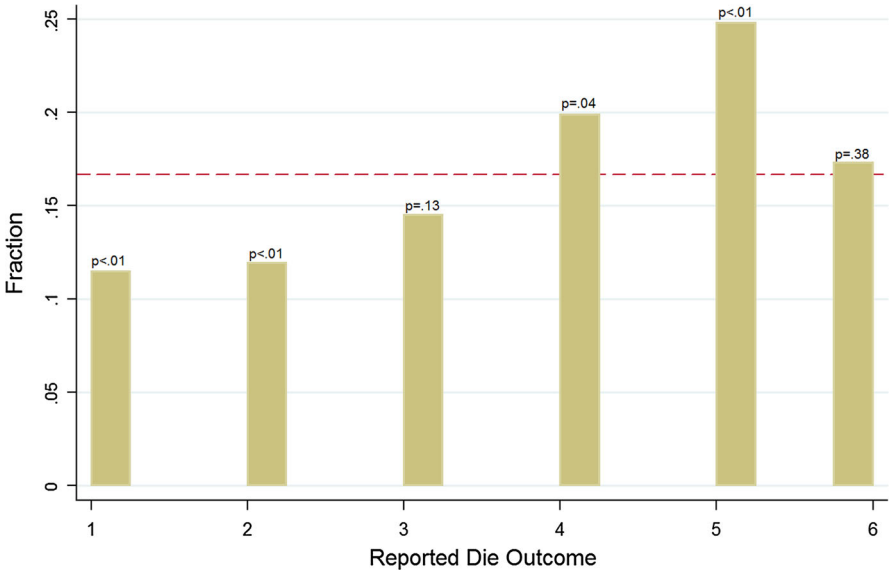

Fig. 1 Histogram of soldiers' reported die outcomes. *Note*: p value from one-sided binomial test that observed frequency of each die outcome is less (greater) than .166 appears above each *bar*

3.2 Honesty and military entrance score

Months prior to recruitment to the Israeli military, every candidate soldier is evaluated on the basis of his or her educational background and a series of computerized psycho-technical exams. In addition to cognitive-ability measures, numerous questions on the psycho-technical exams are designed to evaluate the soldier's honesty by, for instance, framing the same question in multiple ways or otherwise asking as many ten variations of the same question to test for consistent responses. Furthermore, males undergo a lengthy personal interview in which the candidate's honesty is evaluated through several channels. First, highly skilled female interviewers aim to assess the male soldier's "body language, to identify lies and individuals who are unreliable" (Hebrew Wikipedia under "recruitment to the Israeli military").[8] What is more, candidate soldiers are asked a battery of questions, the answers to which are either already known (e.g., "Have you been caught stealing or otherwise been in trouble with the police?"), can be verified by contacting the appropriate authority (e.g., "Have you ever skipped school?", "What would your high school teachers say about you?") or can be cross-checked with the candidate's responses on the written exam (e.g., "Have you ever used drugs?", "Are you seeking the ideal partner or willing to compromise?").

[8] Interestingly, Daniel Kahneman developed in large part the structured interview protocol, which remains largely intact to this day (Kahneman and Daniel 2002).

For male soldiers, the final test score (known as *kaba* in Hebrew and to be subsequently referred to as such for brevity) is made up of the interview (33 %), the psycho-technical exams (50 %) and the candidate's educational background (e.g., school attended, absences, any recorded discipline issues) and achievements (e.g., grades, clubs, distinctions) (17 %). Women do not undergo the interview. Instead, their *kaba* is based on the psycho-technical exams (60 %) and their educational background and achievements (40 %).

An individual's *kaba* determines the unit and job to which he is assigned for his military service. The *kaba* scores range from 41 to 56. Special significance is accorded to 52, the required cutoff to qualify for an officer's course that is offered only after non-combat soldiers have completed a full year of mandatory military service and combat soldiers have completed two years of service.[9] All of the soldiers in our sample were serving in their first year of military service during which the military training and experiences of those who are eligible and those who are ineligible to become officers are identical. It is noteworthy that a *kaba* of 54 constitutes the minimum score to qualify to be an interviewer, with most having attained a rare perfect score of 56. What is more, the criteria for selecting interviewers are considered to be more stringent than any other unit in the IDF (Lerer and Zeev 2009, pp. 20–32).

Through each unit's commanding officer and unbeknown to the soldier, we obtained every participating soldier's *kaba*. Figure 2 displays the distribution separately for females and males. The most unusual feature of these distributions is the paucity of near-miss scores of 50 and 51, accounting for a combined 21–22 % of the sample. At the same time, the highest frequency is associated with 52 (the threshold to qualify for an officer), followed by 49, which together constitute about 30 % of the sample.[10] These features suggest that the Israeli military wishes to enlarge the pool of eligible officers. To do so, it uses the subjective components of the *kaba* score, namely, the personal interview in the case of males and the evaluation of the candidate's educational background, to boost to 52 the scores of soldiers deemed suitable.

The distribution of entrance scores reveals that 48.2 % of soldiers in our sample qualify to be officers. Interestingly, Lerer and Zeev (2009) reports an identical figure (48 %) for the fraction of soldiers with a *kaba* of 52 or higher in 1995 (the most recent year for which he obtained data).[11]

The weighted scatter plot in Fig. 3 displays the distribution of reported die outcomes for each *kaba* score. The size of the circle reflects the number of observations at this *kaba*-die outcome pair. The figure also includes two regression

[9] In fact, the *kaba* exam is only the first of several screening devices used to determine eligibility to become an officer. Only at the end of the first year of military service are additional selection criteria applied to those eligible soldiers with *kaba* scores of 52 or more, such as the recommendations of commanding officers, a personal interview with the soldiers' commanding officers and a sergeant's course.

[10] Neither the rank-sum Wilcoxon-Mann–Whitney test nor the Kolmogorov–Smirnov test rejects the equality of the female and male distributions of military entrance scores ($p = .20$ and $p = .66$, respectively).

[11] The Israel Defense Forces do not make publicly available the distribution of military entrance scores.

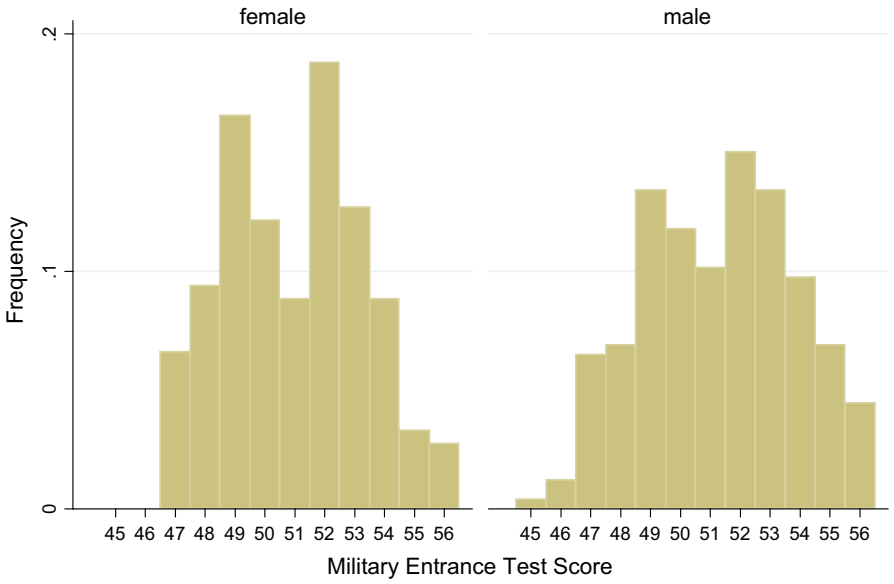

Fig. 2 Distribution of Kaba test scores by sex

curves fitted to the data. The OLS line displays a negative slope, in other words, the higher a soldier's *kaba*, the lower his reported die outcome on average. So as not to impose a linear fit on the data, we also estimate a kernel regression (shown in red in Fig. 3). This curve too is negatively sloped up to and including a *kaba* of 54, after which a modest upturn is observed.[12]

The OLS regression in (1) of Table 2 pinpoints the magnitude and statistical significance of the negative relationship: for every additional point a soldier obtained on his *kaba*, he reports .13 points less on the die outcome ($p < .01$).[13] Regression (2) demonstrates that this highly significant negative relationship between a soldier's reported die outcome and his *kaba* remains robust to the inclusion of numerous controls. These controls include the soldier's self-reported willingness to pay for one half-hour early release from the military base, indicator variables for whether the soldier is female, religious, and an interaction term between these two, and from a city, and the soldier's response to a question about

[12] Warner and Pleeter (2001) also observe unique behavior among the two highest categories of entrance exam scores in the U.S. military. In particular, they exploit a natural experiment conducted by the U.S. Department of Defense to reduce military personnel in which mid-career personnel were offered the choice between a lump-sum separation payment and an annuity valued at considerably more in present terms. Personnel belonging to the top groups display lower rates of discount (i.e., more patience) than their peers, as evidenced by their higher likelihood of preferring an annuity to a lump-sum retirement payment.

[13] The regressor is expressed as soldier *i*'s kaba minus 52 for ease of interpretation. Thus, the constant of 3.76 refers to the average die outcome reported by a soldier with a *kaba* of 52.

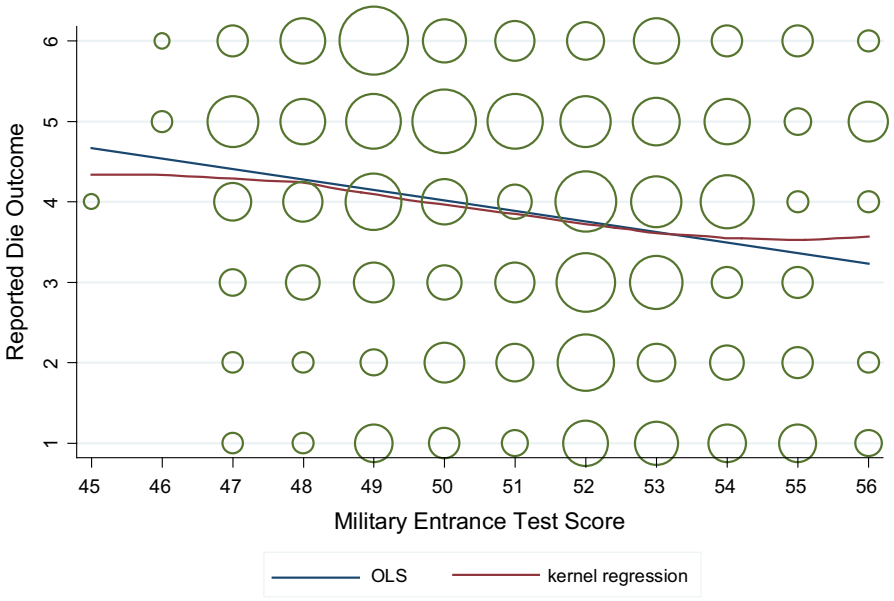

Fig. 3 Histogram of soldiers' reported die outcomes by test score

the extent to which he generally tells the truth.[14] The majority of the controls do not differ significantly from zero in this or any subsequent regression. One exception is that secular female soldiers report weakly significantly higher die outcomes than secular males in this regression only.[15] The mean die outcomes reported by males and females do not differ significantly from one another in any of the other regressions. Soldiers from cities (defined by the Israeli Central Bureau of Statistics as settlements with more than 20,000 residents) report significantly higher die outcomes than those from rural areas in this and all subsequent regressions.

A soldier's willingness to pay for early release is not quite marginally statistically significant in predicting a soldier's decision to claim additional early release ($p = .12$). According to Table 1, the mean willingness to pay for a half-hour early release is 42.7 NIS. A standard deviation of 67.2 that exceeds the mean by more than 50 % and a median of 30 NIS both attest to outliers. If we exclude observations that deviate from the mean by more than two standard deviations, the median remains unchanged, while the mean and standard deviation drop to 33.9 NIS and 28.1 NIS respectively ($N = 412$). Regression (3) excludes these 15 outlying

[14] Also included but not shown are measures of the soldier's military unit and military base peer effects (neither of these measures is significantly different from zero in any of the regressions), as well as indicators for the day of the week on which the experiment was conducted with Sunday as the omitted day (all of the other days of the week are positive and significantly different from zero in all regressions and are discussed in detail, along with the peer effects variables, in Ruffle and Tobol 2014).

[15] Contrast this with Dreber and Johannesson (2008) who find that men are more likely than women to send deceptive, self-serving messages to their partner in a sender-receiver game modeled after Gneezy (2005).

Table 2 Regression analysis on soldiers' reported die outcomes

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	3.76***	1.92**	1.91**	3.78***	1.64	3.67***	2.24
	(.08)	(.93)	(.98)	(.14)	(1.51)	(.13)	(1.47)
Δ *kaba from 52*	−.13***	−.10***	−.10***	–	–	−.18***	−.14**
	(.03)	(.03)	(.03)			(.05)	(.06)
Δ *kaba guess from 52*	–	–	–	.09	.06	–	–
				(.08)	(.09)		
WTP for half hour early release	–	.001	.007***	–	.006	–	.006
		(.001)	(.003)		(.005)		(.005)
Female	–	.33*	.24	–	.39	–	.36
		(.18)	(.18)		(.31)		(.31)
Religious	–	−.21	−.15	–	.50	–	.49
		(.29)	(.28)		(.63)		(.62)
Religious female	–	−.12	−.15	–	−.66	–	−.66
		(.36)	(.36)		(.70)		(.68)
City resident	–	.37**	.30*	–	.68**	–	.62**
		(.17)	(.17)		(.29)		(.30)
Self-reported honesty	–	.06	.08	–	.29*	–	.24
		(.08)	(.08)		(.15)		(.15)
Others think	–	–	–	–	.08	–	.08
					(.07)		(.07)
Adj. R^2	.04	.08	.10	.00	.12	.05	.14
Includes day-of-week, peer-effects controls	No	Yes	Yes	No	Yes	No	Yes
Excludes *WTP* Outliers	No	No	Yes	No	Yes	No	Yes
N	427	427	412	185	174	185	174

Dependent variable: soldier *i* 's reported die outcome

Regressors: Δ *kaba from 52* is soldier *i*'s military test score (*kaba*) minus 52; Δ *kaba guess from 52* is soldier *i*'s guess about his kaba minus 52; the soldier's willingness-to-pay for half an hour early release from the base; indicator variables for whether the soldier is female, religious, and an interaction term between the two, from a city (or a rural area); his self-reported honesty (question 4 in Online Appendix A), the importance he attributes to what others think of him (question 5 in Online Appendix A), and measures of soldier *i*'s military unit and military base peer effects calculated as the mean reported die outcome of all members of soldier *i*'s unit and base, respectively, excluding soldier *i*

Heteroskedasticity-robust standard errors in parentheses

Regressions (2), (3), (5) and (7) include indicator variables for the day of the week on which soldier *i* participated in the experiment and measures of military unit and military base peer effects

Regressions (3), (5) and (7) exclude observations more than two standard deviations above the mean "*WTP for half hour early release*"

Coefficient significantly different from 0 at the 1 % level ***, at the 5 % level **, at the 10 % level *

observations, but is otherwise identical to (2). The coefficient on the willingness-to-pay variable increases sevenfold to .007 and is highly significant ($p = .01$). At the same time, the *kaba* coefficient of −.11 in (3) remains highly significant.

Figure 4 displays histograms of die reports separately for soldiers with a *kaba* score of 52 or more and those below 52. The figure distinctly displays that soldiers with a *kaba* score below 52 report higher outcomes on average (mean die report of 4.21, median and mode of 5) than those with a *kaba* above 52 (mean report of 3.5, median and mode of 4). A *t* test soundly rejects the equality of the mean reports for these two groups of soldiers ($t = 4.66$, $p < .001$).

Lower die reports do not necessarily imply increased honesty. It could be that soldiers with a high *kaba* score underreport 5 s and 6 s to such an extent that they are actually less honest than their counterparts with a low *kaba* scores. This turns out not to be the case. The left panel of the Fig. 4 shows clearly that soldiers below 52 over-report on average; the reported frequencies of 1, 2 and 3 s are all grossly underrepresented, whereas the frequencies of 5 s and 6 s are significantly greater than 1/6 (all $p < .05$). Only the frequency of reported 4 s does not significantly differ from 1/6 ($p = .20$). The result is that the Pearson Chi square test rejects that the die reports of the low *kaba* soldiers are uniformly distributed (χ^2 (5) = 24.19, $p < .001$). By contrast, the right panel shows that soldiers with a *kaba* greater than or equal to 52 neither unambiguously under-report nor over-report on average. In fact, the Pearson Chi square test cannot reject the null hypothesis that their reported die outcomes are uniformly distributed (χ^2 (5) = 3.24, $p = .66$). None of the observed frequencies of reported 1, 2, 3, 4 and 5 s is significantly different from 1/6. Only the reports of 6 s (25/206 or 12.1 % of the observations) differ significantly from 1/6 ($p = .045$).

Whether we compare the distributions of reported die outcomes for soldiers above and below a *kaba* of 52, or first invert the distribution of outcomes for those with a *kaba* 52 or more such that the transformed distribution, Y, equals 7 minus the outcome from the original distribution, both the non-parametric Kolmogorov–Smirnov test (based on the maximum difference between the two cumulative distribution functions) and non-parametric rank-sum Wilcoxon-Mann–Whitney strongly reject the equivalence of the outcomes reported by soldiers below and soldiers above the threshold of 52 ($p < .001$ for all four tests).[16] We conclude that lower die reports among soldiers with higher *kaba* scores indeed reflect significantly more honesty.

There are three possible explanations for the positive association between a soldier's honesty and his entrance score. First, with its emphasis on the value of honesty among its soldiers, the Israeli military has designed a test that successfully predicts honesty. Second, soldiers behave reciprocally toward the military: soldiers know their entrance scores and use our experiment to reward or to punish the military for a high or low score received, respectively. Third, reputational concerns: a soldier whose high *kaba* score qualifies him to be an officer may be considering a

[16] To see why we need to invest the reports for high *kaba* soldiers, suppose we wish to determine whether a group that systematically under-reports is more or less honest than a group that systematically over-reports. To render the two distributions of die reports comparable, we first need to invert one of them before performing the appropriate statistical test. Because soldiers with a *kaba* of 52 or more neither unambiguously under-report nor unambiguously over-report, we compare both their original and their inverted die-report distributions with that of soldiers with a *kaba* below 52. Both methods lead to the same conclusion: soldiers with high *kaba* scores are more honest.

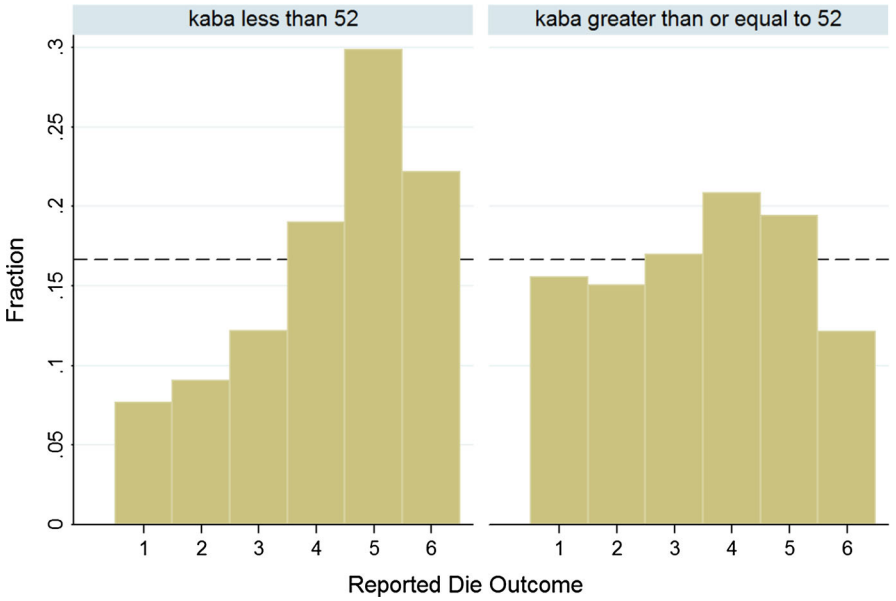

Fig. 4 Histogram of soldiers' reported die outcomes by Kaba score

career in the military and does not want to jeopardize his future by being perceived as greedy or dishonest by his fellow soldiers or commanding officer.[17]

In the next section, we present a follow-up die-rolling experiment on civilians who have completed their military service. This follow-up experiment will allow us to: (1) determine whether the predictive power of the *kaba* is robust to an alternative sample population and setting; (2) control for the reciprocity and reputation explanations; and (3) evaluate which, if any, of the various components of *kaba* predict honesty. In the meantime, our dataset from the experiment on soldiers allows us to further explore the reciprocity explanation from above.

To test whether reciprocity can explain our main result, we introduced two questions midway through the data collection process: 217 soldiers were asked whether they knew their entrance score and, if so, what it was (see questions 8 and 9 of Online Appendix A). In our sample, 15 % of soldiers admitted to not knowing their entrance score. Among those who claimed to know their score, only 47 % indicated the correct score. Of the 53 % of the soldiers who incorrectly guessed their score, the vast majority (92/97) overestimated it. The average guess among

[17] Fischbacher and Föllmi-Heusi (2013) also report the results of a double-anonymous version of their die-rolling experiment in which a subject's reported die outcome is unknown to other subjects and to the experimenter. They find little difference in the distribution of reported outcomes across anonymity conditions. Mazar et al. (2008) report a similarly negligible difference in the number of matrices subjects claim to have solved when anonymity vis-à-vis the experimenter is added. Reputational concerns may nonetheless be more important in our setting in which the payment is different and a subject's reported die outcome is observable by both his commanding officer and fellow soldiers with whom he interacts on a daily basis.

soldiers who mistakenly guessed their score was 1.98 points higher than the actual score (SD = .14).[18]

If soldiers use our experiment to express their appreciation for a high entrance score and the enhanced opportunities that such a score furnishes or to express their dissatisfaction with their low score, then we would expect a significant negative relationship between a soldier's guess about his score and his reported die outcome. Regression (4) replaces a soldier's true entrance score with his guess. The estimated coefficient is actually positive (.09), but not significantly different from zero (p = .24). The coefficient remains close to and not significantly different from zero when our set of control variables is included in (5). By contrast, the parallel regressions in (6) and (7) that use the same sample of 185 soldiers, replacing their guess with their true score, reveal highly significant coefficients of −.18 (p = .001) and −.13 (p = .05), respectively. In short, reciprocity cannot explain the observed relationship between honesty and soldiers' test scores.

4 Follow-up experiment on civilians

4.1 Motivation, experimental design and procedures

What specifically accounts for the seeming effectiveness of the Israeli military's entrance exam in classifying soldiers in terms of their degree of honesty? The calculation of the *kaba* is sufficiently complex that the specific components of the test that are most effective remain shrouded in mystery. Moreover, we have yet to test the hypothesis that reputational concerns account for our observed relationship between a soldier's reported die outcome and his *kaba* score. With these questions in mind, we replicated our experiment on a sample of civilian subjects who had completed their military service. Between March 8, 2015 and May 4, 2015, 156 subjects took part in our experiment. The payment of early release from the military base was replaced with a monetary payment. In addition to a 20 NIS show-up payment, participants receive 10 NIS for every pip they report. Thus, payments range from 30 NIS to 80 NIS ($8 to $21 USD) for an experiment that took between 18 and 22 min to complete.

In an effort to avoid the reputational concerns possibly present in a faculty-led experiment conducted on students and to recruit a more diverse subject pool, we bypassed the university, choosing instead to conduct the experiments at the Malha Shopping Mall in Jerusalem. In collaboration with the shopping mall's management, we placed a reception desk and a table with cardboard partitions (the type used for voting in Israeli elections) in the mall's concourse for the purpose of recruiting and conducting the experiment. Before permitting a candidate subject to

[18] Our observation that as many as 40 % (.47 × (1 − .15)) of soldiers correctly guessed their *kaba* and those who guessed incorrectly were off by "only" 2 points on average are not surprising. Before entering the military, every recruit provides a preference ordering over military units in which he wishes to serve. Since different units require different *kaba* thresholds, a recruit's acceptance to or rejection from his preferred unit(s) provides him with an update about the possible range of his *kaba*.

participate in the experiment, we asked whether the subject had completed mandatory military service (in order that we can later obtain the subject's *kaba* score). Upon submitting the completed questionnaire the subject received payment.

Because we required completion of military service to participate in our experiment, this civilian subject pool is necessarily older (mean age = 25, minimum age = 20, SD = 2.6 years) than our sample of 18-year-old soldiers. Moreover, 52 % are male, 90 % completed high school, 47 % hold a college or university degree, 67 % report having at least one parent with a college or university degree, 72.4 % are born in Israel, and 81.8 % are single, 16.9 % married and 1.3 % divorced. The right panel of Table 1 provides additional summary statistics for the civilian sample.

4.2 Experimental results

4.2.1 Overall distribution

Figure 5 displays the distribution of reported die outcomes for the civilian sample (N = 156). Similar to the corresponding distribution for soldiers discussed in Sect. 3.1 and displayed in Fig. 1, a Pearson Chi square test rejects that all subjects report their true die outcome; namely, the empirical distribution is significantly different from a uniform distribution (χ^2 (5) = 19.8, p = .001). Moreover, one-sided Binomial tests show that the frequencies of reported 1 s and 2 s are significantly less than 16.67 %, while the frequencies of reported 4, 5 and 6 s are significantly greater than 16.67 % (all p < .001). Only the frequency of reported 3 s

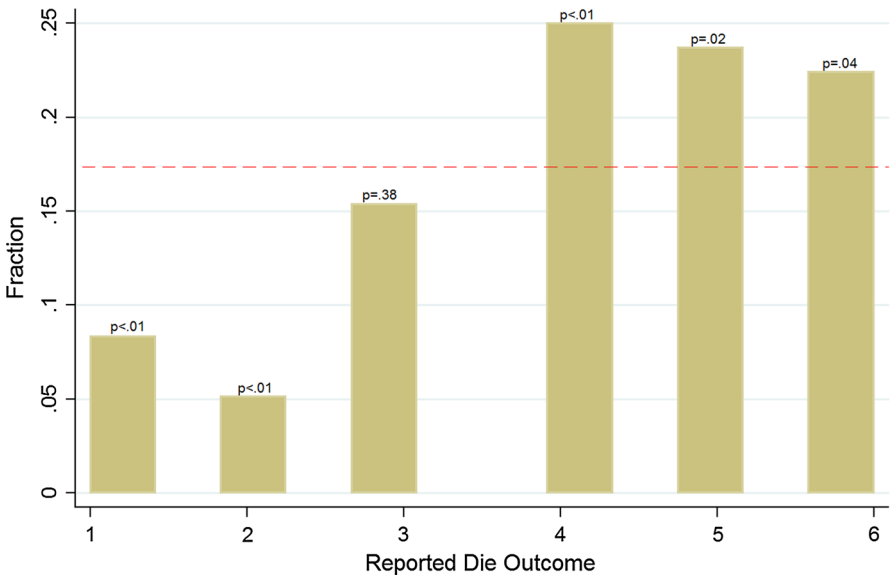

Fig. 5 Histogram of civilians' reported die outcomes. *Note*: p value from one-sided binomial test that observed frequency of each die outcome is less (greater) than .166 appears above each *bar*

cannot be rejected as significantly different from 16.67 % ($p = .70$). Accounting for 25 % of reported die outcomes, 4 is the modal outcome, with the frequency of reports declining modestly for 5 and again for 6. Thus, even with the lack of an ongoing relationship with the experimenter and unobservability by their peers, the overwhelming majority of subjects still refrain from maximizing their monetary earnings.

Whether the soldiers' and civilians' distributions of reported outcomes differ significantly from one another depends on the choice of statistical test. The Kolmogorov–Smirnov test fails to reject the equality of distributions ($p = .18$), whereas the Wilcoxon–Mann–Whitney test rejects their equality ($p = .05$). To the extent that there is a difference, it appears to be driven largely by the relative abundance of soldiers willing to report 1 and 2 s. In fact, if we exclude die reports of 1 and 2 and perform the same Wilcoxon–Mann–Whitney test on the truncated distributions, we no longer come close to rejecting the equality of soldiers' and civilians' distributions ($p = .81$).

4.2.2 Robustness of results on Civilian sample with monetary payoffs

The weighted scatter plot in Fig. 6 displays the distribution of civilians' reported die outcomes for each *kaba* score. The OLS and kernel regressions displayed in this figure as well as the regressions in Table 3 reveal that the observed significant and negative relationship between a soldier's *kaba* and his reported die outcome continues to hold on this sample of civilians who receive cash payments for their die

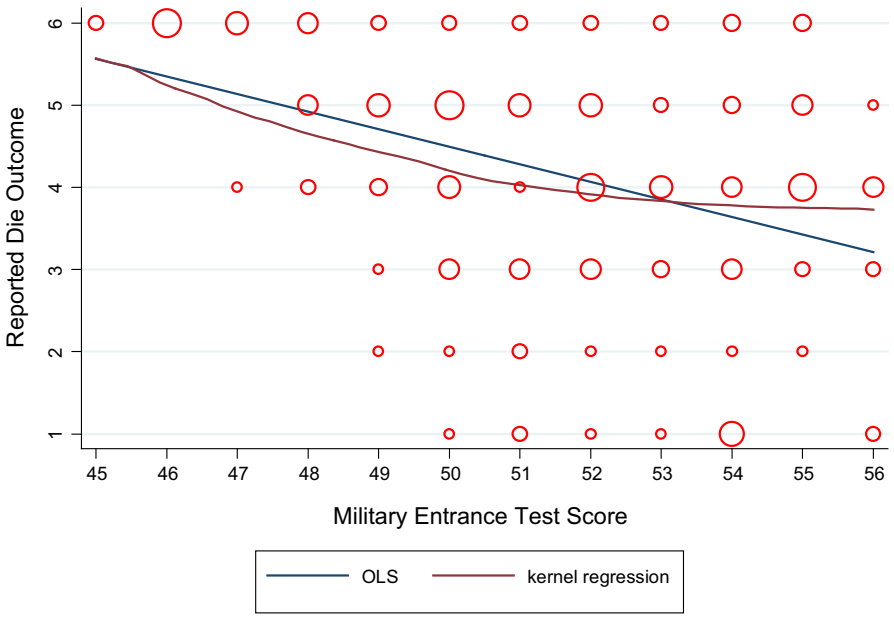

Fig. 6 Histogram of civilians' reported die outcomes by test score

Table 3 Regression analysis on civilians' reported die outcomes

Variable	(8)	(9)	(10)
Constant	4.07***	4.86***	5.94***
	(.12)	(.45)	(.87)
Δ *kaba from 52*	−.21***	−.20***	−.23***
	(.03)	(.04)	(.05)
Female	–	.25	.20
		(.25)	(.27)
Religious	–	−.63*	−.51
		(.34)	(.40)
Religious female	–	.99	−.79
		(.60)	(.67)
City resident	–	−.36	−.45
		(.30)	(.31)
Self-reported honesty	–	−.19	−.12
		(.12)	(.12)
Others think	–	−.08	−.09
		(.07)	(.08)
High-school graduate	–	–	−.81*
			(.44)
Includes controls from civilian questionnaire	No	No	Yes
Adj. R^2	.16	.21	.17
N	156	155	143

Dependent variable: civilian *i* 's reported die outcome

Regressors: Δ *kaba from 52* is civilian *i*'s military test score (*kaba*) minus 52; Δ *kaba guess from 52* is soldier *i*'s guess about his kaba minus 52; indicator variables for whether the participant is female, religious, and an interaction term between the two, from a city (or a rural area); his self-reported honesty (question 4 in Online Appendix B), the importance he attributes to what others think of him (question 5 in Online Appendix B) and whether he finished high school. The controls included, but not displayed, in regression (10) are the participant's self-reported high-school matriculation grade, whether he holds a college or university degree, whether at least one of his parents does, whether he was born in Israel, his rating of the neighborhood in which he grew up, whether he works, his income and a self-reported measure of how hardworking he is

Heteroskedasticity-robust standard errors in parentheses

Coefficient significantly different from 0 at the 1 % level ***, at the 5 % level **, at the 10 % level *

reports. More precisely, for every additional point a civilian subject obtained on his *kaba*, regression (8) shows that he reports about 0.2 points less on the die ($p < .001$). This holds regardless of the collection of control variables that we include alongside the subject's *kaba* score. Regression (9) contains the same set of control variables as those reported on the soldier sample in Table 2.[19] The coefficient on Δ *kaba from 52* remains unchanged at .2 and highly significant

[19] The military variables "WTP for half-hour early release" and military unit peer effects are of course absent from the civilian sample as are the day-of-the-week indicators since all civilian participants received payment immediately after participating and not on Thursday afternoon like the soldier sample.

($p < .001$). At the same time, none of the socio-demographic controls, the self-report responses about one's own honesty and about the importance of what others think is significantly different from zero. The lone exception is that religious male subjects report .63 points less on the die than their secular male counterparts ($p = .07$). Recall that in the soldier sample, the only consistently significant socio-demographic or self-report variable was whether the participant was a city resident: city residents claimed more. The sign on the city-resident indicator variable flips to negative among civilians, but is not significantly different from zero ($p = .20$). It is thus noteworthy that the *kaba* score stands out as the only significant predictor of the participant's reported die outcome in both the soldier and civilian samples.

In addition to the controls present in regression (9), regression (10) includes questions asked only in the civilian survey, several of which were summarized in Sect. 4.1. The estimate on the Δ *kaba from 52* variable remains highly significant ($p < .001$) and its magnitude increases slightly to .23. None of the other variables that were also estimated in (9), including the civilian's religiosity, is significantly different from zero. Among the ten variables unique to regression (10), none is significant with the exception of whether the subject completed high school. As displayed in (10), high school graduates report .81 pips less on the die than those who did not finish high school ($p = .07$).[20] This result foretells one of the central findings from the next subsection and main lessons from this follow-up experiment.

In Sect. 3.2 we presented evidence inconsistent with the hypothesis that reciprocity accounts for the observed positive relationship between the *kaba* score and honesty. That the *kaba* score continues to predict subjects' reported die outcomes on a civilian subject pool drives another nail in the coffin of the reciprocity hypothesis and provides strong evidence against the officer reputation hypothesis. As far as subjects were concerned, this experiment had nothing to do with the military or *kaba* scores. Moreover, the payment was in cash. Yet, strikingly, years after taking the military entrance exam and completing their military service, these civilian subjects' scores continue to forecast their degree of honesty in this die-rolling experiment.

4.2.3 Decomposing the military entrance score

What remains is to understand which *kaba* items predict honesty in our experiment. The self-report honesty question that appears in both the soldier survey (question 4 of Online Appendix A) and civilian survey (question 31 of Online Appendix B) is never significantly different from zero in any of the regression specifications. In the civilian survey, we included additional self-report honesty measures (questions 39 and 44) as well as a disguised variation of question 31 (question 46) to determine whether respondents' consistency in answering questions 31 and 46 can explain their die report.[21] Also, in an attempt to mimic the psycho-technical nature of the

[20] The significance and lack thereof of each of these variables is robust to whichever subset of regressors is included in the specification.

[21] In question 31, we ask "Which of the following sentences best describes you?" with "a. I always tell the truth," "b. I almost always tell the truth," "c. I usually tell the truth," and "d. I tell the truth when it is convenient for me" as the possible responses. Question 46 reads, "Do you speak the truth in your daily

kaba exam, the civilian questionnaire includes: (1) the 12-item abbreviated version of the Raven Advanced Progressive Matrices Test developed by Arthur and Day (1994) and based on the original 60-item test (Raven and John 1936) for the purpose of measuring cognitive ability (questions 19–30); (2) the 3-item cognitive reflection task (abbreviated as CRT hereafter) developed by Frederick (2005) to test subjects' ability to overcome their impulse to write the intuitive and incorrect response in order to think through the problem and arrive at the correct answer (see questions 33, 38, and 44).[22]

The first two columns in Tables 4 and 5 report the distributions of scores for the Raven test and CRT, respectively. The mean score is 10.36 (SD = 1.60) on the Raven test and 1.33 (SD = 1.04) on the CRT. Frederick (2005) reports a mean CRT score of 1.24 based on a sample of 3,428 university students and non-students. The Raven and CRT scores are highly correlated in our sample (Kendall's $\tau = .46$, $p < .001$). The right-hand columns in Tables 4 and 5 display the mean and full distributions of reported die outcomes for each Raven and CRT test score, respectively. Raven test scores of 5 through 9 have been combined so that all cells contain at least 25 observations. Take notice of the pronounced shift from reporting a 6 for low Raven and CRT scores to reporting a 1, 2 or 3 as these test scores increase. To illustrate, 47.5 % of subjects who score between a 5 and a 9 on the Raven test claim to have rolled a 6. This percentage plunges to 21.4 % for Raven scores of 10, and drops further to about 11 % for the highest Raven scores of 11 and 12. At the same time, the combined outcomes of 1, 2 and 3 account for a mere 10 % for Raven scores between 5 and 9, more than doubling to 21.5 % for Raven scores of 10, and again nearly doubling to 37–42 % for scores of 11 and 12.

This shift in outcomes from a 6 to a 1, 2 or 3 as subjects' cognitive abilities increase is no less dramatic for the CRT scores. While 41.5 % of those who answered no questions correctly on the CRT reported a 6, this percentage progressively shrivels to 18.4 % for a score of 1, to 17.5 % for a score of 2 and to 7.7 % for a perfect score of 3. Concurrently, only 14.7 % of subjects who scored a 0 claim a 1, 2 or 3, compared to 22.4 % for those who score a 1, 35 % for a score of 2 and 53.9 % for a score of 3.

Footnote 21 continued
life?" with the set of answers, "a. always," "b. generally," "c. sometimes," and "d. when I stand to gain from it." The absence of a one-to-one correspondence between the two sets of responses requires us to be liberal in our definition of consistency and minimizes the likelihood of a type-1 error in incorrectly inferring that a subject is lying or inconsistent in responding to the two questions. While 31a corresponds perfectly to 46a, 31b may be consistent with either 46a or 46b; 31c matches 46b or 46c, and 31d may correspond to 46c or 46d. Even with this charitable definition of consistency, we still find that 18 % of subjects unambiguously contradict themselves in responding to the two questions with 57 % of the inconsistent choices being 31a ("I always tell the truth") and 46c ("sometimes").

[22] Numerous studies in economics demonstrate that higher cognitive ability predicts a number of desirable traits and outcomes, such as lower risk aversion and more patient time preferences (see, e.g., Frederick 2005 and the references therein as well as Burks et al. 2009, Dohmen et al. 2010, Oechssler et al. 2009). Oechssler et al. (2009) also show that subjects with high CRT scores are less prone than their low CRT-score peers to both the conjunction fallacy and to conservatism in probability updating, while the two groups are equally susceptible to anchoring.

Table 4 Distribution of Raven test scores and reported die outcomes

Correct Answers	Subjects (%)	Distribution of Reported Die Outcomes						Mean Outcome
		1	2	3	4	5	6	
5	1 (0.6%)	.050	.025	.025	.175	.250	.475	4.98 (1.35)
6	4 (2.6%)							
7	4 (2.6%)							
8	13 (8.3%)							
9	18 (11.5%)							
10	28 (18.0%)	0	.036	.179	.250	.321	.214	4.50 (1.14)
11	43 (27.6%)	.116	.093	.209	.303	.163	.116	3.65 (1.48)
12	45 (28.9%)	.133	.044	.200	.267	.244	.111	3.78 (1.51)
Mean	10.36 (1.60)	.083	.051	.154	.250	.237	.224	4.18 (1.49)

Notes: The distribution of scores from abbreviated Raven advanced progressive matrices test, frequencies and mean (SD) of reported die outcomes for each Raven test score

Table 5 Distribution of CRT scores and reported die outcomes

Correct answers	Subjects (%)	Distribution of reported die outcomes						Mean outcome
		1	2	3	4	5	6	
0	41 (26.3 %)	0	.049	.098	.146	.293	.415	4.93 (1.19)
1	49 (31.4 %)	.061	.041	.122	.286	.306	.184	4.29 (1.35)
2	40 (25.6 %)	.100	.075	.175	.350	.125	.175	3.85 (1.49)
3	26 (16.7 %)	.231	.385	.269	.192	.192	.077	3.31 (1.62)
Mean	1.33 (1.04)	.083	.051	.154	.250	.237	.224	4.18 (1.49)

Notes: The distribution of scores from the cognitive reflection task (CRT), frequencies and mean (SD) of reported die outcomes and for each CRT score

Figure 7 displays a weighted scatter plot of the civilians' reported die outcomes for each Raven test score along with fitted OLS and kernel regressions. Figure 8 is an analogous weighted scatter plot of die outcomes for each CRT score. Both figures display a marked downward tendency in reported die outcomes as the number of correctly answered questions on the Raven test and CRT increases.

The regressions in Table 6 quantify these relationships and confirm the strong predictive power of the Raven and CRT test scores. Regression (11) shows that for each additional correct answer on the Raven test, a subject reports .28 pips less on the die outcome ($p < .001$). Similarly, for every CRT question answered correctly subjects claim .29 fewer pips on the die ($p = .047$). Regression (12), on the other hand, includes as regressors only the self-report honesty questions and a response-consistency check among two of these questions. The signs of these five estimates vary from positive to negative and none is different from zero at conventional

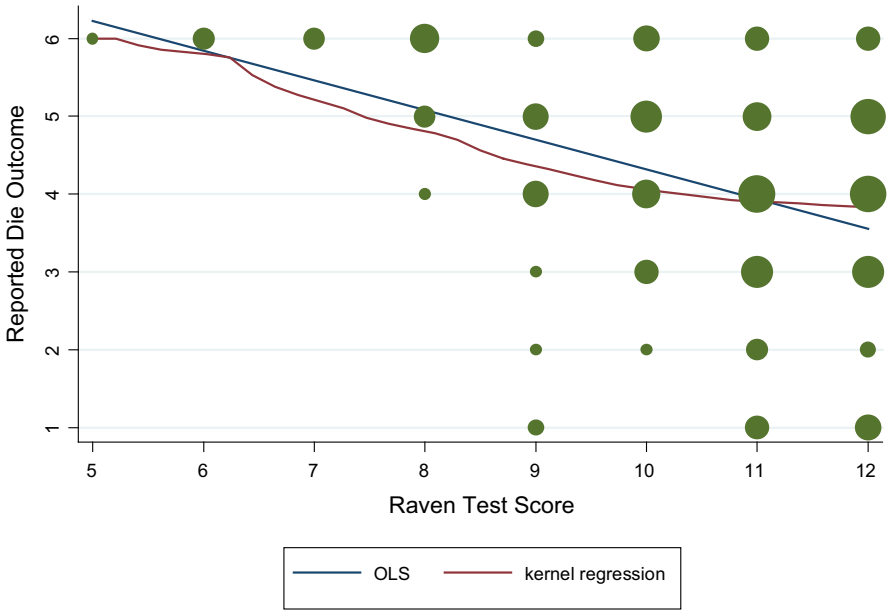

Fig. 7 Histogram of civilians' reported die outcomes by Raven test score

significance levels.[23] Finally, regression (13) includes the full set of controls that were also included (both those displayed and not displayed) in regression (10) of Table 3. Among these are the civilian's self-reported income, employment status and whether he holds a university degree and at least one of his parents does, none of which comes close to be significantly different from zero, thereby ruling out the hypothesis that the role of cognitive ability is rooted in an individual's income or income-earning potential. Moreover, none of the self-report honesty measures differs significantly from zero—p values range from .27 to .79—whereas the estimates and the significance of the Raven and CRT scores remain virtually unchanged—although the CRT estimate of −.29 is now only weakly significantly different from 0 ($p = .064$).[24]

The implication of these findings is that the *kaba* score predicts honesty in our experiments because it incorporates cognitive ability. Yet it could be that the *kaba*

[23] Responses to the four self-report honesty questions are coded (as in Online Appendix B) such that higher values correspond to less honesty. The coefficient of −.29 on the self-report honesty variable has a *p* value of .11; but its negative sign implies that the *less* truthful a subject claims to be, the lower his reported die outcome.

[24] In a within-subject online experiment, Hugh-Jones and David (2015) finds that honest behavior is positively correlated in a coin-flip and a quiz experiment. Yet, a self-report honesty question about whether lying in one's self-interest is justifiable fails to predict behavior in either honesty experiment. At the same time, Hugh-Jones also includes self-report questions about whether the respondent had engaged in any one of four ethically questionable actions in the past 12 months (e.g., avoid fare on public transport, fabricate information on a job application). Reports of unethical actions do predict dishonesty in both experiments. These findings suggest that questions about actual participation in specific forms of dishonest behavior may be better predictors of dishonesty in incentivized experiments than general self-report questions about honesty.

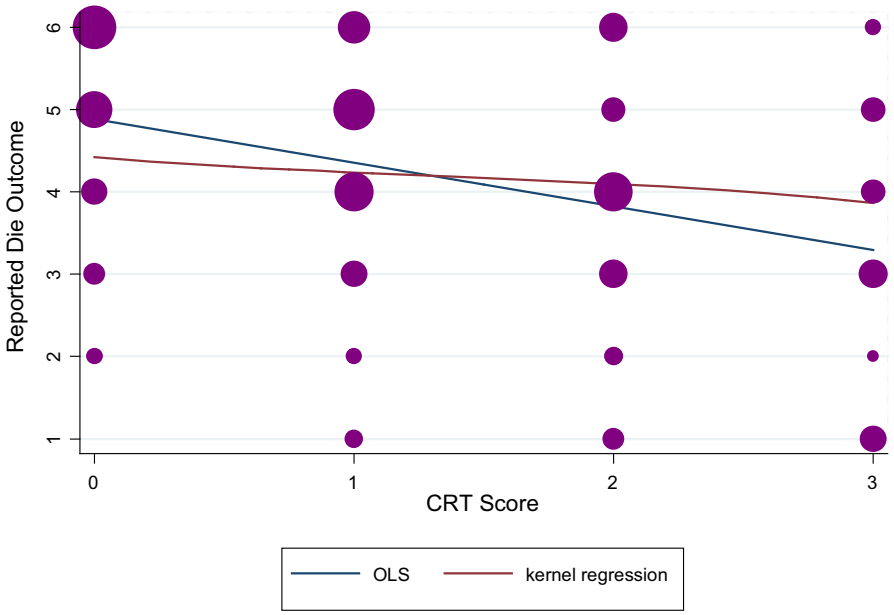

Fig. 8 Histogram of civilians' reported die outcomes by CRT score

score and our two measures of cognitive ability independently explain variation in reported die outcomes. The Pearson correlation coefficient of .80 between the civilian's *kaba* and Raven scores and a similarly high correlation of .66 between the *kaba* and CRT scores indicate that the *kaba* score's effectiveness indeed operates through its testing of cognitive ability. By contrast, the absolute value of the Pearson correlation coefficient between the *kaba* score and any of the five self-report honesty measures never exceeds .09.

One plausible, almost tautological, explanation for the observed relationship between CRT scores and honesty is that subjects who overcome their impulse to write down the intuitive, but incorrect, answers in the CRT in order to think their way through to the correct answers are the same subjects who contemplate the costs and benefits of different die reports and resist reporting a 6.[25] Similarly, subjects who succeed in solving the progressively trickier questions on the Raven test also think more about their choice in the die-reporting experiment and come to the realization that the monetary-payoff-maximizing report of 6 is not necessarily utility maximizing.

Why might someone who (over)thinks his decision in the die-reporting experiment choose not to inflate his report or claim a 6? One possibility is that such subjects suspect that the true outcome of their die roll may be observable by others. Perhaps a hidden camera records their roll. Second and even more

[25] Somewhat relatedly, numerous social psychology studies show that high self-control and the ability to overcome impulses are associated with higher grades, better relationships and interpersonal skills (see, for e.g., Tangney et al. 2004 and the references therein). Similarly, Shalvi et al. (2012) show that dishonesty increases when subjects face time pressure in the form of insufficient time to fully contemplate their reporting decision.

Table 6 *kaba* items that predict civilians' reported die outcomes

Variable	(11)	(12)	(13)
Constant	7.43***	4.25***	5.94***
	(.66)	(.62)	(.87)
Raven score	−.28***	–	−.27***
	(.08)		(.09)
CRT score	−.29**	–	−.29*
	(.15)		(.16)
Consistent honesty responses	–	−.05	.12
		(.43)	(.42)
Self-reported honesty (Q31)	–	−.29	−.16
		(.18)	(.19)
Friends' honesty (Q39)	–	.28	.24
		(.21)	(.21)
Honesty importance (Q44)	–	.00	.08
		(.08)	(.09)
Honesty in life (Q46)	–	.09	−.08
		(.30)	(.32)
Includes controls from civilian questionnaire	No	No	Yes
Adj. R^2	.19	.00	.19
N	156	153	140

Dependent variable: civilian *i* 's reported die outcome

Regressors: civilian *i*'s total scores on the Raven advanced progressive matrices test and CRT; indicator variable for whether civilian *i*'s responses to two self-reported honesty questions (31 and 46) are consistent; responses to four self-reported honesty questions (corresponding question numbers appear in parentheses). The controls included, but not displayed, in regression (13) are the same as those that appear in regression (10) as well as those listed in Note 2 of Table 3

Heteroskedasticity-robust standard errors in parentheses

Coefficient significantly different from 0 at the 1 % level ***, at the 5 % level **, at the 10 % level *

plausibly, subjects who score high on the cognitive tests recognize that inflating their report to the maximum possible outcome of 6 forces them to confront the reality that they are cheaters. These subjects decide that preserving their positive self-image is worth more than the additional shekels to be gained from succumbing to the impulse to cheat.[26] Whichever of these two overthinking explanations is correct, both point to subjects' concern with their image, whether it be in the eyes of others or self.[27]

[26] The forward-looking orientation implicit in this concern for one's future self-image is consistent with the observed link found in the literature between CRT scores and more patient time preferences (see the references in footnote 23) and with the theory posited by Gottfredson and Hirschi (1990) that the primary cause of deviancy is low self-control, namely, the tendency of individuals to pursue short-term gratification without consideration of the long-term consequences of their acts.

[27] This explanation raises the question whether the same relationship between cognitive ability and honesty would continue to hold in a high-stakes experiment in which the benefit to cheating is considerably higher.

A second possible explanation is that high-cognitive ability individuals are faster at both learning and applying the social norm of honesty. More specifically, when placed in a novel situation, an individual needs first to interpret the situation before deciding how to respond. High cognitive-ability subjects are better able to identify the die-rolling experiment as a test of their honesty and thus opt to obey the social norm.[28]

5 Conclusions

The Israel Defense Forces requires every candidate soldier to take an entrance exam that measures cognitive ability and honesty, among other traits. We find that a soldier's exam score predicts his degree of honesty in an incentivized experiment in which soldiers receive early release from the military base in accordance with their reported outcome from a privately observed die roll. Moreover, in a parallel experiment on civilians in which financial incentives are substituted, this same relationship between exam score and honesty continues to hold. Closer inspection of the components of the entrance exam reveals that self-report measures of one's overall honesty and a consistency check among them are of no value in explaining subjects' honesty. Rather, two distinct measures of cognitive ability both predict the variation in subjects' honesty. The nature of the Fischbacher and Föllmi-Heusi die-rolling task proffers a sensible explanation for this result. Namely, the benefits from cheating are clear and salient (e.g., earlier release from the base, an extra 40 NIS in payment), while the costs are not immediately apparent and may even be altogether uncertain (e.g., detection, suspicion of having cheated, eroded (self-)image). We propose that higher cognitive-ability subjects think through these costs and, as a result, resist the temptation to inflate their die reports.

The question then arises of whether this result generalizes to other tests of honesty. We argue below that it might, with two caveats. In most truth-telling dilemmas, the benefit to lying is tangible and immediate (e.g., impress another, avoid reprimand or punishment) and, by definition, serves as the very motivation for not telling the truth. At the same time, the cost of lying is uncertain (e.g., probability of detection, can I live with myself?) and likely borne in the (possibly distant) future (e.g., after detection, following years of soul-searching). As a result, while everyone is aware of the gain from telling a lie in a given situation, higher cognitive-ability types may be more aware or better able to assess the expected costs. One caveat concerns the assumption that the benefit to lying is always obvious: if the costs of lying are salient and transparent to all, while the gains from dishonesty are hidden, perhaps those of higher cognitive ability are more likely to recognize and pursue cheating opportunities. A second caveat concerns access to cheating opportunities: to the extent that higher cognitive-ability individuals attain greater success in their careers, they will have greater access to resources and power and, accordingly, more frequent occasions and larger temptations to cheat.

[28] Simon (1990) provides a theoretical rationale for the evolutionary success of social norms such as honesty based on docility and an inability to distinguish between socially prescribed behaviors that contribute to group fitness from those that reduce individual fitness.

This discussion hints at several juicy directions for future research. If higher-cognitive types prove to be no more honest in other truth-telling tasks, then the salience of the gains from cheating and obscurity of the expected costs in the Fischbacher and Föllmi-Heusi die-rolling experiment may be a special case. Such a finding would call into question the representativeness of the die-rolling task as a gauge of honesty.

Acknowledgments We thank Johannes Abeler, Yuval Arbel, Ofer Azar, Ronen Bar-El, Bram Cadsby, Danny Cohen-Zada, Leif Danziger, Nadja Dwenger, Naomi Feldman, Lan Guo, Shachar Kariv, Jonathan Mamujee, Mattia Pavoni, Chet Robie, Tata Pyatigorsky-Ruffle, Jonathan Schulz, Ze'ev Shtudiner, Justin Smith, Fei Song, Michal Kolodner-Tobol, Ro'i Zultan, an editor of this journal, David Cooper, two anonymous referees and numerous seminar participants for helpful comments. We also are grateful to Capt. Sivan Levi and Meytal Sasson for research assistance, Capt. Itamar Cohen for facilitating the soldier experiments and all of the commanding officers for granting us access to their units. A preliminary version of this paper circulated under the title, "Screening for Honesty".

References

Abeler, J., Becker, A., & Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics, 113*, 96–104.

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven advanced progressive matrices test. *Educational and Psychological Measurement, 54*, 394–403.

Azar, O. H., Yosef, S., & Bar-Eli, M. (2013). Do customers return excessive change in a restaurant? A field experiment on dishonesty. *Journal of Economic Behavior & Organization, 93*, 219–226.

Brooks, C. (2013). Employee theft on the rise and expected to get worse. *Business News Daily*, June 19, 2013, Retrieved from, http://www.businessnewsdaily.com/4657-employee-theft-rising.html.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences, 106*(19), 7745–7750.

Charness, G., & Dufwenberg, M. (2006). Promises and partnerships. *Econometrica, 74*(6), 1579–1601.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review, 100*(3), 1238–1260.

Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters, 99*(1), 197–199.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science, 58*(4), 723–733.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association, 11*(3), 525–547.

Fosgaard, T. R., Hansen, J. G., & Piovesan, M. (2013). Separating will from grace: An experiment on conformity and awareness in cheating. *Journal of Economic Behavior & Organization, 93*, 279–284.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Gino, F., & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology, 102*(3), 445–459.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95*(1), 384–394.

Goette, L., Huffman, D., & Meier, S. (2012). The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics, 4*(1), 101–115.

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime*. Stanford: Stanford University Press.

Hao, L. & Houser, D. (2013). Perceptions, intentions, and cheating. Unpublished manuscript.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character, vol 1: Studies in deceit*. New York: Macmillan.

Hugh-Jones, D. (2015). Way to measure honesty: A new experiment and two questionnaires. Unpublished manuscript.

Kahneman, D. (2002). The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2002 Daniel Kahneman, Vernon L. Smith. Retrieved from, http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahneman-bio.html.

Lahav, E., Benzion, U., & Shavit, T. (2011). The effect of military service on soldiers' time preferences—Evidence from Israel. *Judgment and Decision Making, 6*(2), 130–138.

Lerer, Z. (2009). *Groups of quality: The social history of the IDF selection system*. Ph.D. dissertation, Tel Aviv University.

Levitt, S. D. (2006). White-collar crime writ small: A case study of bagels, donuts, and the honor system. *American Economic Review Papers and Proceedings, 96*(2), 290–294.

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.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*(3), 683–729.

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization, 72*(1), 147–152.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*(4), 995–1027.

Pruckner, G. J., & Sausgruber, R. (2013). Honesty on the streets: A natural field experiment on newspaper purchasing. *Journal of the European Economic Association, 1*(3), 661–679.

Raven, J. C. (1936). *Mental tests used in genetic studies: the performance of related individuals on tests mainly educative and mainly reproductive*. MSc Thesis, University of London, London.

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.

Ruffle, B. J., & Tobol, Y. (2014). Honest on mondays: Honesty and the temporal distance between decisions and payoffs. *European Economic Review, 65*, 126–135.

Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes, 115*, 181–190.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science, 23*(10), 1264–1270.

Simon, H. A. (1990). A mechanism for social selection and successful altruism. *Science, 250*(4988), 1665–1668.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72*(2), 271–324.

Unger, S. M. (1964). Relation between intelligence and socially-approved behavior: A methodological cautionary note. *Child Development, 35*(1), 299–301.

Warner, J. T., & Pleeter, S. (2001). The personal discount rate: Evidence from military downsizing programs. *American Economic Review*, 33–53.

Wikipedia, http://he.wikipedia.org/wiki/.