The impact of statistics training and education on reasoning performance

Pavle Valerjev (valerjev@unizd.hr)
Department of Psychology, University of Zadar, Croatia

Marin Dujmović (marin.dujmovic@bristol.ac.uk)
School of Psychological Science, University of Bristol, United Kingdom

Abstract

Recent studies have shown that analytical reasoning is related to a number of individual factors including IQ, lower conservativism, as well as other cognitive and personality factors. These studies have been broad, without aiming at specific influences on the development of analytical reasoning. The aim of this study was to determine whether education level and statistics training affect performance in reasoning tasks. Large samples from Croatia and the UK completed the Test of Statistical Reasoning (TSR), as well as a set of modified reasoning tasks. Results revealed that participants with some statistics training performed better in both the reasoning tasks and the TSR. The main finding was a country by education level interaction. Education level had a significant effect on both reasoning and TSR performance in the UK sample (higher level related to better performance), but a non-significant effect in the Croatian sample. An interesting finding was that Croatian participants performed significantly better than their UK counterparts at earlier stages of education, but then plateaued. UK participants reached the same level of performance at later stages. These differences may be explained by the breadth and depth of high school education in Croatia compared to the UK. Overall, statistics training and higher education levels relate to more analytical and statistical reasoning.

Keywords: *analytical reasoning; statistical reasoning; dual-process theories; education; statistics training*

Introduction

The tradition of researching reasoning within a dual-process approach has developed rapidly in the past decade with new findings and integration with metacognitive measures. Classic models, such as the default-interventionist view of dual-processing, relied on early empirical findings which found that heuristic-based reasoning was fast, cognitively undemanding and automatic. Analytic reasoning was defined as slow, deliberate and cognitively demanding (Evans, 2012). Thus, the two types of processes were named Type 1 and Type 2, while the systems are usually referred to as System 1 and System 2. Additionally, early findings seemed to indicate exclusivity – responses cued by System 2 were not available to System 1. Limiting cognitive resources by imposing a secondary task should then manifest as a substantial decrease in System 2 responses. However, recent findings using this approach have shown that analytical responses are given at only slightly lower frequency (Bago & De Neys, 2017; Lawson et al., 2020). Further evidence comes from studies relying on the rethinking paradigm. Participants are

required to give a quick initial response (sometimes under cognitive load via a secondary task). Then they are instructed to take time and rethink their responses. Findings show that the vast majority of analytical responses after rethinking come from trials in which the analytical response was already provided as the initial response (Dujmović et al., 2020; Thompson et al., 2011). Indeed, there are examples of rethinking increasing heuristic-based responses rather than analytical responses. The introduction of metacognitive measures (e.g., judgments of confidence) has expanded reasoning to what is known as the field of meta-reasoning. Evidence suggests participants can be just as quick and confident when giving analytical responses as when giving heuristic responses (De Neys & Pennycook, 2019; Dujmović & Valerjev, 2018). These findings have gradually led to more modern views on the dual-process nature of reasoning.

Considering that System 1 can generate both heuristic and analytical types of responses, modern theories posit that intuitive processes of various types exist. Some fit traditional heuristics, such as the representativeness heuristic or belief-based reasoning. However, others may be based on mathematical principles, probability estimates or formal logic (De Neys, 2012). A typical reasoning task will be designed to cue two responses generated by different processes. In the modern dual-process framework, it is presumed that different intuitive processes generate initial responses. These initial responses can be congruent – meaning both processes generate the same response, or they can be in conflict – they generate competing responses. When responses are congruent, there is no conflict and no uncertainty, and a final response can be given without more processing. When responses are conflicting, then depending on the level of uncertainty, System 2 may be triggered to resolve the situation. Initial responses will usually differ in strength, one may be dominant and, depending on the relative difference in strength, it will be more or less likely for System 2 to trigger. System 2 may then result in rationalization of the more dominant response, sampling more evidence for one or both responses, decoupling from the dominant response or in generating a novel response (De Neys, 2022; Pennycook, Fugelsang et al., 2015).

Early reasoning tasks, such as the Linda problem (Tversky & Kahneman, 1983) resulted in mainly quick heuristic responses, and very rare, slow analytical responses. This was due to the analytical intuitive response being relatively weak or not being generated at all. Stanovich (2018) proposes that *mindware* determines the strength of intuitive processes and responses. Mindware for any type of processing is acquired through a lifetime of experiences and learning. A strong mindware will make it more likely for a particular type of process to generate automatic, strong responses. Representativeness or belief-based responses are usually a product of strong mindware. On the other hand, responses based on estimates or computation of probability, formal logic or complex maths are usually weak, considering the mindware required for these types of processes is typically not as strong. The strength of intuitive responses has also been shown to be sensitive to the instruction type (Valerjev & Dujmović, 2017) and combining modalities in which information is presented (Dujmović & Valerjev, 2017). This indicates that attentional and cognitive resources play a mediating role between mindware and response strength.

There has been a rise in the number of studies which focus on individual differences in reasoning. Usually, the goal is to determine good predictors of analytical reasoning. This, in effect, is a search for correlates of strong analytical mindware. Results have shown that more analytical reasoners tend to be less religious (Gervais & Norenzayan, 2012), less susceptible to *pseudo-profound bullshit* (Pennycook, Cheyne et al., 2015), more intelligent (Kaufman, 2011), and differ compared to more heuristic-based reasoners on a number of other cognitive and personality traits. However, research into the impact of education has been missing from this line of inquiry. There are some studies that ask the question of how education in general and for specific domains transfers to general cognitive capacities, but not to reasoning from a dual-process perspective. Bunge and Leib (2020) review some of these findings, showing that there is a clear gap. First, the studies

mostly look at specific types of reasoning, like relational reasoning, or use IQ tests as shorthand for testing reasoning. And second, the impact of domain specific training mostly relates to very narrow interventions, e.g., tax reasoning skills or law school entrance exam preparation rather than broader statistics education. The aim of this study was to specifically determine whether education level and statistics training relate to reasoning performance on two sets of tasks. One set is contained in the Test of Statistical Reasoning (Rapan & Valerjev, 2020), and the other set is a combination of various classical reasoning tasks. Education, especially statistics training, should strengthen the type of mindware which leads to more analytical responses. This should be particularly pronounced for the statistical reasoning tasks (since the type of mindware and task match more closely), but should also be present for classic reasoning tasks, considering that performance is strongly correlated between the two sets of tasks (Rapan & Valerjev, 2021).

Methods

Design

The study is a 2(country of origin) by 2(type of reasoning task) quasi-experiment at its core. Two further independent variables are the highest completed level of education (high school, undergraduate, graduate or higher), and whether or not the participants received any statistics training (yes/no). Considering the sample size and group allocation, independent 2×2×3 and 2×2×2 analyses will be conducted, depending on whether general education level or statistics training is included as the third independent variable.

Participants

Participants from Croatia and the United Kingdom were recruited online. Croatian participants were recruited via Facebook groups and student contacts, while UK participants were recruited via the Prolific platform (see Table 1). A high proportion of Croatian participants were current undergraduate students, with quite a large cohort from Psychology departments, which accounts for disparities in completed education and proportion with statistics training between the two countries.

Table 1. Sample characteristics.

	Croatia	United Kingdom
N	292	298
M_{age} (*SD*)	28.38 (11.51)	31.42 (11.11)
Gender		
Male	27.74%	28.52%
Female	71.92%	70.47%
Other/refused	0.34%	1.01%
Education level		
High-school	52.05%	41.61%
Undergraduate	18.15%	40.93%
Graduate or higher	29.79%	17.45%
Statistics training	64.73%	17.79%

Materials

Participants completed a modified version of the Cognitive Reflection Test (CRT), one item each of a conflict Base Rate task (BR), the Linda problem, Covariation Detection task (CD), as well as the Test of statistical reasoning (TSR).

The CRT (Fredrick, 2005) is a widely used set of tasks designed to induce a misleading heuristic response and an analytical response. It originally consisted of three tasks with an open-ended response mode. Here, the task is modified both in terms of content and response mode. Four new tasks have been created due to overuse of the original. Additionally, the task was changed to a 4-alternative forced choice format, two of which were the heuristic and analytical responses, and the remaining two were distractors (Valerjev, 2019). An example can be seen below.

The ages of Mark and Andy add up to 28 years. Mark is 20 years older than Andy.
How old is Andy?
A) 4 B) 8 C) 6 D) 10

The modified BR task (De Neys & Glumicic, 2008; Dujmović & Valerjev, 2018) pits a response based on typicality of a stereotypical characteristic against the one based on probability. An example can be seen below.

Imagine a group of people consisted of 993 mathematicians and 7 hospitality workers (bartenders, hotel receptionists etc.). One person is chosen at random from that group of a 1000 people.

The person turns out to be very sociable and extroverted. Is it more probable that the person is a mathematician or a hospitality worker?

Probability indicates the person is much more likely to be a mathematician, but the personality traits are much more typical of hospitality workers.

The classic Linda problem pits the representativeness heuristic against probabilistic reasoning. The task content has been modified here, considering how well known the original has become.

John is a professional poker player and spent his college days studying computer science.

A) John is a member of an independent theatre group.

B) John has an above average IQ and is a member of an independent theatre group.

Participants had to indicate the probability of the two options (A and B), given the description of John. The description is more representative of option B, however, the response is incorrect as the conjunction of two events is less probable than either of the two events individually.

The CD task (Stanovich et al., 2016; Valerjev & Dujmović, 2019) pits absolute magnitudes against ratios. An example can be seen below.

Imagine clinical trials for a new COVID-19 vaccine have been started with human patients. Patients are assigned to one of two groups. One receives the vaccine, while the control group gets vitamin boosters. Outcomes are evaluated after 4 weeks, and the results are shown below.

	Negative	*Infected*
Treatment	*200*	*100*
Control	*150*	*50*

On a scale from -3 to 3 evaluate to what degree the vaccine correlates with a more positive outcome.

The absolute number of people testing negative for the virus 4 weeks after trials started is higher in the treatment group (200 vs 150). However, the ratio of negative to infected patients is better in the control group (3:1 vs 2:1). The heuristic response is for the vaccine to be correlated with the more positive outcome. The correct response is that it is negatively correlated with the more positive outcome.

The TSR consists of 11 tasks which do not pit heuristic and analytical responses against each other. Rather, it is a timed multiple-choice test primarily concerned with probability and proportion judgments. Each task is limited to 45 seconds (an example can be seen below).

A box contains 4 white, 6 blue and 8 black balls. A single ball is drawn. What is the probability that the ball is blue?
A) 25% B) 33.3% C) 50% D) 66.6%

The study was designed and conducted online via the PsyToolkit platform (Stoet, 2010).

Procedure

This study was part of a larger project in which participants answered a number of demographic-related questions and questions related to beliefs/behaviour concerning the COVID-19 pandemic. The presented reasoning tasks were completed in the order presented here as the final stage of that

study. The order of items within each task (for the CRT and TSR) was randomized for each participant. After they completed the TSR, participants gave a metacognitive judgment about their performance by indicating how many of the 11 tasks they solved correctly. Correct responses on the CRT, BR, Linda and CD tasks were summed and a percentage of correct responses computed for each participant. The same was done for the TSR.

Results

Performance for both TSR and reasoning tasks was normally distributed across groups (Table 2), and no further data processing was required prior to analysis.

Table 2. Performance descriptive statistics across tasks and groups

TSR			
	M (*SD*)	Skewness	Kurtosis
Overall	64.23 (19.52)	-0.29	-0.49
Croatia	65.41 (18.66)	-0.31	-0.43
UK	63.09 (20.29)	-0.26	-0.56
Reasoning tasks			
	M (*SD*)	Skewness	Kurtosis
Overall	49.58 (24.75)	0.02	-0.83
Croatia	52.89 (23.64)	-0.17	-0.75
UK	46.36 (25.43)	0.23	-0.77

In order to test whether country of origin, type of task and statistics training affected performance, a 2×2×2 mixed ANOVA was conducted (Table 3). The only two significant effects were a large effect of type of task with participants performing better in the TSR than in the set of reasoning tasks, and a small effect of statistics training, which indicates participants with at least some statistics training performed better (see Figure 1).

In order to test whether general education level affected performance, a 2(country)×2(task)×3(education) ANOVA was performed. Education levels were: completed high school, completed undergraduate program and completed graduate or higher level. Results can be seen in Table 4.

Table 3. Group, task and statistical training ANOVA of performance results

Effect	F (1, 588)	η_p^2	p
Country	0.08	<.01	.78
Task	163.84	.22	<.001
Statistics training	14.44	.02	<.001
Country*Task	1.65	<.01	.20
Country*Statistics training	3.56	<.01	.06
Task*Statistics training	0.77	<.01	.38
Three-way interaction	2.33	<.01	.13

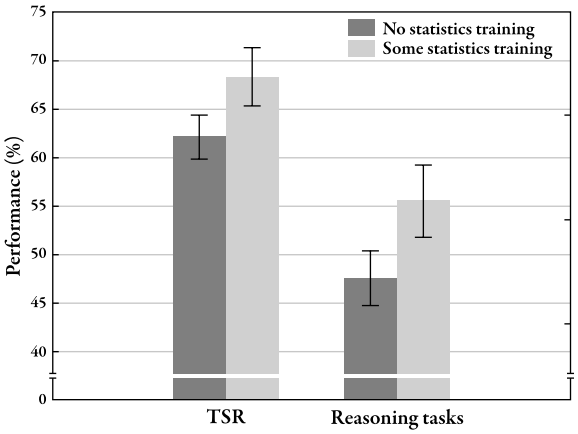

Figure 1. Performance depending on task and statistics training (error bars represent 95% CI)

Apart from the previously shown effect of task type, both education level and the country by education level interaction was significant. Participants with a higher completed level of education tended to have better performance in general (post-hoc Tukey HSD tests indicate this was due to a significant difference in performance between the highest level of education and the remaining two levels). However, the country by education interaction effect was also significant (Figure 2). The interaction effect is significant due to the difference in performance of participants from the two countries being significant only at the lowest achieved education level. In this study, Croatian participants were significantly better at this stage and further education was not accompanied by a significant improvement, while participants from the UK did improve to reach the same level of performance.

An additional analysis was conducted in order to determine whether the effect of statistical training was robust across all levels of education by

conducting a 2(task type) × 2(statistics training) × 3(education level) ANOVA. Apart from the effects already demonstrated, statistics training remained a significant factor ($F(1, 584) = 11.77$, $p < .001$, $\eta_p^2 = .02$) with no interaction effects. Therefore, statistics training seemingly increased performance, irrespective of education level.

Table 4. Group, task and education ANOVA of performance results

Effect	F (df_1, df_2)	η_p^2	p
Country	3.24 (1, 584)	<.01	.07
Task	225.71 (1, 584)	0.28	<.001
Education	7.01 (2, 584)	.02	<.001
Country*Task	3.60 (1, 584)	<.01	.06
Country*Education	4.41 (2, 584)	.01	.01
Task*Education	0.22 (2, 584)	<.01	.80
Three-way interaction	0.16 (2, 584)	<.01	.81

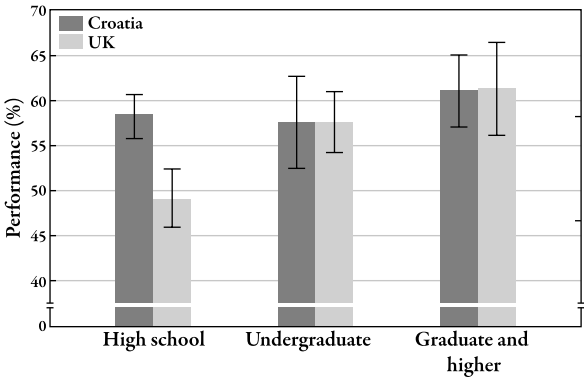

Figure 2. Performance depending on country of origin and achieved level of education (error bars represent 95% CI)

Finally, a metacognitive calibration score was computed for each participant, as the difference between the actual number of correct responses on the TSR and the metacognitive judgment of how many responses they felt were correct. First, a 2(country) × 2(statistics training) ANOVA resulted in no significant effects on metacognitive calibration (all $F(1, 588) < 1.74$, $p > .18$, $\eta_p^2 < .01$). However, the 2(country) × 3(education level) ANOVA revealed a slight but significant country by education level interaction effect ($F(2, 584) = 4.35$, $p = .01$, $\eta_p^2 = .015$). The interaction is a result of education level having no effect on calibration for participants from the UK, but a small to moderate effect on calibration for participants from Croatia ($\eta_p^2 = .04$). Croatian participants at the highest achieved education level exhibit better calibration than other groups (see Figure 3). Note that perfect calibration would have a score of 0 on this variable since it would indicate a perfect match between actual performance and perceived performance. Positive scores indicate underestimating and negative scores indicate overestimating performance. Therefore, all the groups tend to underestimate their performance.

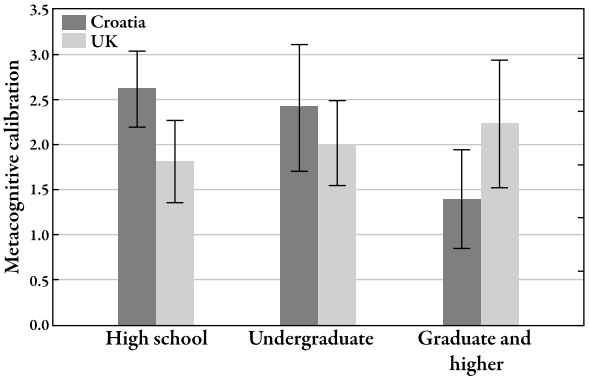

Figure 3. Metacognitive calibration depending on country of origin and education level (error bars indicate 95% CI)

Discussion

The aim of this study was to determine whether education level and, more specifically, statistics training contributed to the strengthening of analytical mindware, and related to better performance on a number of reasoning tasks. The findings show that both general education level and statistics training do indeed relate to increased performance in the TSR and reasoning tasks.

Our findings support the hypothesis that stronger mindware relates to the probability an individual is going to generate and select an analytical response. One possible explanation is that education and statistics training strengthen this mindware and, therefore, performance. However, given the nature of the quasi-experimental design, it is difficult to judge how much education and statistics training contribute to mindware. This would require a longitudinal approach. It is certainly the case that participants who already had developed stronger analytical processes are more likely to

choose to continue their education and study programs which require statistics training. Therefore, it is unknown how much the differences observed in this study are due to the differences occurring earlier in development, and how much due to specific education and training. Results depicted in Figure 2 may give further insight into this question. The key finding is that there is an increase in performance with the increase of education level in the UK, but not in Croatia. Considering how participants were sampled (see Methods), it is plausible that UK participants with a completed high school education include a larger proportion of those who did not continue with their education. The Croatian sample likely contains more undergraduate students who also have a high school education as their highest achieved level. If this is the case, then the fact that performance does not increase with education level in Croatia, and that there is a significant effect of country of origin only at the high school level, may indicate that the propensity toward analytical reasoning rather than education is driving the observed effect. Alternatively, the differences in performance patterns across countries may lie in the differences between the two education systems. The Croatian system is much more comprehensive, with much more information to assimilate and many more concepts to cover and understand. Education in the UK, due to available resources, makes it possible for students to make more choices, and the material is not covered in the comprehensive, old-fashioned manner, as it is in Croatia. This may force Croatian students to develop more analytical strategies for navigating the education system at an earlier age, leading to more of a plateau effect by the end of high school.

On the other hand, statistics training seems to have a robust effect regardless of education level. Though it was expected that participants with statistics training would exhibit a higher probability of choosing analytical responses, it was expected that performance on the TSR would be affected more than performance on the reasoning tasks. Considering the TSR aims to test statistical reasoning, and that statistics training instantiates this specific type of mindware, the effect should have been larger than for reasoning tasks. This is true even when taking into account that reasoning tasks certainly benefit from a strong statistics mindware, even though they include other types of analytical intuitions as well. It is important to note that the TSR is currently still under development in the hopes of becoming a part of a larger, easy–to-administer, reasoning battery. The results here indicate that additional aspects of statistical reasoning need to be covered by the instrument in the future, at which point performance will likely be influenced more by strengthening statistics-based mindware.

The fact that statistics training does not aid metacognitive calibration in the TSR task further indicates the need for modifications. On the other hand, overall education level does impact metacognitive calibration of Croatian participants, but not participants from the UK. Croatian participants tend to be more accurate at assessing their performance with the increase of education level (with a noted improvement between undergraduate and graduate level – see Figure 3). It is an interesting pattern. Croatian participants across education levels have similar levels of performance (Figure 2), but get better at assessing performance. On the other hand, performance of UK participants increases with education levels, but metacognitive calibration remains equally inaccurate across levels. This, again, may be caused by a number of factors. On one hand, Croatian higher education has numerous midterms and exams compared to higher education in the UK. The sheer volume of test taking and getting feedback may lead to better calibration in general. The alternative explanation would indicate that confidence simply increases as Croatian participants reach higher education levels, regardless of performance (as stated, performance is fairly constant across levels). In that case, calibration would improve as a side-effect of general increase in confidence rather than a more nuanced matching of performance and perception.

This preliminary study has uncovered many potentially interesting effects which will be informative for future research. First, the effect of education on analytical reasoning is likely much more complex than a simple increase in the propensity for analyt-

ical reasoning with higher education. For a more thorough analysis of how the education system may affect analytical reasoning, many more factors need to be measured and controlled (e.g., the area of study, exact current level attended by the participants, details about the particular education system etc.). Second, the TSR task did not benefit more from statistics training than other reasoning tasks, indicating more development is required. However, a more detailed inquiry, longitudinal, or action study in particular types of statistics training would lead to better understanding of how it contributes to development of statistical reasoning processes. Finally, interesting cultural differences may point to interesting opportunities from the perspective of higher education in Croatia. The results indicate that at the point of entering higher education, there may be a window of opportunity to better develop analytical skills than is currently the case.

References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90–109.

Bunge, S.A., & Leib, E.R. (2020). How does education hone reasoning ability?, *Current Directions in Psychological Science, 29*(2), 167-173.

De Neys, W. (2012). Bias and conflict a case for logical intuitions. *Perspectives on Psychological Science, 7*, 28–38.

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, E111.

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science, 28*, 503–509.

Dujmović, M., & Valerjev, P. (2017). An image is worth a thousand words, but what of numbers? The impact of multi-modal processing on response times and judgments of confidence in base-rate tasks. In O. Tošković, K. Damnjanović, & Lj. Lazarević (Eds.), *Proceedings of the XXIII science conference Empirical studies in psychology* (pp. 30-36). University of Belgrade.

Dujmović, M., & Valerjev, P. (2018). The influence of conflict monitoring on meta-reasoning and response times in a base rate task. *Quarterly Journal of Experimental Psychology, 71*(12), 2548-2561.

Dujmović, M., Valerjev, P., & Bajšanski, I. (2021). The role of representativeness in reasoning and metacognitive processes: An in-depth analysis of the Linda problem. *Thinking & Reasoning, 27*(2), 161-186.

Evans, J. S. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning, 18*(1), 5–31.

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

Gervais, W.M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science, 336*(6080), 493–496.

Kaufman, S. B. (2011). Intelligence and the cognitive unconscious. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence*. Cambridge University Press.

Lawson, M.A., Larrick, R.P., & Soll, J.B. (2020). Comparing fast thinking and slow thinking: The relative benefits of interventions, individual differences, and inferential rules. *Judgment and Decision Making, 15*, 660–684.

Pennycook, G., Cheyne, J.A., Barr, N., Koehler, D.J., & Fugelsang, J.A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making, 10*(6), 549–563.

Pennycook G., Fugelsang J.A., Koehler D.J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34–72.

Rapan, K., & Valerjev, P. (2020). Test statističkog rasuđivanja [Test of Statistical Reasoning]. In: V. Ćubela Adorić, I. Burić, I. Macuka, M. Nikolić Ivanišević, & A. Slišković (Eds.), *Zbirka psihologijskih skala i upitnika, Svezak 10*. University of Zadar.

Rapan, K., & Valerjev, P. (2021). Is automation of statistical reasoning a suitable mindware in

a base-rate neglect task? *Psihologijske teme 30*(3),447-466.

Stanovich, K.E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning, 24*, 423-444.

Stanovich, K.E., West, R.F., & Toplak, M.E. (2016). *The rationality quotient. Toward a test of rational thinking*. Cambridge, MA: The MIT Press.

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096-1104.

Thompson, V.A., Turner, J.A.P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*, 107–140.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.

Valerjev, P. (2019). Chronometry and meta-reasoning in a modified Cognitive Reflection Test. In K. Damnjanović, O. Tošković, & S. Marković (Eds.), *Proceedings of the XXV scientific conference Empirical studies in psychology* (pp. 31-34). University of Belgrade.

Valerjev, P., & Dujmović, M. (2017). Instruction type and believability influence on metareasoning in a base rate task In: G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *Proceedings of the 39th annual meeting of the Cognitive Science Society* (pp. 3429-3434). Cognitive Science Society.

Valerjev, P., & Dujmović, M. (2019). Performance and metacognition in scientific reasoning: The covariation detection task. *Psihologijske Teme 28*(1), 93-113.