Neyse, Levent; Fossen, Frank M.; Johannesson, Magnus; Dreber, Anna

Cognitive reflection and 2D:4D: Evidence from a large population sample

Journal of Economic Behavior and Organization

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Neyse, Levent; Fossen, Frank M.; Johannesson, Magnus; Dreber, Anna (2023) : Cognitive reflection and 2D:4D: Evidence from a large population sample, Journal of Economic Behavior and Organization, ISSN 1879-1751, Elsevier, Amsterdam, Vol. 209, pp. 288-307, https://doi.org/10.1016/j.jebo.2023.03.020

This Version is available at:
https://hdl.handle.net/10419/333208

Journal of Economic Behavior and Organization 209 (2023) 288–307

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Cognitive reflection and 2D:4D: Evidence from a large population sample

Levent Neyse[a,b,c], Frank M. Fossen[d,c], Magnus Johannesson[e], Anna Dreber[e,f,*]

[a] WZB, Social Science Center, Berlin, Germany
[b] SOEP at German Institute of Economic Research DIW, Berlin, Germany
[c] Institute of Labor Economics (IZA), Bonn, Germany
[d] University of Nevada, Reno, USA
[e] Department of Economics, Stockholm School of Economics, Sveavägen 65, Stockholm 113 83, Sweden
[f] Department of Economics, University of Innsbruck, Innsbruck, Austria

ARTICLE INFO

Article history:
Received 30 March 2022
Revised 18 September 2022
Accepted 19 March 2023
Available online 30 March 2023

Keywords:
Cognitive reflection test
2d:4d
Replication
Prenatal testosterone
Sex

ABSTRACT

Bosch-Domènech et al. (2014) reported a negative association between 2D:4D, a suggested marker of prenatal testosterone exposure, and the Cognitive Reflection Test (CRT) in a sample of 623 university students. In this pre-registered study, we test if we can replicate their findings in a general population sample of over 2,500 individuals from Germany. We find no statistically significant association between 2D:4D and the CRT in any of our primary hypothesis tests, or in any of our pre-registered exploratory analyses and robustness tests. The evidence is strong (based on the 99.5% confidence intervals in all three primary hypothesis tests) against effect sizes in the hypothesized direction larger than 0.075 CRT units (0.073 of the CRT standard deviation) for a one standard deviation change in 2D:4D.

1. Introduction

Dual process theories (e.g. Epstein, 1994; Kahneman and Frederick, 2002) attempt to explain decision making in two systems. In this framework, System I decisions are practical, fast and they make people's lives easier. Yet, they have the potential risk of being wrong as they are intuitive, unconscious and usually biased. System II decisions on the other hand require higher cognitive effort and more detailed analytical thinking. Many important decisions need a switch to System II to reduce decision errors. The Cognitive Reflection Test (CRT; Frederick, 2005) is one of the fundamental tools in social scientists' toolbox to study dual process theories.[1] CRT consists of three questions that test the respondent's ability to avoid an intuitive incorrect answer (System I) and to switch to the more analytical System II. The CRT has been reported to predict various behaviors and decision patterns including time and risk preferences (Frederick, 2005; Oechssler et al., 2009), performance on heuristics and biases in decision tasks (Toplak et al., 2011) or creativity (Barr et al., 2015). See the meta-analysis of Brañas-Garza et al. (2019a) for a detailed literature review.

* Corresponding author: Department of Economics, Stockholm School of Economics, Sveavägen 65, Stockholm 113 83, Sweden.
E-mail address: Anna.Dreber@hhs.se (A. Dreber).

[1] There are also studies criticizing dual process theories (see, e.g., Mercier and Sperber, 2011).

https://doi.org/10.1016/j.jebo.2023.03.020

Cumulative evidence suggests that men perform better than women on the CRT (e.g. Frederick, 2005; Kahan, 2012; Skagerlund et al., 2018). Following the repeatedly observed sex difference, several studies have set out to investigate the role of prenatal androgen exposure on CRT scores. The reason is that the prenatal exposure to testosterone is sexually dimorphic and it is suggested to have organizing effects on the brain and endocrinological development that might contribute to shaping personality traits and decision patterns (Phoenix et al., 1959; Arnold, 2009; Lombardo et al., 2012). Since male fetuses are exposed to higher levels of testosterone than female fetuses, it is hypothesized that prenatal testosterone exposure can explain sex differences in human behavior and psychology. The 2D:4D ratio (also referred to as the digit ratio), measured by dividing the lengths of the index and the ring fingers of human hands, has been suggested as a bio-marker of prenatal exposure to testosterone (Manning et al., 1998). A lower 2D:4D ratio is argued to indicate higher levels of prenatal testosterone exposure, with men usually having lower 2D:4D ratios than women (Hönekopp and Watson, 2010). However, there are also studies that did not find any significant sex differences in the 2D:4D ratio (e.g. Apicella et al., 2016). Using different methods including direct amniotic fluid draws, twin studies and investigation of diseases related to hormone deficiencies, there are examples of studies reporting a negative relationship between 2D:4D and prenatal testosterone exposure (e.g. Lutchmaya et al., 2004; van Anders et al., 2006; Manning et al., 2013; Ventura et al., 2013), but there are also many null results (e.g. Buck et al., 2003; Hollier et al., 2015; Hiraishi et al., 2012; Medland et al., 2008; Nave et al., 2020), with the latter ones tending to have larger sample sizes.

There are several papers studying the relationship between CRT scores and 2D:4D. In a sample of 623 university students, Bosch-Domènech et al. (2014) report a statistically significant negative relationship between the two.[2] The result holds for both hands and is stronger among women. In a sample of 243 male college students, Nave et al. (2017), among other things, test the relationship between 2D:4D and CRT scores, but fail to replicate the findings of Bosch-Domènech et al. (2014). Similarly, neither Millet and Aydinli (2019) nor Neyse et al. (2016) could detect a significant relationship between 2D:4D and CRT; but as the Nave et al. (2017) study they relied on samples of students that were smaller than Bosch-Domènech et al. (2014), with $n = 218$ and $n = 281$ respectively.

Considering the inconclusive results in the previous studies, we perform a replication study where we re-test the relationship between 2D:4D and CRT in a large, representative sample of more than 2500 respondents in the German Socio-Economic Panel's Innovation Sample (SOEP-IS). SOEP-IS is a panel study that has run with a representative sample of more than 4000 adult individuals for 10 years. There are at least three advantages of conducting this replication study in a large-scale household panel. First, the external validity issue of experiments (see Levitt and List, 2007) can partly be tackled in representative samples. Second, the sample sizes are usually much larger than in laboratory experiments. For example, the largest sample that investigates the relationship between 2D:4D and CRT scores is the one of Bosch-Domènech et al. (2014) (n=623). There is evidence that some share of experiments in experimental economics are potentially underpowered (see, e.g., Zhang and Ortmann, 2013; Maniadis et al., 2014; Bilén et al., 2021), thus it is important to work with larger sample sizes in order to determine which results are true versus false positive, and also to estimate effect sizes, which tend to be inflated even for true positive results when studies are underpowered (Gelman and Carlin 2014; Camerer et al., 2016, 2018). Finally, university students are now often experienced with the CRT as many of them are exposed to it in different experiments and questionnaires. Researchers have therefore started to re-phrase the original CRT questions to avoid familiarization (e.g. Capraro et al., 2017; Sirota et al., 2021). Familiarization with the CRT questions can be expected to be less of a problem in a household survey, making it possible to use the original CRT questions. Indeed, our results show that only 131 out of 2529 respondents indicated that they knew the CRT before participating in the study. Among those, 57 could answer all questions correctly. To test if we can replicate the findings of Bosch-Domènech et al. (2014), we design our pre-registered analysis as close as possible to the original study. We find no statistically significant association between 2D:4D and the CRT in any of our primary hypothesis tests, or in any of our pre-registered exploratory analyses and robustness tests.

2. Methods

2.1. Data collection and participants

The data collection was performed through the Innovation Sample of the German Socio-Economic Panel (SOEP-IS) (Richter and Schupp, 2015). SOEP-IS is a longitudinal survey that was established in 2012 and it is aimed at collecting new data in cooperation with external researchers. The SOEP-IS survey typically includes standard panel questions (e.g. income, work, family, health status) as well as experimental and survey modules that can be suggested by the international research community from various fields. The panel structure enables the combination of datasets and variables of different years and also the possibility to track changes in respondents' personal and professional lives. SOEP-IS has more than 4000 adult respondents, representative of the German population. The data and innovative modules are open to the scientific community two years after the delivery of the data. For the current study we use the data of the 2D:4D module from the 2018 wave and the CRT module from the 2020 wave. Both of these modules were collected as a result of the initiative and planning of

[2] Cueva et al. (2016) use a collection of datasets from various studies and find a negative relationship between 2D:4D and CRT scores only for the right hand ($p<0.05$). The statistical significance, however, disappears with inclusion of control variables in their regression analysis.

the authors of this study. The 2D:4D module has been available for researchers since April 2021 and the CRT data will be open to the research community in April 2023.

2.1.1. 2D:4D module

The 2D:4D data were collected in 2018 and used in three studies to investigate the relationship between 2D:4D and i) economic preferences (Neyse et al., 2021), ii) entrepreneurial career choice (Fossen et al., 2022) and depressive disorders (Lautenbacher and Neyse, 2020).[3]

The 2D:4D measurements were done by 263 interviewers with digital calipers. Both hands of the respondents were measured. The interviewers were thoroughly trained and were provided with printed guidelines. These guidelines were also tested by two research assistants, and their measurements were compared to those done with a flatbed scanner and an image editing software. The results were indistinguishable between the two methods. To minimize 2D:4D measurement errors, we included notes to interviewers in the guidelines, which can be found at https://osf.io/5vpdn/. The 2D:4D measurement guidelines indicated that if the 2D:4D ratio is outside the typical range of 0.8 and 1.1, the measurement should be done again. For this reason, some participants' hands were measured twice. Following the previous studies by Neyse et al. (2021) and Fossen et al. (2022) we construct three 2D:4D samples and repeat the analyses for all of these sample specifications:

(i) *The Main Sample* is based on the first 2D:4D measurement for all individuals and no 2D:4D observations have been excluded. The Main Sample is used for the main analyses.
(ii) *Corrected Sample* involves the second 2D:4D measurement instead of the first if the interviewer recorded two measures. We repeat the whole analysis using the Corrected Sample in our robustness tests and present them in the Appendix. By definition, the sample size remains the same as the size of the main sample.
(iii) *The Restricted Sample* involves participants with a 2D:4D in the range between 0.8 and 1.2 based on the first measurement of 2D:4D (this range corresponds to about +/- three to four STDs away from the mean in our data). We also rerun the whole analysis using the Restricted Sample in robustness tests and present them in the Appendix. The fraction of outliers excluded in the Restricted Sample was low (19 out of 2529 observations for left-hand 2D:4D and 8 out of 2522 observations for right-hand 2D:4D).

Appendix Table A1 presents descriptive 2D:4D statistics for the Main Sample, the Corrected Sample, and the Restricted Sample for individuals with data on gender and CRT (i.e. the samples included in the analyses below). It shows descriptive statistics for the left and right hands and the average of both hands for all three samples, and descriptive statistics are also reported separately for men and women. The mean 2D:4D and the standard deviation are somewhat higher in our study than in the study of Bosch-Domènech et al. (2014), and in line with Bosch-Domènech et al. (2014), 2D:4D is slightly higher for women than men.

2.1.2. CRT module

1. The CRT module was a part of the 2020 wave of the SOEP-IS. The module involved 6 questions in total. The first three questions were standard CRT items:
   A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (Intuitive answer: 10 cents. Correct answer: 5 cents)
2. If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets? (Intuitive answer: 100 min. Correct answer: 5 min)
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (Intuitive answer: 24 days. Correct answer: 47 days)

In the fourth question, respondents were asked whether they knew the answers before. The last two questions elicited respondents' beliefs about their own and other respondents' average scores. As noted in the pre-analysis plan, we do not use the prediction data in the current study. In line with Bosch-Domènech et al. (2014), we did not incentivize correct responses in the CRT. The meta-analysis of Brañas-Garza et al. (2019a) also concluded that the monetary incentives did not have a significant effect on CRT performance. Although the number of respondents who answered the three CRT questions were 3021 in total, we can analyze only 2529 since the remaining respondents did not participate in the 2D:4D module in 2018.

2.2. Statistical analysis and variables

In line with Benjamin et al. (2018), we use $p<0.005$ as our threshold for statistically significant evidence and $p<0.05$ for suggestive evidence. Our tests are based on two-sided p-values and our analyses are based on linear OLS regressions estimated with robust (Huber-White) standard errors. Note that our decision to use two-sided instead of one-sided p-values is based on the recommendation of Benjamin et al. (2018).

[3] The pre-analysis plans of the first two studies can be found at https://osf.io/5vpdn and at https://osf.io/t94fv respectively.

Our dependent variable in the regressions is the CRT score of the respondent. The main independent variable is 2D:4D. We also control for the respondent's sex in all the regressions. In a robustness test we also control for impatience and educational background information that are available in the panel data.

We pre-registered our hypotheses and published the detailed pre-analysis plan in a public repository before the CRT data collection was completed and before we had access to any CRT data (https://osf.io/259wx/). The pre-analysis plan, which we strictly follow, reports all the details about the variables and the analyses that we perform in this study. The pre-analysis plan was prepared in line with the analysis of Bosch-Domènech et al. (2014). The authors of Bosch-Domènech et al. (2014) also reviewed the pre-analysis plan and confirmed that it followed their analyses.

2.2.1. Variables

CRT-score: Our dependent variable is the *CRT-score* [0,3] in all our analysis.

2D:4D: Our main independent variable is *2D:4D*. We run the analysis for the right and left hands as well as the average of the both hands for each hypothesis.

Female: We will include sex (1 for female and 0 for male respondents) as a control variable in all regressions as the 2D:4D ratio can differ between men and women. Sex was also shown to be associated with CRT in previous studies (e.g. Frederick, 2005; Kahan, 2012; Skagerlund et al., 2018).

2D:4D X female: As a pre-registered exploratory analysis we test whether the association between 2D:4D and CRT differs between men and women. To do so, we add an interaction variable between the binary female variable and 2D:4D in the three OLS regressions used to test primary hypothesis 1a, 1b, and 1c (see below). Note that we do not specify a hypothesized direction of the interaction coefficient as these analyses are exploratory and would need confirmation in other studies to carry much weight. We also report the statistical significance of the 2D:4D effect separately for men and women as part of these analyses, but these tests are only relevant if the interaction coefficient is statistically significant.

Impatience & *education variables*: In another pre-registered robustness tests, we control for *Impatience* and *education variables* in the regressions. The education variables are four not mutually exclusive dummy variables. The motivation is that Bosch-Domènech et al. (2014) ran additional analyses with time discounting and mathematical skills as control variables.[4] While our dataset does not include these exact measures, we use the self-reported patience variable [0,10] included in the panel in our robustness test (controlling for *Impatience*), which was experimentally validated by Vischer et al. (2013). As we have the educational background of all respondents, we also include educational degree dummies as control variables in our robustness tests as a proxy variable for mathematical skills. The four dummy variables that we refer to as *education variables* are: The first indicating a high school degree that qualifies for university entrance in Germany ("Fachhochschulreife" or "Abitur"), referred to as "higher secondary school"; the second indicating completion of an apprenticeship, referred to as "apprenticeship"; the third indicating a degree from a university, a university of applied sciences, or a college or university outside Germany, referred to as "university degree"; and the fourth indicating completion of a vocational school, health care school, technical school, civil service training, or other vocational degree, referred to as "vocational degree".[5]

2.3. Main hypothesis

Our primary hypothesis relies on the study of Bosch-Domènech et al. (2014), which found that 2D:4D was negatively associated with CRT scores for both hands. We carry out the primary hypothesis test for the left-hand 2D:4D, the right-hand 2D:4D and for the average of the left-hand and the right-hand 2D:4D, using separate OLS regressions. The binary sex variable, *female*, is included in all regressions. Even though Bosch-Domènech et al. (2014) did not report results for the average of the left-hand and right-hand 2D:4D, we report results for this 2D:4D measure as well because taking the mean of these two measurements will reduce measurement noise. We refer to the three tests of the primary hypothesis as Primary Hypotheses 1a, 1b, and 1c.

Primary Hypothesis 1a. : Lower left-hand 2D:4D is associated with more correct answers in the CRT.

Primary Hypothesis 1b. : Lower right-hand 2D:4D is associated with more correct answers in the CRT.

Primary Hypothesis 1c. : Lower average of the left-hand and right-hand 2D:4D is associated with more correct answers in the CRT.

The primary hypothesis tests are carried out on the Main Sample described above, and all participants in this sample with data on 2D:4D, CRT and gender will be included in the analyses.

[4] Inclusion of these two control variables (time discounting and mathematical skills) in Bosch-Domènech et al. (2014) did not importantly change their results for the association between 2D:4D and CRT; the association was reported as statistically significant with and without these two control variables.

[5] We use the dummy variables for educational degrees that can be accumulated rather than dummies only capturing the highest degree obtained in order to make use of more information about individuals. For example, some individuals in Germany first obtain an apprenticeship before starting university studies. Allowing both the apprenticeship dummy and the university degree dummy to have the value one captures potential additional effects from the apprenticeship that would otherwise be ignored.

Table 1
CRT:% of correct answers by sex (for the left-hand 2D:4D sample, n=2529).

	All	Men	Women	p-value
CRT question 1 (bat & ball)	16.57%	19.76%	13.78%	0.0001
CRT question 2 (machine)	41.72%	47.84%	36.37%	<0.0001
CRT question 3 (lilies)	34.05%	43.68%	25.63%	<0.0001
0 correct answers	46.78%	38.00%	54.44%	
1 correct answer	24.71%	26.97%	22.74%	
2 correct answers	17.91%	20.78%	15.41%	
3 correct answers	10.60%	14.25%	7.41%	

Note: The p-values presented in the table are based on Fisher's Exact Tests of differences between men and women.

2.4. Pre-Registered exploratory analyses and robustness tests

As explained above, we also carried out a pre-registered exploratory analysis adding an interaction variable between 2D:4D and the female dummy variable in the above three regression equations.

We furthermore carry out three pre-registered robustness tests. In the first of those we repeat the analyses in the Corrected Sample and the Restricted Sample (including *all participants in each sample with data on 2D:4D, CRT and gender)*. In the second robustness test, we add the *impatience* and *education* variables and re-estimate the results in the Main Sample, the Corrected Sample and the Restricted Sample (including *all participants in each sample with data on 2D:4D, CRT, gender, impatience and education)*. Finally, we repeat all the analyses excluding respondents with previous CRT knowledge (i.e. respondents answering yes to the question: "Did you know the answers of these questions before?").

In total, we report 12 regression tables with 72 pre-registered OLS models including the 3 main hypothesis tests. Our main hypothesis tests and the three models with the gender interaction variable are presented in the main text, while the remaining 11 tables with the robustness test results are presented in the Appendix.

3. Results

3.1. Descriptive CRT results

Our analysis is based on the SOEP-IS respondents with data on both CRT and 2D:4D (in addition to data on the binary gender variable). For left-hand 2D:4D we have data for n=2529 individuals (1350 women), for right-hand 2D:4D we have data for n=2522 individuals (1346 women), and for the mean of the left-hand and right hand 2D:4D we have data for n=2503 individuals (1338 women). Table 1 presents the distribution of correct answers to each CRT question separately in percentages both for men and women based on the n=2529 left-hand 2D:4D sample. The upper section of the table displays the percentages of the participants who responded to each question correctly. 16.57% of all participants answered the first question (ball and bat) correctly, 41.72% responded to the second (machine and widget) correctly and 34.05% of them answered the third question (lake and lilies) correctly. Men scored statistically significantly higher than women in all questions ($p\text{-value}<0.005$ in all cases).

The bottom half of Table 1 presents the distribution of the number of correct answers in percentages. 46.78% of the respondents could not solve any of the questions correctly. 24.71% scored 1, 17.91% scored 2 and 10.60% responded to all questions correctly. We pre-registered the specification of all the descriptive results in Table 1 including the tests of a difference between men and women; but we did not specify in the pre-registration if these results would be reported for the left-hand 2D:4D sample (n = 2529), the right-hand 2D:4D sample (n = 2522), or the mean of the left and right-hand 2D:4D sample (n = 2503). In Appendix Tables A2-A3 we therefore also report this table for the two other samples showing very similar numbers.[6]

The number of correct answers on the CRT is our dependent variable in the regression analysis and it is 0.923 on average (STD=1.032) for the left-hand 2D:4D sample; 1.113 (STD=1.071) for men and 0.758 (STD=0.968) for women. For the right-hand 2D:4D sample the mean CRT score is 0.922 (STD=1.032) and for the mean of the left and right-hand 2D:4D sample the mean CRT score is 0.927 (STD=1.033).

In Bosch-Domènech et al. (2014) the fraction of men and women with 0 correct answers is 43.46% and 61.43% and in our study it is 38.00% and 54.44%. In Bosch-Domènech et al. (2014) the fraction of men and women with all correct answers is 11.54% and 5.23% and in our study it is 14.25% and 7.41%. These numbers are thus quite similar in the two studies. In the meta-analysis of Brañas-Garza et al. (2019a), the fraction of men and women with 0 correct answers is 27.01% and 45.09% and the fraction of men and women with all correct answers is 25.21% and 12.79%. CRT performance in the Brañas-Garza et al. (2019a) meta-analysis is hence somewhat better than in our study. As we include a general population sample and the meta-analysis includes a higher fraction of college students (42.28% students in the meta-analysis versus 2.97%

[6] As we have a general population sample, it can be argued that the external validity is higher in our study than in Bosch-Domènech et al. (2014) and Brañas-Garza et al. (2019a).

Table 2
Primary hypotheses tests and exploratory analysis.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.525			0.342		
	(0.304)			(0.481)		
Right-hand 2D:4D		−0.322			−0.524	
		(0.367)			(0.497)	
Mean 2D:4D			0.327			−0.048
			(0.423)			(0.633)
Left-hand 2D:4D X female				0.304		
				(0.621)		
Right-hand 2D:4D X female					0.445	
					(0.738)	
Mean 2D:4D X female						0.677
						(0.851)
Female	−0.360**	−0.344**	−0.352**	−0.664	−0.788	−1.028
	(0.041)	(0.041)	(0.041)	(0.622)	(0.739)	(0.852)
Constant	0.590	1.428**	0.789	0.772	1.629**	1.162
	(0.304)	(0.368)	(0.423)	(0.480)	(0.496)	(0.632)
Observations	2529	2522	2503	2529	2522	2503
R^2	0.031	0.028	0.029	0.031	0.028	0.029
Test of 2D:4D terms sum (F)				2.715	0.021	1.222
Test of 2D:4D terms sum (p)				0.100	0.885	0.269
MDE $p < 0.005$	0.074	0.074	0.075			
MDE $p < 0.05$	0.057	0.057	0.057			

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$
** $p<0.005$.MDE is the effect size the study had 80% power to detect at the $p<0.005$ and $p<0.05$ levels (measured in terms of CRT score units for a one standard deviation change in the 2D:4D variable).

students in our SOEP sample) that perform better according to the meta-analysis, a somewhat lower CRT score would be expected in our study. To assess how much this might matter, we report the CRT results of the small subset of respondents in our SOEP sample who were college students when CRT was measured in 2020 (75 respondents) in Appendix Table A4. Due to the small sample the CRT results are much less precise but are consistent with a higher CRT score among students. CRT performance is also affected by prior knowledge of the answers. In our study we directly measured this in a survey question and only about 5% of the respondents claimed that they knew the CRT answers prior to participating, while the corresponding number in the meta-analysis is not reported and may be higher. One difference in our study compared to both Bosch-Domènech et al. (2014) and Brañas-Garza et al. (2019a) is that the fraction of correct answers is highest on question 2 rather than on question 1. One potential explanation for this could be that the familiarity with the CRT is lower in our sample and familiarity with the CRT may disproportionally increase the number of correct answers on the first question (that may be the most commonly known question).

3.2. Main hypothesis tests and pre-registered exploratory analysis

Table 2 presents our primary hypothesis tests. Columns 1, 2 and 3 are dedicated to left-hand 2D:4D, right-hand 2D:4D and the mean 2D:4D analyses respectively. Columns 4, 5 and 6 present the pre-registered exploratory analyses where we repeat the same analysis with an additional interaction variable (2D:4D X female). For the interaction models the table also presents the tests of statistical significance of 2D:4D for women, i.e. tests of the null hypothesis that the sum of the coefficients of 2D:4D and its interaction with *female* equals zero.

Finally, the table also shows the minimum detectable effect size (MDE); i.e. the effect size the study had 80% power to detect at the $p<0.005$ level (3.65 times the standard error of the 2D:4D coefficient) and $p<0.05$ level (2.8 times the standard error of the 2D:4D coefficient). To make it easier to interpret the MDE, we multiplied it by the standard deviation of the 2D:4D for the sample included in the respective regression equation (so that it is measured as the change in the CRT score for a one standard deviation change in 2D:4D). The inclusion and estimation of the MDEs was also pre-registered. For our three primary hypotheses we have 80% power to detect an effect size of 0.074 (0.057) CRT score units at the $p<0.005$ ($p<0.05$) level for a one standard deviation change in 2D:4D. As the standard deviation of the CRT score variable is close to 1 (1.032), these MDEs in CRT score units correspond approximately to CRT standard deviation units. Our MDE estimations highlight that we have sufficient statistical power to detect even small potential effects.

We do not find statistically significant or suggestive evidence of an association between 2D:4D and CRT in any of our three primary hypothesis tests. For left-hand 2D:4D and the mean 2D:4D the sign of the coefficient is even in the opposite direction of the hypothesis (the hypothesis implies a negative coefficient). In Fig. 1 we plot the 95% and 99.5% confidence intervals of the 2D:4D regression coefficients for our three primary hypothesis tests; to ease the interpretation we have multiplied the point estimates of the coefficient and the confidence intervals by the standard deviation of the 2D:4D measure for the sample included in the respective regression (so that, as for the MDE above, they are measured as the change in the

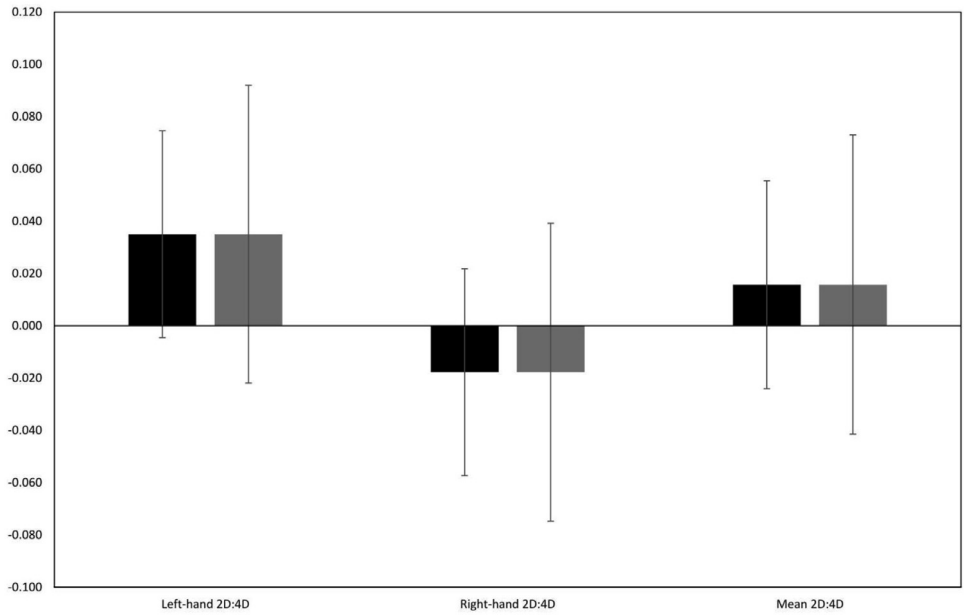

Fig. 1. Effect sizes for primary hypotheses tests (95% and 99.5% confidence intervals).
Note: The units of the effect sizes are the changes in the number of correctly answered CRT questions for a one standard deviation change in 2D:4D.

CRT score for a one standard deviation change in 2D:4D). Based on the 99.5% confidence intervals we find strong evidence against effect sizes in the hypothesized direction larger than 0.022 for left-hand 2D:4D, 0.075 for right-hand 2D:4D, and 0.042 for the mean 2D:4D.

In the pre-registered exploratory analyses testing for an interaction between 2D:4D and gender, reported in columns 4–6 of Table 2, we find no evidence of a difference in the 2D:4D coefficient between men and women ($p>0.05$).

3.3. Robustness tests

3.3.1. Robustness tests 1–2: corrected and restricted samples

As our analyses in Table 2 are based on the Main Sample, they involve outliers and/or a number of observations where the interviewers decided to repeat the first measurement. This is why we repeat the same analyses with the Corrected Sample and the Restricted Sample as explained in Section 2.1. The results remain the same as none of the significance levels change in neither of the Appendix Tables A5 and A6.

3.3.2. Robustness test 3: control variables: education and patience

In our third robustness test, we include education and patience variables in our models and repeat the same set of analyses that we presented in Table 2 and Appendix Tables A5-A6. Appendix Tables A7-A9 show the main hypothesis tests and the exploratory analyses including the control variables. Analyses in Appendix Table A7 are based on the Main Sample, while Appendix Tables A8 and A9 are based on the Corrected Sample and Restricted Sample, respectively. The results for the *2D:4D* variables remain the same with no statistically significant or suggestive evidence of an association between 2D:4D and CRT.

3.3.3. Robustness test 4: previous CRT knowledge

In our final robustness test, we exclude respondents who claimed that they knew the answers to the CRT questions before participating regardless of their actual scores ($n = 131$). Appendix Tables A10-A15 repeat all the analyses performed above excluding these respondents. Once again, results remain the same for the *2D:4D* variables.

In sum, there is no statistically significant or suggestive evidence of an association between 2D:4D and CRT in any of the robustness tests.

3.4. Non-Pre-Registered analysis: interviewer fixed effects

In this and the following subsection, we conduct additional analyses that were not pre-registered.[7] The first of these is a robustness test checking the potential concern that different enumerators visiting the households to measure the

[7] We thank the anonymous reviewers and the editor for suggesting these additional analyses.

Table 3
Post-study probability (PSP-rep) that the hypothesis is true after our replication.

Prior	80% power	90% power
0.10	0.022	0.011
0.25	0.063	0.032
0.50	0.167	0.091
0.75	0.376	0.232
0.90	0.644	0.475

Note: The PSP-rep is based on equation (4) in Maniadis et al. (2017). The PSP-rep is shown for different priors (the probability of the hypothesis being true prior to the replication), a statistical power of the replication of 80% and 90% and a statistical significance threshold of 0.005.

2D:4D ratio (and also conducting the interview) could make a difference to our results. Interviewer (enumerator) fixed effects are only expected to affect results if interviewers systematically affect both 2D:4D values and CRT scores, which seems unlikely, especially as the interviewers that measured 2D:4D in 2018 are not the same as the interviewers that measured CRT in 2020. To be sure, we repeat regressions (1) to (6) in Table 2 (the three primary hypotheses tests and the three exploratory analyses) with interviewer fixed effects, i.e., fixed effects for the interviewers that measured 2D:4D in 2018 (see Appendix Table A16). The results with interviewer fixed effects are similar to the results in Table 2 without interviewer fixed effects; there is no statistically significant or suggestive association between 2D:4D and CRT in any of the regressions, and there is no statistically significant or suggestive evidence of an interaction between gender and 2D:4D in the three exploratory analyses either. This confirms robustness of the results with respect to interviewer fixed effects.

3.5. Non-Pre-Registered analysis: priors

Maniadis et al. (2017) propose that replications should report by how much the replication affects the probability of the hypothesis being true. They propose a formula for this, equation (4) in their paper, that they refer to as the post-study probability, replication (PSP-rep). PSP-rep depends on the prior (the PSP after the original study has been published), the statistical power of the replication, the significance threshold used in the replication, and whether the replication found a significant effect in the same direction as the original study or not. We have estimated the PSP-rep based on five different priors (0.10, 0.25, 0.50, 0.75 and 0.90), and for 80% and 90% statistical power and the significance threshold of 0.005 used in our study.[8] The PSP-rep will be the same for the three primary hypotheses as we failed to reject the null hypothesis for all three hypotheses. The PSP-rep results are reported in Table 3 and show that the probability of the hypothesis being true decreases substantially due to the original result not replicating in our study. If, for instance, the probability of the hypothesis being true (the prior) was 50% after the publication of the original study, it decreases to about 17% after the replication for 80% power and to 9% for 90% power.

4. Conclusion

We test for an association between 2D:4D and CRT in a large, general population sample. Our study can be viewed as a replication of the study by Bosch-Domènech et al. (2014) who reported a negative association between 2D:4D and CRT in a sample of 623 university students (for both left-hand and right-hand 2D:4D). We modeled our pre-registered analyses on the analyses reported in their paper, but we fail to confirm their findings. We cannot reject the null hypothesis of no association irrespective of whether we use the left-hand 2D:4D, the right-hand 2D:4D, or the average of the two. We furthermore find no evidence of an interaction between 2D:4D and gender in a pre-registered exploratory analysis. These results are robust in a number of pre-registered as well as non-pre-registered robustness tests.

Our study is well powered with a sample size of about $n = 2500$, which is about four times as large as the sample used in Bosch-Domènech et al. (2014). For a one standard deviation change in 2D:4D, we had 80% power to detect a CRT effect size of at least 0.074 (0.057) CRT score units at the 0.5% (5%) level in our three primary hypothesis test regressions. Our estimated 99.5% confidence intervals also provide strong evidence against effect sizes larger than between 0.022–0.075 CRT score units in the hypothesized direction in the three main regression equations. In terms of CRT standard deviations this corresponds to 0.021–0.073 standard deviations. We cannot rule out effect sizes smaller than this. However, it should also be noted that we cannot rule out the possibility that the different results of the two studies might be due to heterogeneity in the association between 2D:4D and CRT between the two populations.

[8] Using 80% power implies that the estimations are done for a hypothesized effect size equal to the effect size we had 80% power to detect, which is the MDE we report for $p<0.005$ in Table 2; 0.074 for primary hypotheses 1a and 1b and 0.075 for primary hypothesis 1c measured in terms of CRT score units for a one standard deviation change in the 2D:4D variable. The corresponding effect sizes for 90% power are 0.083 for primary hypotheses 1a and 1b, and 0.084 for primary hypothesis 1c.

There have been attempts to link 2D:4D to a number of behavioral outcomes and traits. For example, behavioral economists have investigated 2D:4D and economic preferences such as risk taking and social preferences (see Neyse et al., 2021, and Parslow et al., 2019, for reviews), psychologists have investigated the relationships with personality traits (e.g. Voracek, 2011; Fink et al., 2004) and cognitive scientists have studied cognitive skills and decision patterns (e.g. Austin et al., 2002; Puts et al., 2008). The results do not provide strong evidence for clear associations. For economic preferences, the largest study to date is Neyse et al. (2021) who also use the SOEP-IS to obtain a sample of about 3450 respondents. In their pre-registered study, they find no statistically significant associations between 2D:4D and preferences related to risk taking, altruism, positive reciprocity, negative reciprocity and trust. Similarly, Brañas-Garza et al. (2019b) do not detect a significant correlation between 2D:4D and generosity, bargaining or trust-related behavior in a sample of 560 students. To the extent that previous studies report statistically significant associations, these may partly be due to reasons such as small sample sizes, publication bias and the various "researcher degrees of freedom" such as for example "forking" (Simmons et al., 2011; Gelman and Loken, 2013) where even researchers who are testing clear hypotheses have flexibility in finding statistically significant results, making such results more likely to be false positive ones. How much the results vary because some studies are done in the lab and some not (like our study) is however not clear.

The evidence linking 2D:4D to prenatal testosterone exposure is also inconsistent. For example, while Lutchmaya et al. (2004) found a link between the testosterone-to-estradiol ratio in amniotic fluid and right-hand but not left-hand 2D:4D in a sample of 29 children and Ventura et al. (2013) found a link between maternal plasma testosterone levels and 2D:4D for both hands for 106 newborns, Hollier et al. (2015) found no such evidence for hormone measures from umbilical cord blood in a sample of 341 children. Similarly, there are studies based on the theory of sex hormone transfer in utero, testing whether women with male twins have lower 2D:4D than women with female twins. While there are positive results (e.g. van Anders et al., 2006), larger studies report no statistically significant differences (Hiraishi et al., 2012; Medland et al., 2008).

The 2D:4D literature investigating human behavior needs large-scale, pre-registered studies to draw firm conclusions. This study helps bridge this gap and finds no evidence of an association between 2D:4D and the CRT. This is in line with three recent related large pre-registered studies on the association between 2D:4D and economic preferences and entrepreneurship (Van Leeuwen et al., 2020; Neyse et al., 2021; Fossen et al., 2022). A key issue moving forward is to test if there actually is an association between prenatal testosterone exposure and 2D:4D in a well-powered pre-registered study, to establish if further work on 2D:4D and human behavior is warranted.

Funding

We thank Riksbankens Jubileumsfond (P21–0168), the Jan Wallander and Tom Hedelius Foundation, the Knut and Alice Wallenberg Foundation and the Marianne and Marcus Wallenberg Foundation (Anna Dreber is a Wallenberg Scholar), and the Austrian Science Fund (FWF, SFB F63). Levent Neyse acknowledges the financial support of Leibniz Association (SAW-10868) and German Research Foundation (DFG) through project CRC TRR 190 Rationality and Competition.

Declaration of Competing Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

We thank the authors of Bosch-Domènech et al. (2014) for reviewing the pre-analysis plan and providing feedback.

Appendix

Table A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12,A13,A14,A15,A16.

Table A1
2D:4D descriptive statistics table.

	All			Men			Women		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
Left-hand 2D:4D	1.001	0.067	2529	0.996	0.062	1179	1.005	0.071	1350
LH 2D:4D corrected	0.998	0.054	2529	0.994	0.053	1179	1.001	0.054	1350
LH 2D:4D restricted	1.000	0.054	2510	0.995	0.054	1171	1.004	0.054	1339
Right-hand 2D:4D	0.999	0.055	2522	0.997	0.060	1176	1.001	0.051	1346
RH 2D:4D corrected	0.998	0.052	2522	0.996	0.057	1176	1.000	0.048	1346
RH 2D:4D restricted	0.998	0.048	2514	0.996	0.047	1172	1.001	0.049	1342
Mean 2D:4D	1.000	0.048	2503	0.996	0.047	1165	1.003	0.049	1338
Mean 2D:4D corrected	0.998	0.042	2503	0.995	0.044	1165	1.001	0.041	1338
Mean 2D:4D restricted	0.999	0.042	2478	0.996	0.041	1153	1.002	0.042	1325

Note: The left-hand (LH) and right-hand (RH) 2D:4D and the mean 2D:4D of the two hands (Mean) are presented for each sample criterion (Main, Corrected and Restricted).

Table A2
CRT:% of correct answers by sex (for the right-hand 2D:4D sample, n=2522).

	All	Men	Women	p-value
CRT question 1 (bat & ball)	16.42%	19.39%	13.82%	0.0002
CRT question 2 (machine)	41.75%	47.70%	36.55%	<0.0001
CRT question 3 (lilies)	34.06%	43.54%	25.78%	<0.0001
0 correct answers	46.79%	38.35%	54.16%	
1 correct answer	24.78%	26.87%	22.96%	
2 correct answers	17.84%	20.58%	15.45%	
3 correct answers	10.59%	14.20%	7.43%	

Note: The p-values presented in the table are based on Fisher's Exact Tests of differences between men and women.

Table A3
CRT:% of correct answers by sex (for the mean 2D:4D sample, *n*=2503).

	All	Men	Women	p-value
CRT question 1 (bat & ball)	16.54%	19.57%	13.90%	0.0002
CRT question 2 (machine)	41.99%	48.07%	36.70%	<0.0001
CRT question 3 (lilies)	34.20%	43.78%	25.86%	<0.0001
0 correct answers	46.54%	37.94%	54.04%	
1 correct answer	24.85%	27.04%	22.94%	
2 correct answers	17.94%	20.69%	15.55%	
3 correct answers	10.67%	14.33%	7.47%	

Note: The p-values presented in the table are based on Fisher's Exact Tests of differences between men and women.

Table A4
CRT:% of Correct Answers by Sex (for the subsample of college students in 2020, *n*=75).

	All	Men	Women	p-value
CRT question 1 (bat & ball)	48.00%	48.72%	47.22%	1.0000
CRT question 2 (machine)	66.67%	76.92%	55.56%	0.0852
CRT question 3 (lilies)	61.33%	82.05%	38.89%	0.0002
0 correct answers	17.33%	7.69%	27.78%	
1 correct answer	21.33%	20.51%	22.22%	
2 correct answers	29.33%	28.21%	30.56%	
3 correct answers	32.00%	43.59%	19.44%	

Note: The p-values presented in the table are based on Fisher's Exact Tests of differences between men and women.

Table A5
First robustness test: corrected 2D:4D measure.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.608			0.474		
	(0.379)			(0.560)		
Right-hand 2D:4D		−0.254			−0.464	
		(0.389)			(0.523)	
Mean 2D:4D			0.310			0.008
			(0.481)			(0.685)
Left-hand 2D:4D X female				0.246		
				(0.760)		
Right-hand 2D:4D X female					0.472	
					(0.783)	
Mean 2D:4D X female						0.595
						(0.963)
Female	−0.359**	−0.344**	−0.351**	−0.605	−0.815	−0.945
	(0.041)	(0.041)	(0.041)	(0.759)	(0.783)	(0.962)
Constant	0.509	1.359	0.806	0.641	1.569**	1.106
	(0.378)	(0.389)	(0.480)	(0.557)	(0.522)	(0.683)
Observations	2529	2522	2503	2529	2522	2503
R^2	0.030	0.028	0.029	0.030	0.028	0.029
Test of 2D:4D terms sum (F)				1.961	0.000	0.794
Test of 2D:4D terms sum (p)				0.161	0.989	0.373
MDE $p < 0.005$	0.074	0.074	0.075			
MDE $p < 0.05$	0.057	0.057	0.057			

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$
** $p<0.005$.MDE is the effect size the study had 80% power to detect at the $p<0.005$ and $p<0.05$ levels (measured in terms of CRT score units for a one standard deviation change in the 2D:4D variable).

Table A6
Second robustness test: restricted sample.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.307			0.355		
	(0.377)			(0.555)		
Right-hand 2D:4D		−0.098			−0.297	
		(0.424)			(0.631)	
Mean 2D:4D			0.257			0.282
			(0.492)			(0.729)
Left-hand 2D:4D X female				−0.090		
				(0.755)		
Right-hand 2D:4D X female					0.362	
					(0.852)	
Mean 2D:4D X female						−0.046
						(0.988)
Female	−0.360**	−0.345**	−0.356**	−0.270	−0.706	−0.310
	(0.041)	(0.041)	(0.041)	(0.756)	(0.851)	(0.988)
Constant	0.807*	1.207	0.860	0.758	1.405*	0.835
	(0.376)	(0.423)	(0.491)	(0.553)	(0.629)	(0.726)
Observations	2510	2514	2478	2510	2514	2478
R^2	0.030	0.028	0.029	0.030	0.028	0.029
Test of 2D:4D terms sum (F)				0.267	0.013	0.126
Test of 2D:4D terms sum (p)				0.605	0.910	0.723
MDE $p < 0.005$	0.074	0.074	0.075			
MDE $p < 0.05$	0.057	0.057	0.057			

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.MDE is the effect size the study had 80% power to detect at the $p<0.005$ and $p<0.05$ levels (measured in terms of CRT score units for a one standard deviation change in the 2D:4D variable).

Table A7
Third robustness test: with control variables.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.228			0.003		
	(0.290)			(0.460)		
Right-hand 2D:4D		−0.261			−0.522	
		(0.356)			(0.478)	
Mean 2D:4D			0.109			−0.334
			(0.406)			(0.609)
Left-hand 2D:4D X female				0.373		
				(0.592)		
Right-hand 2D:4D X female					0.584	
					(0.716)	
Mean 2D:4D X female						0.796
						(0.816)
Female	−0.316**	−0.303**	−0.308**	−0.689	−0.887	−1.104
	(0.039)	(0.039)	(0.040)	(0.592)	(0.717)	(0.816)
Patience	−0.009	−0.010	−0.009	−0.009	−0.010	−0.009
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Higher sec. school	0.597**	0.606**	0.601**	0.597**	0.607**	0.601**
	(0.051)	(0.050)	(0.051)	(0.051)	(0.050)	(0.051)
Apprenticeship	0.006	−0.003	0.006	0.006	−0.004	0.005
	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)	(0.054)
Vocational degree	0.133*	0.124*	0.126*	0.133*	0.122*	0.125*
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
University degree	0.185**	0.173**	0.182**	0.185**	0.172**	0.181**
	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)
Constant	0.646*	1.136**	0.765	0.869	1.396**	1.206*
	(0.299)	(0.365)	(0.413)	(0.464)	(0.485)	(0.613)
Observations	2474	2467	2448	2474	2467	2448
R^2	0.134	0.133	0.132	0.134	0.133	0.133
Test of 2D:4D terms sum (F)				1.018	0.014	0.723
Test of 2D:4D terms sum (p)				0.313	0.907	0.395

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.

Table A8
Third robustness test: corrected 2D:4D measure with control variables.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.270			0.174		
	(0.362)			(0.536)		
Right-hand 2D:4D		−0.265			−0.565	
		(0.372)			(0.499)	
Mean 2D:4D			0.064			−0.299
			(0.461)			(0.657)
Left-hand 2D:4D X female				0.175		
				(0.727)		
Right-hand 2D:4D X female					0.673	
					(0.748)	
Mean 2D:4D X female						0.713
						(0.921)
Female	−0.316**	−0.303**	−0.308**	−0.491	−0.974	−1.019
	(0.039)	(0.039)	(0.040)	(0.726)	(0.748)	(0.920)
Patience	−0.009	−0.010	−0.009	−0.009	−0.010	−0.009
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Higher sec. school	0.597**	0.606**	0.601**	0.597**	0.607**	0.601**
	(0.051)	(0.050)	(0.051)	(0.051)	(0.050)	(0.051)
Apprenticeship	0.006	−0.003	0.006	0.006	−0.004	0.006
	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)	(0.054)
Vocational degree	0.133*	0.124*	0.126*	0.133*	0.122*	0.125*
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
University degree	0.185**	0.174**	0.182**	0.185**	0.172**	0.181**
	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)
Constant	0.605	1.139**	0.810	0.700	1.438**	1.171
	(0.368)	(0.380)	(0.466)	(0.538)	(0.504)	(0.659)
Observations	2474	2467	2448	2474	2467	2448
R^2	0.134	0.133	0.132	0.134	0.133	0.133
Test of 2D:4D terms sum (F)				0.508	0.038	0.411
Test of 2D:4D terms sum (p)				0.476	0.846	0.521

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.

Table A9
Third robustness test: restricted sample with control variables.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.009			0.102		
	(0.359)			(0.531)		
Right-hand 2D:4D		−0.143			−0.521	
		(0.404)			(0.602)	
Mean 2D:4D			−0.013			−0.126
			(0.469)			(0.696)
Left-hand 2D:4D X female				−0.172		
				(0.721)		
Right-hand 2D:4D X female					0.690	
					(0.813)	
Mean 2D:4D X female						0.207
						(0.943)
Female	−0.314**	−0.302**	−0.308**	−0.142	−0.991	−0.515
	(0.039)	(0.039)	(0.040)	(0.721)	(0.812)	(0.942)
Patience	−0.009	−0.010	−0.009	−0.009	−0.009	−0.009
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Higher sec. school	0.601**	0.606**	0.603**	0.601**	0.607**	0.603**
	(0.051)	(0.050)	(0.051)	(0.051)	(0.051)	(0.051)
Apprenticeship	0.014	−0.000	0.017	0.015	−0.001	0.017
	(0.053)	(0.053)	(0.054)	(0.053)	(0.053)	(0.054)
Vocational degree	0.138*	0.127*	0.135*	0.139*	0.126*	0.134*
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
University degree	0.199**	0.176**	0.197**	0.199**	0.175**	0.197**
	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)
Constant	0.851*	1.015*	0.873	0.759	1.391*	0.986
	(0.367)	(0.411)	(0.475)	(0.534)	(0.605)	(0.698)
Observations	2455	2461	2425	2455	2461	2425
$_R2$	0.136	0.133	0.135	0.136	0.133	0.135
Test of 2D:4D terms sum (F)				0.020	0.096	0.016
Test of 2D:4D terms sum (p)				0.887	0.757	0.898

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.

Table A10
Fourth robustness test: excluding respondents who knew answers.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.344			−0.019		
	(0.304)			(0.477)		
Right-hand 2D:4D		−0.435			−0.654	
		(0.360)			(0.487)	
Mean 2D:4D			0.070			−0.465
			(0.420)			(0.624)
Left-hand 2D:4D X female				0.612		
				(0.618)		
Right-hand 2D:4D X female					0.486	
					(0.724)	
Mean 2D:4D X female						0.978
						(0.843)
Female	−0.341**	−0.327**	−0.335**	−0.952	−0.813	−1.312
	(0.040)	(0.040)	(0.041)	(0.619)	(0.725)	(0.844)
Constant	0.693*	1.464**	0.969*	1.054*	1.683**	1.502*
	(0.304)	(0.360)	(0.419)	(0.475)	(0.486)	(0.622)
Observations	2398	2393	2374	2398	2393	2374
$_R2$	0.029	0.028	0.028	0.029	0.028	0.028
Test of 2D:4D terms sum (F)				2.262	0.098	0.817
Test of 2D:4D terms sum (p)				0.133	0.754	0.366

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$,.
** $p<0.005$.

Table A11
Fourth robustness test: excluding respondents who knew answers; corrected 2D:4D measure.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.481			0.111		
	(0.377)			(0.555)		
Right-hand 2D:4D		−0.339			−0.542	
		(0.382)			(0.512)	
Mean 2D:4D			0.138			−0.328
			(0.476)			(0.675)
Left-hand 2D:4D X female				0.686		
				(0.757)		
Right-hand 2D:4D X female					0.458	
					(0.769)	
Mean 2D:4D X female						0.927
						(0.953)
Female	−0.341**	−0.328**	−0.335**	−1.025	−0.785	−1.260
	(0.040)	(0.040)	(0.041)	(0.755)	(0.768)	(0.951)
Constant	0.558	1.368**	0.902	0.925	1.571**	1.365*
	(0.376)	(0.381)	(0.475)	(0.552)	(0.511)	(0.672)
Observations	2398	2393	2374	2398	2393	2374
R^2	0.029	0.027	0.028	0.030	0.028	0.028
Test of 2D:4D terms sum (F)				2.405	0.022	0.794
Test of 2D:4D terms sum (p)				0.121	0.883	0.373

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$,.
** $p<0.005$.

Table A12
Fourth robustness test: excluding respondents who knew answers; restricted sample.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.288			−0.014		
	(0.374)			(0.551)		
Right-hand 2D:4D		−0.267			−0.544	
		(0.418)			(0.624)	
Mean 2D:4D			0.127			−0.217
			(0.488)			(0.724)
Left-hand 2D:4D X female				0.559		
				(0.751)		
Right-hand 2D:4D X female					0.504	
					(0.841)	
Mean 2D:4D X female						0.632
						(0.981)
Female	−0.341**	−0.329**	−0.338**	−0.900	−0.831	−0.969
	(0.041)	(0.040)	(0.041)	(0.751)	(0.840)	(0.980)
Constant	0.749*	1.299**	0.915	1.049	1.575*	1.258
	(0.373)	(0.417)	(0.487)	(0.549)	(0.622)	(0.721)
Observations	2382	2385	2352	2382	2385	2352
R^2	0.029	0.027	0.028	0.029	0.028	0.028
Test of 2D:4D terms sum (F)				1.145	0.005	0.393
Test of 2D:4D terms sum (p)				0.285	0.943	0.531

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.

Table A13
Fourth robustness test: excluding respondents who knew answers with control variables.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.074			−0.297		
	(0.290)			(0.456)		
Right-hand 2D:4D		−0.372			−0.625	
		(0.350)			(0.470)	
Mean 2D:4D			−0.117			−0.680
			(0.403)			(0.601)
Left-hand 2D:4D X female				0.622		
				(0.590)		
Right-hand 2D:4D X female					0.570	
					(0.705)	
Mean 2D:4D X female						1.023
						(0.810)
Female	−0.308**	−0.298**	−0.302**	−0.930	−0.867	−1.324
	(0.039)	(0.039)	(0.039)	(0.591)	(0.705)	(0.810)
Patience	−0.008	−0.008	−0.008	−0.008	−0.008	−0.008
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Higher sec. school	0.588**	0.596**	0.592**	0.588**	0.597**	0.593**
	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)
Apprenticeship	−0.007	−0.015	−0.006	−0.007	−0.017	−0.008
	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)
Vocational degree	0.096	0.092	0.093	0.095	0.090	0.091
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
University degree	0.169**	0.160*	0.168*	0.169**	0.158*	0.167*
	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)
Constant	0.753*	1.197**	0.942*	1.121*	1.450**	1.503*
	(0.299)	(0.359)	(0.410)	(0.460)	(0.476)	(0.604)
Observations	2350	2345	2326	2350	2345	2326
$_R2$	0.133	0.134	0.133	0.134	0.134	0.134
Test of 2D:4D terms sum (F)				0.749	0.011	0.400
Test of 2D:4D terms sum (p)				0.387	0.917	0.527

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.

Table A14
Fourth robustness test: excluding respondents who knew answers; corrected 2D:4D measure with control variables.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.170			−0.102		
	(0.361)			(0.533)		
Right-hand 2D:4D		−0.344			−0.619	
		(0.366)			(0.490)	
Mean 2D:4D			−0.078			−0.551
			(0.456)			(0.648)
Left-hand 2D:4D X female				0.503		
				(0.725)		
Right-hand 2D:4D X female					0.621	
					(0.736)	
Mean 2D:4D X female						0.939
						(0.913)
Female	−0.309**	−0.298**	−0.302**	−0.810	−0.917	−1.238
	(0.039)	(0.039)	(0.039)	(0.723)	(0.736)	(0.911)
Patience	−0.008	−0.008	−0.008	−0.008	−0.008	−0.008
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Higher sec. school	0.588**	0.596**	0.592**	0.588**	0.597**	0.593**
	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)
Apprenticeship	−0.007	−0.015	−0.006	−0.007	−0.016	−0.007
	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)
Vocational degree	0.096	0.092	0.094	0.095	0.090	0.092
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
University degree	0.169**	0.160*	0.168*	0.169**	0.159*	0.168*
	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)
Constant	0.657	1.169**	0.903	0.926	1.444**	1.373*
	(0.367)	(0.373)	(0.462)	(0.534)	(0.495)	(0.650)
Observations	2350	2345	2326	2350	2345	2326
$_R2$	0.133	0.134	0.133	0.133	0.134	0.134
Test of 2D:4D terms sum (F)				0.667	0.000	0.363
Test of 2D:4D terms sum (p)				0.414	0.998	0.547

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.

Table A15
Fourth robustness test: excluding respondents who knew answers; restricted sample with control variables.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.040			−0.155		
	(0.358)			(0.529)		
Right-hand 2D:4D		−0.295			−0.709	
		(0.400)			(0.596)	
Mean 2D:4D			−0.094			−0.502
			(0.467)			(0.693)
Left-hand 2D:4D X female				0.360		
				(0.718)		
Right-hand 2D:4D X female					0.752	
					(0.804)	
Mean 2D:4D X female						0.746
						(0.937)
Female	−0.306**	−0.297**	−0.301**	−0.666	−1.047	−1.046
	(0.039)	(0.039)	(0.040)	(0.718)	(0.803)	(0.936)
Patience	−0.007	−0.008	−0.007	−0.007	−0.008	−0.007
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Higher sec. school	0.589**	0.596**	0.592**	0.589**	0.598**	0.592**
	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)
Apprenticeship	−0.003	−0.013	−0.000	−0.004	−0.013	−0.001
	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)
Vocational degree	0.094	0.095	0.095	0.094	0.094	0.094
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
University degree	0.182**	0.163*	0.182**	0.182**	0.161*	0.182**
	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)	(0.060)
Constant	0.777*	1.118*	0.909	0.971	1.529*	1.313
	(0.365)	(0.406)	(0.472)	(0.532)	(0.598)	(0.694)
Observations	2334	2339	2306	2334	2339	2306
R^2	0.135	0.134	0.135	0.135	0.134	0.135
Test of 2D:4D terms sum (F)				0.177	0.006	0.150
Test of 2D:4D terms sum (p)				0.674	0.936	0.699

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$.
** $p<0.005$.

Table A16
Not pre-registered robustness test: with interviewer fixed effects.

	(1)	(2)	(3)	(4)	(5)	(6)
Left-hand 2D:4D	0.337			0.402		
	(0.317)			(0.494)		
Right-hand 2D:4D		−0.424			−0.394	
		(0.383)			(0.504)	
Mean 2D:4D			0.067			0.058
			(0.443)			(0.650)
Left-hand 2D:4D X female				−0.108		
				(0.631)		
Right-hand 2D:4D X female					−0.069	
					(0.762)	
Mean 2D:4D X female						0.017
						(0.870)
Female	−0.348**	−0.333**	−0.341**	−0.241	−0.264	−0.358
	(0.040)	(0.040)	(0.040)	(0.632)	(0.762)	(0.871)
Interviewer Fixed Effects	YES	YES	YES	YES	YES	YES
Observations	2529	2522	2503	2529	2522	2503
R^2	0.185	0.183	0.185	0.185	0.183	0.185
Test of 2D:4D terms sum (F)				0.526	0.639	0.016
Test of 2D:4D terms sum (p)				0.469	0.424	0.899

Note: OLS regressions; standard errors in parentheses.
* $p<0.05$
** $p<0.005$.Dummy variables for each interviewer in 2018 are added as independent variables to the regressions; apart from that, the models and samples are the same as in Table 2.

References

Apicella, C.L., Tobolsky, V.A., Marlowe, F.W., Miller, K.W., 2016. Hadza hunter-gatherer men do not have more masculine digit ratios (2 D: 4 D). Am. J. Phys. Anthropol. 159 (2), 223–232.

Arnold, A.P., 2009. The organizational–activational hypothesis as the foundation for a unified theory of sexual differentiation of all mammalian tissues. Horm Behav. 55 (5), 570–578.

Austin, E.J., Manning, J.T., McInroy, K., Mathews, E., 2002. A preliminary investigation of the associations between personality, cognitive ability and digit ratio. Pers. Individ. Differ. 33 (7), 1115–1124.

Barr, N., Pennycook, G., Stolz, J.A., Fugelsang, J.A., 2015. Reasoned connections: a dual-process perspective on creative thought. Think Reason 21 (1), 61–75.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Johnson, V.E., 2018. Redefine statistical significance. Nature Hum. Behav. 2 (1), 6–10.

Bilén, D., Dreber, A., Johannesson, M., 2021. Are women more generous than men? a meta-analysis. J. Econ. Sci. Assoc. 7, 1–18.

Bosch-Domènech, A., Brañas-Garza, P., Espín, A.M., 2014. Can exposure to prenatal sex hormones (2D: 4D) predict cognitive reflection? Psychoneuroendocrinology 43, 1–10.

Brañas-Garza, P., Kujal, P., Lenkei, B., 2019a. Cognitive reflection test: whom, how, when. J. Behav. Exp. Econ. 82, 101455.

Brañas-Garza, P., Espín, A.M., García-Muñoz, T., Kovářík, J., 2019b. Digit ratio (2D: 4D) and prosocial behaviour in economic games: no direct correlation with generosity, bargaining or trust-related behaviours. Biol. Lett. 15 (8), 20190185.

Buck, J.J., Williams, R.M., Hughes, I.A., Acerini, C.L., 2003. In-utero androgen exposure and 2nd to 4th digit length ratio—Comparisons between healthy controls and females with classical congenital adrenal hyperplasia. Hum. Reprod. 18 (5), 976–979.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-.H., Huber, J., Johannesson, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351, 1433–1436.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-.H., Huber, J., Johannesson, M., Wu, H., 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat. Hum. Behav. 2, 637–644.

Capraro, V., Corgnet, B., Espín, A.M., Hernán-González, R., 2017. Deliberation favours social efficiency by making people disregard their relative shares: evidence from USA and India. R. Soc. Open Sci. 4 (2), 160605.

Cueva, C., Iturbe-Ormaetxe, I., Mata-Pérez, E., Ponti, G., Sartarelli, M., Yu, H., Zhukova, V., 2016. Cognitive (ir) reflection: new experimental evidence. J. Behav. Exp. Econ. 64, 81–93.

Epstein, S., 1994. Integration of the cognitive and the psychodynamic unconscious. Am. Psychol. 49 (8), 709.

Fink, B., Manning, J.T., Neave, N., 2004. Second to fourth digit ratio and the 'big five' personality factors. Pers. Individ. Differ. 37 (3), 495–503.

Fossen, F.M., Neyse, L., Johannesson, M., Dreber, A., 2022. 2D:4D and self-employment: a preregistered replication study in a large general population survey. Entrep. Theory Pract. 46 (1).

Frederick, S., 2005. Cognitive reflection and decision making. J. Econ. Perspect. 19 (4), 25–42.

Gelman, A., Carlin, J., 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect. Psychol. Sci. 9 (6), 641–651.

Gelman, A., Loken, E., 2013. The Garden of Forking paths: Why multiple Comparisons Can Be a problem, Even When There is No "fishing expedition" Or "p-hacking" and the Research Hypothesis Was Posited Ahead of Time. Department of Statistics, Columbia University, p. 348.

Hiraishi, K., Sasaki, S., Shikishima, C., Ando, J., 2012. The second to fourth digit ratio (2d:4d) in a Japanese twin sample: heritability, prenatal hormone transfer, and association with sexual orientation. Arch. Sex. Behav. 41 (3), 711–724.

Hollier, L.P., Keelan, J.A., Jamnadass, E.S., Maybery, M.T., Hickey, M., Whitehouse, A.J., 2015. Adult digit ratio (2d: 4d) is not related to umbilical cord androgen or estrogen concentrations, their ratios or net bioactivity. Early Hum. Dev. 91 (2), 111–117.

Hönekopp, J., Watson, S., 2010. Meta-analysis of digit ratio 2D: 4D shows greater sex difference in the right hand. Am. J. Hum. Biol. 22 (5), 619–630.

Kahan, D.M., 2012. Ideology, motivated reasoning, and cognitive reflection: an experimental study. Judgm. Decis. Mak. 8, 407–424.

Kahneman D., & Frederick S. (2002). Representativeness revisited: attribute substitution in intuitive judgment. Heuristics and biases: The Psychology of Intuitive Judgment, 49, 81.

Lautenbacher, L.M., Neyse, L., 2020. Depression, neuroticism and 2D: 4D ratio: evidence from a large, representative sample. Sci. Rep. 10 (1), 1–10.

Levitt, S.D., List, J.A., 2007. What do laboratory experiments measuring social preferences reveal about the real world? J. Econ. Perspect. 21 (2), 153–174.

Lombardo, M.V., Ashwin, E., Auyeung, B., Chakrabarti, B., Taylor, K., Hackett, G., Baron-Cohen, S., 2012. Fetal testosterone influences sexually dimorphic gray matter in the human brain. J. Neurosci. 32 (2), 674–680.

Lutchmaya, S., Baron-Cohen, S., Raggatt, P., Knickmeyer, R., Manning, J.T., 2004. 2nd to 4th digit ratios, fetal testosterone and estradiol. Early Hum. Dev. 77 (1–2), 23–28.

Maniadis, Z., Tufano, F., List, J.A., 2014. One swallow doesn't make a summer: new evidence on anchoring effects. Am. Econ. Rev. 104 (1), 277–290.

Maniadis, Z., Tufano, F., List, J.A., 2017. To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. Econ. J. 127 (605), F209–F235.

Manning, J.T., Scutt, D., Wilson, J., Lewis-Jones, D.I., 1998. The ratio of 2nd to 4th digit length: a predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. Hum. Reprod. 13 (11), 3000–3004.

Manning, J.T., Kilduff, L.P., Trivers, R., 2013. Digit ratio (2D: 4D) in Klinefelter's syndrome. Andrology 1 (1), 94–99.

Medland, S.E., Loehlin, J.C., Martin, N.G., 2008. No effects of prenatal hormone transfer on digit ratio in a large sample of same-and opposite-sex dizygotic twins. Pers. Individ. Differ. 44 (5), 1225–1234.

Mercier, H., Sperber, D., 2011. Why do humans reason? Arguments for an argumentative theory. Behav. Brain Sci. 34, 57–111.

Nave, G., Nadler, A., Zava, D., Camerer, C., 2017. Single-dose testosterone administration impairs cognitive reflection in men. Psychol. Sci. 28 (10), 1398–1407.

Nave, G., Koppin, C.M., Manfredi, D., Richards, G., Watson, S.J., Geffner, M.E., Yong, J.E., Kim, R., Ross, H.M., Serrano-Gonzalez, M., Kim, M.S., 2020. No difference in 2D:4D ratio between youth with elevated prenatal androgen exposure due to congenital adrenal hyperplasia and controls. bioRxiv.

Neyse, L., Bosworth, S., Ring, P., Schmidt, U., 2016. Overconfidence, incentives and digit ratio. Sci. Rep. 6, 23294.

Neyse, L., Johanneson, M., Dreber, A., 2021. 2D:4D does not predict economic preferences: evidence from a large, representative sample. J. Econ. Behav. Organ 185, 390–401.

Oechssler, J., Roider, A., Schmitz, P.W., 2009. Cognitive abilities and behavioral biases. J. Econ. Behav. Organ 72 (1), 147–152.

Parslow, E., Ranehill, E., Zethraeus, N., Blomberg, L., von Schoultz, B., Hirschberg, A.L., Dreber, A., 2019. The digit ratio (2D: 4D) and economic preferences: no robust associations in a sample of 330 women. J. Econ. Sci. Assoc. 5 (2), 149–169.

Phoenix, C.H., Goy, R.W., Gerall, A.A., Young, W.C., 1959. Organizing action of prenatally administered testosterone propionate on the tissues mediating mating behavior in the female guinea pig. Endocrinology 65 (3), 369–382.

Puts, D.A., McDaniel, M.A., Jordan, C.L., Breedlove, S.M., 2008. Spatial ability and prenatal androgens: meta-analyses of congenital adrenal hyperplasia and digit ratio (2D: 4D) studies. Arch. Sex. Behav. 37 (1), 100.

Richter, D., Schupp, J., 2015. The SOEP Innovation Sample (SOEP IS). Schmollers Jahrb. 135 (3), 389–399.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366.

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., Marshall, A.C., 2021. Measuring cognitive reflection without maths: development and validation of the verbal cognitive reflection test. J. Behav. Decis. Mak. 34 (3), 322–343.

Skagerlund, K., Lind, T., Strömbäck, C., Tinghög, G., Västfjäll, D., 2018. Financial literacy and the role of numeracy–How individuals' attitude and affinity with numbers influence financial literacy. J. Behav. Exp. Econ. 74, 18–25.

Toplak, M.E., West, R.F., Stanovich, K.E., 2011. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Mem. Cogn. 39 (7), 1275.
van Anders, S.M., Vernon, P.A., Wilbur, C.J., 2006. Finger-length ratios show evidence of prenatal hormone-transfer between opposite-sex twins. Horm. Behav. 49 (3), 315–319.
Van Leeuwen, B., Smeets, P., Bovet, J., Nave, G., Stieglitz, J., Whitehouse, A., 2020. Do sex hormones at birth predict later-life economic preferences? Evidence from a pregnancy birth cohort study. Proc. R. Soc. B 287 (1941), 20201756.
Ventura, T., Gomes, M.C., Pita, A., Neto, M.T., Taylor, A., 2013. Digit ratio (2D: 4D) in newborns: influences of prenatal testosterone and maternal environment. Early Hum. Dev. 89 (2), 107–112.
Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., Wagner, G.G., 2013. Validating an ultra-short survey measure of patience. Econ. Lett. 120 (2), 142–145.
Voracek, M., 2011. Special issue preamble: digit ratio (2D: 4D) and individual differences research. Pers. Individ. Differ. 51 (4), 367–370.
Zhang L., & Ortmann A. (2013). Exploring the meaning of significance in experimental economics. Australian School of Business Research Paper No. 2013 ECON 32.