Decision-making skills and styles as predictors of managerial performance

Erceg, Nikola

University of Zagreb

Faculty of Humanities and Social Sciences

Nikola Erceg

DECISION-MAKING SKILLS AND STYLES AS PREDICTORS OF MANAGERIAL PERFORMANCE

Doctoral Thesis

Zagreb, 2023

University of Zagreb

Faculty of Humanities and Social Sciences

Nikola Erceg

DECISION-MAKING SKILLS AND STYLES AS PREDICTORS OF MANAGERIAL PERFORMANCE

DOCTORAL THESIS

Supervisor: prof. dr. sc. Zvonimir Galić

Zagreb, 2023

Sveučilište u Zagrebu

Filozofski fakultet

Nikola Erceg

VJEŠTINE I STILOVI ODLUČIVANJA KAO ODREDNICE USPJEHA U POSLU RUKOVODITELJA

DOKTORSKI RAD

Mentor: prof. dr. sc. Zvonimir Galić

Zagreb, 2023

Podaci o mentoru

Titula, ime i prezime nastavnika: prof. dr. Zvonimir Galić

Naziv ustanove u kojoj je zaposlen: Filozofski fakultet, Zagreb

E-mail adresa i adresa osobne mrežne stranice: zgalic@ffzg.hr;

http://psihologija.ffzg.unizg.hr/zvonimir-galic;

https://www.researchgate.net/profile/Zvonimir-Galic

Životopis

Zvonimir Galić rođen je 9. lipnja 1980. godine u Požegi gdje je završio osnovnu školu i prirodoslovno matematičku gimnaziju. Studij psihologije diplomirao je 2003. godine na Filozofskom fakultetu u Zagrebu s prosjekom ocjena 5,00. Neposredno po završetku studija zaposlen je na Odsjeku za psihologiju Filozofskog fakulteta u Zagrebu u svojstvu znanstvenog novaka-asistenta pri Katedri za psihologiju rada i ergonomiju. Trogodišnji doktorski studij psihologije na istom je fakultetu upisao 2004. godine. Naziv doktora znanosti iz znanstvenog područja društvenih znanosti, znanstvenog polja psihologija, grana psihologija rada stekao je 2008. godine obranom doktorskog rada pod nazivom „*Nezaposlenost, traženje posla i zapošljavanje: longitudinalna analiza psiholoških aspekata*“. U znanstveno-nastavno zvanje docenta izabran je 2009. godine, a u znanstveno-nastavno zvanje izvanrednog profesora 2016. godine. U znanstveno zvanje znanstvenog savjetnika u području društvenih znanosti, polje psihologija izabran je 2021., a 2022. u znanstveno nastavno zvanje redovitog profesora. Trenutačno je predstojnik Katedre za psihologiju rada i ergonomiju na Filozofskom fakultetu u Zagrebu te suvoditelj (zajedno s prof. dr. Ninom Pološki Vokić) poslijediplomskog specijalističkog studija upravljanje ljudskim potencijalima.

Z. Galić usavršavao se na nekoliko međunarodnih institucija. Pohađao je 9., 10. i 11. Švicarsku ljetnu školu „*Metode u socijalnim znanostima*“ u Luganu, (2005., 2006. i 2007.) i tako usavršio primjenu složenih kvantitativnih metoda u društvenim znanostima. Sudjelovao je u radu 1. i 2. postdoktorske škole Europske asocijacije za psihologiju rada i organizacijsku psihologiju (2008. u Berlinu i 2010. u Valenciji) koje su bile posvećene novim spoznajama iz psihologije rada i organizacijske psihologije. Kao dobitnik Fulbrightove stipendije ak. godinu 2012./2013. proveo je u SAD-u na Sveučilištu Purdue, West Lafayette, a kao dobitnik Endeavour stipendije šest mjeseci tijekom ak. god. 2017/2018. proveo je u Australiji na Australian National University (Research School of Management), Canberra. 2017. i 2019. boravio je na kraćim istraživačkim boravcima na the Pennsylvania State University.

Glavni istraživački interesi Z. Galića vezani su uz procjenu ličnosti, odnos ličnosti i radnog ponašanja, psihosocijalne aspekte nezaposlenosti i kvalitetu radnog života, te rukovođenje i donošenje odluka u organizacijskom kontekstu. Sudjelovao je u radu više znanstvenih projekata koje su financirali MZOŠ i Sveučilište u Zagrebu, a vodio je i uspostavni istraživački projekt „*Implicit personality and work behaviour*“ Hrvatske zaklade za znanost. Trenutno je voditelj projekta „ *Implicit personality, decision-making and organizational leadership*“ kojeg također financira Hrvatska zaklada za znanosti. Nalaze svojih istraživanja priopćio je na brojnim konferencijama u zemlji i inozemstvu te ih objavio u uglednim domaćim i stranim časopisima. Član je uredništava poznatih međunarodnih časopisa: *Journal of Vocational Behavior, International Journal of Selection and Assessment i Journal of Personnel Psychology*.

Z. Galić na svom matičnom fakultetu predaje više različitih kolegija iz područja psihologije rada i organizacijskog ponašanja na preddiplomskoj, diplomskoj i doktorskoj razini. Također, nastavnik je i jedan od osnivača interdisciplinarnog specijalističkog poslijediplomskog studija Upravljanja ljudskim potencijalima na Sveučilištu u Zagrebu. Povrh toga, nastavu održava na doktorskom studiju Edukacijsko rehabilitacijskog fakulteta u Zagrebu te na poslijediplomskom specijalističkom studiju Medicine rada i športa Medicinskog fakulteta u Zagrebu. Mentor je brojnih diplomskih radova i više doktoranda. Član je Hrvatskog psihološkog društva te tri međunarodne strukovne udruge: European Association for Work and Organizational Psychology, European Network of Selection Researchers, Society for Industrial and Organizational Psychology. Ovlašteni je psiholog i član Hrvatske psihološke komore.

Z. Galić sudjelovao je u radu većeg broja znanstvenih projekata. Prvo je radio na dva projekta MZOS-a, koje je vodio prof. B. Šverko: “Ljudski potencijali u promjenjivom svijetu rada” (2003.-2006) i “Psihološki aspekti nezaposlenosti: longitudinalna studija“ (2006.-2009.). Potom je bio suradnik u nekoliko projekata koje je vodio prof. Ž. Jerneić: „Ličnost i socijalno poželjno odgovaranje“ (MZOS, 2009.-2012.), „Lažiranje odgovora u selekcijskim situacijama“ (Sveučilište u Zagrebu; 2013.-2014.), „Uloga osobina ličnosti u predviđanju različitih aspekata radne uspješnosti“ (Sveučilište u Zagrebu; 2015. godine), „Osobine ličnosti kao prediktori odgovornog i nepoželjnog organizacijskog ponašanja (Sveučilište u Zagrebu; 2016. godine) “Implicitna i eksplicitna ličnost kao odrednice radne uspješnosti” (Sveučilište u Zagrebu; 2017. godine), “Individualne razlike u motivaciji i vještinama donošenja odluka kao prediktori poslovnog i karijernog uspjeha hrvatskih iseljenika (Sveučilište u Zagrebu; 2018. godine) i “Identifikacija psiholoških odrednica proaktivnog radnog ponašanja i uspjeha u poduzetništvu” (Sveučilište u Zagrebu; 2019.; svi voditelj: prof. dr. Ž. Jerneić). Multimodalni indikatori lažiranja u

selekcijskom intervjuu za rukovodeću poziciju (2020., voditeljica doc. dr. M. Parmač Kovačić); „Mjerenje i unaprjeđivanje racionalnog prosuđivanja i donošenja odluka u poslovnom svijetu“ (2021., voditelj Z. Galić) i „Aktivno otvoreno mišljenje (AOM): sustavni pregled literature i eksperimentalna provjera “ozbiljne” računalne igre namijenjene poučavanju AOM-a“ (2022. voditelj Z. Galić). U ak. godini 2012. /2013. u suradnji s prof. J. M. LeBretonom sa Sveučilšta Purdue proveo je projekt „Predicting (dis)honesty at work: a cross-cultural study of two integrity measures”, koji je financirala Fulbrightova zaklada. Od 2014. do 2017. bio je voditelj uspostavnog istraživačkog projekta “Implicitna ličnost i radno ponašanje”, a od 2018. do danas voditelj je projekta „Implicitna ličnost, donošenje odluka i rukovođenje u organizacijama“. Oba projekta su projekti Hrvatske zaklade za znanost (HRZZ). Od 2018. do 2022. surađuje i na projektu HRZZ-a „Doprinos interne komunikacije uspješnosti organizacije: položaj, kanali, mjerenje i odnos s povezanim konceptima“ (voditeljica: prof. A. Tkalac Verčić, Ekonomski fakultet u Zagrebu).

Dosad je objavio jednu knjigu i više od 40 znanstvenih radova.

Abstract

The aim of this multi-paper dissertation (i.e. Scandinavian model PhD thesis) was to address several of the gaps identified in the literature related to the assessment of individual differences in reasoning and decision-making quality, sometimes also referred to as rationality.

In first two studies, we tested the construct validity of the cognitive reflection test (CRT) and its non-numerical counterpart belief-bias syllogisms (BBS; Study 1) and investigated different ways and approaches to solving these tasks (Study 2). We concluded that CRT and another numerical test that does not contain lures are factorially indistinguishable and that lures do not play role in the success on or predictive validity of the CRT or BBS (Study 1). Additionally, we showed that in many cases when CRT and BBS tasks are solved correctly, they are solved relying solely on intuition and not analytical thinking. We conclude that these two tests are therefore not particularly good measure of reflection and analytical thinking engagement as thought before.

In Study 3, across two studies, we examined a factorial structure of different cognitive biases tasks in a bid to understand the structure of rationality captured by these tasks and to validate identified rationality factor(s). In both studies, one-factor solution was the most appropriate and, to a large degree, explained by numerical abilities and thinking disposition called actively open-minded thinking (AOT; 61% of the rationality factor variance in Study 1 and 75% in Study 2). We conclude that cognitive biases tasks are highly heterogeneous and not particularly good solution for measuring rationality. Instead, we argue that individual differences in rational judgments and decisions matter are better captured by AOT.

In Study 4, across three samples (undergraduates, employed participants and entrepreneurs), we showed that decision-making styles matter for various real-life and work-related outcomes, often adding explanatory power over cognitive abilities and personality traits.

Finally, in Study 5, we examined the relevance of managerial AOT for positive organizational and employee-level outcomes. Over two studies, managers' AOT correlated positively with range of positive personal and organizational outcomes. We concluded that concluded that it would be worthwhile to focus on AOT in selection for leadership positions as well as teach current leaders about the benefits and implementation of this way of thinking.

In sum, results of our five studies point to the conclusion that how we think is as important, if not more, for consequential real-life outcomes, than how smart we are. What seem to be the most important is not to avoid decision-making and making them after the process of thinking in an actively open-minded way.
Keywords: decision-making skills; decision-making styles; rationality; actively open-minded thinking; managers.

Extended abstract in Croatian

Uvod

Cilj ove disertacije sačinjene od više radova (tzv. Skandinavski model) bio je pozabaviti se identificiranim nedostacima u literaturi iz domene individualnih razlika u kvaliteti razmišljanja i donošenja odluka, odnosno iz domene racionalnosti. Ovi nedostaci se prvenstveno odnose na pitanja valjanosti nekih od najčešćih načina mjerenja racionalnosti i na gotovo potpuni nedostatak istraživanja individualnih razlika u racionalnosti i kvaliteti donošenja odluka u kontekstu rukovođenja, području u kojem su racionalne prosudbe i odluke od iznimne važnosti.

Metodologija i rezultati pet istraživanja

U prvom istraživanju (n = 506 studenata) provjerili smo konstruktnu valjanost testa kognitivne refleksivnosti (eng. Cognitive Reflection Test; CRT), što je ustvari numerički test koji „mami“ ljude na davanje intuitivnih, ali netočnih odgovora, i silogizama s pristranošću uvjerenja (eng. Belief Bias Syllogisms; BBS) koji također „mame“ ljude nelogične, ali intuitivne odgovore. Iako se CRT najčešće naziva mjerom refleksivnosti ili sklonosti analitičkom mišljenju, neka prethodna istraživanja pokazala su da je visoko povezan s testovima kognitivnih sposobnosti, posebice numeričkih sposobnosti. Pitanje je, stoga, proizlaze li prediktivne sposobnosti ovog testa iz njegovih „mamaca“ ili samo iz činjenice da je dobra mjera kognitivnih sposobnosti. Koristeći pristup strukturalnog modeliranja, zaključili smo da su CRT i klasični numerički test (bez „mamaca“) faktorski identični, što znači da „mamci“ ne igraju ulogu u uspjehu na CRT testu kao ni u njegovoj prediktivnoj valjanosti. Slično tome, „mamci“ nisu bili zaslužni ni za prediktivnu valjanost BBS-a. Zaključak ovog istraživanja je da ova dva testa stoga nisu osobito dobra mjera kognitivne refleksivnosti kao što se prije mislilo.

CRT zadaci se često koriste za ilustriranje teorije dvostrukog procesiranja u praksi. Pritom se uobičajeno misli da se CRT zadaci rješavaju sekvencijalno: osoba prvo proizvede intuitivni, ali netočan odgovor na koji je „namamljena“, a tek zatim, ako uspije detektirati da je nešto neobično s ovim odgovorom, se može

angažirati u analitičkom mišljenju i pokušati izračunati točan odgovor. Međutim, u zadnjih nekoliko godina pojavila su se istraživanja koja su pokazala da nezanemariv broj ljudi CRT i slične zadatke rješava intuitivno točno, bez oslanjanja na analitičko mišljenje. Stoga je cilj našeg drugog istraživanja bio, koristeći dvije različite paradigme, provjeriti pretpostavku postojanja različitih načina rješavanja CRT zadatka, te ispitati koje sposobnosti, dispozicije i znanja stoje u podlozi tih različitih pristupa. Kroz dva istraživanja (n1 = 506, n2 = 83), pokazali smo da se u mnogim slučajevima CRT i BBS zadaci točno rješavaju oslanjajući se isključivo na intuiciju, a ne na analitičko razmišljanje. Ove intuitivno ispravne odgovore u pravilu daju ljudi najviših kognitivnih sposobnosti i najrazvijenije sklonosti prema analitičkom mišljenju. Naši rezultati potvrđuju hibridnu teoriju dvostrukog procesiranja ili tzv. teoriju „logičkih intuicija" i još jednom dovode u pitanje CRT i BBS kao dobre mjere analitičkog angažmana i refleksivnog razmišljanja.

Drugi uobičajeni način mjerenja racionalnog prosuđivanja su zadatci koji mjere sklonost različitim kognitivnim pogreškama i pristranostima prilikom prosuđivanja. Ovi zadaci imaju barem tri prednosti koje ih potencijalno čine dobrim mjerama individualnih razlika u racionalnosti: a) dobri su indikatori znanja koje je nužno da bi osoba mogla donositi kvalitetne prosudbe i odluke (npr. poznavanje logičkih principa, statistike, teorije vjerojatnosti itd.), b) dobri su indikatori kognitivnih sposobnosti koja je također važna za kvalitetno odlučivanje i prosuđivanje, i c) slično CRT zadacima, konstruirani su na način da „mame“ osobu na davanje intuitivnog, heurističkog odgovora, što ih čini i potencijalno dobrim indikatorima dispozicija koje su potrebne da prepoznamo „mamce“ i odupremo im se, a koje također igraju važnu ulogu u racionalnom prosuđivanju i odlučivanju. Upravno zbog ove zadnje karakteristike, neki autori smatraju da ovi zadaci mjere važne karakteristike koje klasični testovi inteligencije ne uspijevaju zahvatiti i da bi se na racionalnost trebalo gledati kao na konstrukt nadređen inteligenciji. Jedno od otvorenih pitanja u ovom području je dijele li različiti tipovi ovakvih zadataka dovoljno varijance koju bismo mogli proglasiti faktorom racionalnosti. Neka od prethodnih istraživanja ne idu u prilog jednom faktoru racionalnosti, ali ta istraživanja nisu bila bez svojih nedostataka.

Stoga smo u trećem istraživanju, na dva neovisna uzorka (n1 = 253, n2 = 210), ispitali faktorsku strukturu deset odnosno sedam zadataka kognitivnih pristranosti kako bismo razumjeli strukturu racionalnosti mjerenu ovakvim zadacima i validirali identificiran(e) faktor(e) racionalnosti. Jednofaktorsko rješenje bilo je najprikladnije na oba uzorka. Drugim riječima, identificirali smo "faktor racionalnosti" iako uspio objasniti samo mali dio varijance u ovim zadacima (12% u prvom i 22% u drugom istraživanju). Faktor racionalnosti bio je negativno povezan s (i)racionalnim vjerovanjima (npr. praznovjerno razmišljanje u

prvom istraživanju) te je čak bio u pozitivnoj korelaciji sa zadovoljstvom karijerom i poslom u drugom istraživanju. Također, numeričke sposobnosti i dispozicija mišljenja nazvana aktivnim otvorenim mišljenjem (eng. Actively Open-minded Thinking; AOT) objasnile su vrlo visok postotak varijance u faktoru racionalnosti (61% u prvom i 75% u drugom istraživanju). Zaključak trećeg istraživanja je, stoga, da su zadaci kognitivnih pristranosti izrazito idiosinkratični i heterogeni te kao takvi nisu najbolje rješenje za mjerenje individualnih razlika u racionalnosti. Umjesto toga, kako racionalnost u velikoj mjeri ovisi o AOT dispoziciji razmišljanja, smislenijim pristupom se čini individualne razlike u racionalnosti operacionalizirati kao sklonost AOT-u i mjeriti tastovima ili upitnicima koji mjere ovu dispoziciju.

U četvrtom istraživanju, slijedeći proizvoljnu ali korisnu podjelu individualnih razlika u kvaliteti donošenja odluka na vještine i stilove donošenja oduka, skrenuli smo pažnju s vještina (mjerenih zadacima maksimalnog učinka kao što su CRT, BBS i zadaci kognitivnih pristranosti) na stilove odlučivanja (mjerenih upitnicima tipičnog učinka). Stilovi odlučivanja mogu se definirati kao tipični obrasci ponašanja i doživljavanja u situacijama donošenja odluka. Najviše dokumentirana u literaturi i istraživana su racionalni i intuitivni stil odlučivanja. Međutim, postoje dodatni stilovi (izbjegavajući, ovisni i spontani) koji nisu toliko istraživani, posebno u radnom kontekstu. Stoga smo, u ovom istraživanju, na tri uzorka (n1 = 253 studenta, n2 = 210 zaposlenih sudionika, n3 = 53 poduzetnika) pokazali da su stilovi odlučivanja važni za različite ishode, kako u svakodnevnom životu, tako i na poslu, često ih objašnjavajući i povrh kognitivnih sposobnosti i osobina ličnosti. Pritom se izbjegavajući stil donošenja odluka posebno istaknuo kao prediktor mnogih negativnih ishoda u različitim kontekstima.

Konačno, u petom istraživanju smo se usredotočili na važnost racionalnog razmišljanja i odlučivanja kod menadžera i moguće pozitivne posljedice takve racionalnosti za organizacijske ishode i ishode na razini zaposlenika. Dva razmatranja su motivirala ovo istraživanje. Prvo, donošenje odluka smatra se jednom od najosnovnijih i najvažnijih zadaća rukovoditelja, dok su u rukovoditelji, istodobno, u tom poslu u prosjeku prilično neuspješni. Drugo, aktivno otvoreno mišljenje, kao važan indikator racionalnosti, moglo bi biti najbolji „lijek“ za neke od čestih menadžerskih pogrešaka u odlučivanju i prosuđivanju identificiranih u literaturi. Stoga smo, kroz dvije studije (n1 = 124 menadžera i 190 njihovih podređenih, n2 = 126 menadžera i 335 njihovih podređenih), mjerili razinu AOT-a menadžera i korelirali je s različitim procjenama njihovih podređenih. AOT menadžera pozitivno je korelirao s njihovom kvalitetom donošenja odluka i intelektualnom poniznošću procijenjenima od strane njihovih podređenih, kao i s ocjenama podređenih o psihološkoj sigurnosti u radnim timovima te zadovoljstvom poslom i percepcijom organizacijske podrške koju doživljavaju podređeni. Zaključak petog istraživanja je da

racionalnost na vodećim pozicijama može donijeti mnoge pozitivne posljedice za tvrtke i njihove zaposlenike te da bi se vjerojatno isplatilo fokusirati na AOT prilikom odabira za vodeće pozicije, kao i na poučavanje postojećih rukovoditelja o prednostima i primjeni ovakvog načina razmišljanja.

Zaključak

Zaključno, rezultati naših pet istraživanja upućuju na zaključak da je način na koji razmišljamo jednako važan, ako ne i važniji, nego kognitivne sposobnosti za bitne poslovne i životne ishode. Ono što se čini najvažnijim je ne izbjegavati donošenje odluka i donositi ih procesom aktivno otvorenog mišljenja. Naši rezultati sugeriraju da bi AOT mogao biti najbolja konceptualizacija i definicija racionalnog prosuđivanja i donošenja odluka te da uobičajeni zadaci koji se koriste u literaturi za mjerenje racionalnosti (npr. CRT ili zadaci kognitivnih pristranosti) imaju ozbiljne nedostatke. Stoga, umjesto korištenja tih zadataka, bolje bi bilo procjenjivati individualne razlike u racionalnosti putem mjerenja sklonosti AOT-u. U prilog tome govore i rezultati na rukovoditeljima gdje se njihova sklonost AOT-u pokazala blagotvornom za organizacije i njihove zaposlenike.

Ključne riječi: stilovi odlučivanja; vještine odlučivanja; racionalnost; aktivno otvoreno mišljenje; rukovođenje.

Table of Contents

1. INTRODUCTION 1

2. STUDY 1: A REFLECTION ON COGNITIVE REFLECTION – TESTING CONVERGENT/DIVERGENT VALIDITY OF TWO MEASURES OF COGNITIVE REFLECTION 13

Introduction 13

Methods 17

Results 22

Discussion 26

Conclusion 29

3. STUDY 2: WHO DETECTS AND WHY – HOW DO INDIVIDUAL DIFFERENCES IN COGNITIVE CHARACTERISTICS UNDERPIN DIFFERENT TYPES OF RESPONSES TO REASONING TASKS? 31

Introduction 31

Study 1 42

Methods 42

Results 45

Study 1 discussion 55

Study 2 58

Methods 58

Results 61

Study 2 discussion 71

General discussion 71

Conclusion 76

4. STUDY 3: NORMATIVE RESPONDING ON COGNITIVE BIAS TASKS - SOME EVIDENCE FOR A WEAK RATIONALITY FACTOR THAT IS MOSTLY EXPLAINED BY NUMERACY AND ACTIVELY OPEN-MINDED THINKING 77

Introduction 77

Study 1 83

Methods 83

Results 92

Study 2 98

Methods 99

Results 103

Discussion 108

Conclusion 113

5. STUDY 4: INCREMENTAL VALIDITY OF DECISION-MAKING STYLES IN PREDICTING REAL-LIFE AND WORK-RELATED OUTCOMES 114
Introduction 114
Study 1 118
Methods 119
Results 120
Study 2 121
Methods 122
Results 124
Study 3 130
Methods 130
Results 132
General discussion 135
6. STUDY 5: TESTING THE THEORY OF GOOD THINKING AND DECIDING IN ORGANIZATIONAL SETTING - MANY BENEFITS OF LEADER'S ACTIVELY OPEN-MINDED THINKING 139
Introduction 139
Study 1 142
Methods 142
Results 145
Study 2 147
Methods 147
Results 150
Discussion 152
Conclusion 155
7. GENERAL DISCUSSION 157
8. CONCLUSION 161
LITERATURE 163
APPENDIX A 183
APPENDIX B 217

1. INTRODUCTION

"Humans are cognitive misers because their basic tendency is to default to processing mechanisms of low computational expense." This sentence comes from recent but influential article by Stanovich (2018, p. 424) on origins and repercussions of human miserliness. What it means is that, unless prompted or encouraged, people will basically tend to superficially process information and respond with a first thing that comes to their mind. These prompts can come in several ways. For example, there are times when it is clear that relying on intuition or "gut feeling" is not an option. When solving a complex math problem (e.g. 17x34) or an IQ test in a process of job selection, it is clear that engagement in analytical thinking and deeper processing is needed, if nothing because almost nobody has a ready intuition or "gut feeling" to follow. So, the lack of intuition is by itself a trigger to engage deeper with the tasks.

However, there are situations where the case is not so clear-cut. For example, oftentimes we possess intuitions that can be faulty and if we do not recognize the need to override them and engage in deliberative thinking, we will make a mistake (e.g., come to wrong conclusion, make a wrong decision, behave irrationally). For example, a couple might choose to buy a house based on one salient attribute that aligns with their stereotype of a nice home ("wow, the ceilings are so high and there is so much light"), but fail to check the quality of installations or the prices of other houses in vicinity. Or a CEO might put too much weight on own "gut feeling" about merging with other company based on hunches that are unrelated with the decision at hand. Furthermore, modern society is full of external agents that try to profit on our intuitive decisions and reactions. Basically, whole advertisement industry exists to lure us into not thinking too much about money we spend.

Clearly, in these situation different people will react differently. Some will not even realize that they should perhaps pause and think more carefully but will mindlessly continue initial course of action or decision. Some will perhaps detect that there are conflicting paths and that more thinking could be warranted. However, even then, not everyone will have motivation to engage in additional thinking, or ability and knowledge to pull it off. Of course, some people will do everything the right way – they will recognize that their "gut feeling" might tempt them into wrong conclusion or decision, engage in additional thinking and overturn their initial impulses. What this all means is that there are individual differences in motivation and ability to engage in deliberate, analytical thinking and that these differences can have significant repercussions in various domains of life.

Some authors claim that these individual differences, although crucial for good decision-making and rational behavior, are not particularly well represented in current personality and cognitive abilities models. Therefore, current IQ tests do not capture these rationality-related traits and we need different kinds of tasks to properly capture them (e.g. Stanovich, 2009a, 2012; Stanovich et al., 2016). Another repercussion of inadequate attention given to these traits is that they have basically been completely missing from some areas in which good decision-making is crucial. For example, although most leadership models list decision-making as one of the crucial leadership skills (e.g. Bartram, 2005; Dierdorff & Rubin, 2006; Tett et al., 2000), almost none of the previous work tried to capture these traits in leaders and investigate their relevance for personal and organizational outcomes.

Therefore, this dissertation aims to identify and address several gaps in the literature related to measures for assessment of rationality or decision-making quality, as well as in applying those measures in predicting real life outcomes such as, for example, those from organizational context. This work was done across five different studies described in subsequent chapters. However, before moving on, I will give a brief introduction about the current state of the art in the reasoning and decision-making literature, as well as describe gaps that were identified and addressed in subsequent studies.

Current model for understanding reasoning errors

Recently, Pennycook (2023) published the most comprehensive framework for describing and understanding reasoning errors that also make it clear what kind of individual differences can play role in such errors. This framework draws on dual process theories that characterize human thinking and reasoning as an interplay of fast, automatic, autonomous and non-conscious System 1, and slower, rule based, effortful deliberate System 2 (DeNeys, 2012; 2015; Evans & Stanovich, 2013; Kahneman, 2011). It also incorporates all of the previous models of the dual process theories (e.g. default-interventionist model, parallel activation model and hybrid logical-intuitions model described in Chapter 4), providing a more complete picture of human reasoning and the one that is best supported by current data.

This framework (originally reported in Pennycook et al., 2015) presumes that reasoning plays out in three subsequent stages (i.e. the three-stage model). In the first stage, System 1 generates several intuitions about the possible responses. These responses emerge autonomously as a direct result of a stimulus-response pairing which could either be evolved (e.g. fear response that emerges from seeing a snake) or

something that has been learned (e.g. stereotypes). As these intuitions are generated autonomously, there can be several different intuitive responses.

A second stage is the metacognition stage. As we can have multiple autonomous outputs in parallel, then there will be cases where these outputs suggest competing responses or actions (Pennycook, 2023). For example, a person deciding whether to buy a flat can simultaneously feel attracted to it and have gut reaction to buy it (e.g. high ceilings and lots of light) and to move on (e.g. high price or bad location). What recent studies show (e.g. Bago and De Neys, 2017; 2019; 2020) suggest is that the difference in the strength of these competing intuitions are crucial for a person to detect the conflict between them and engage in more analytical processing. Zvonimir, for example, has always loved to read lifestyle magazines that have heavily influenced his idea of ideal home (spacious and light). However, he is also price sensitive, meaning that both of the above intuitions will be strong in Zvonimir, increasing the likelihood that Zvonimir will detect the conflict between them and pause to reflect more carefully on the decisions. In contrast, Nikola gives little weight to home esthetics as he never liked lifestyle magazines and always has a price in mind anyways, so he basically lacks the first intuitive response about the beauty of the place. Therefore, when faced with the high price, Nikola will probably intuitively reject the option to pursue a nice apartment, never even feeling the need to weight pros and cons of such apartment. In other words, the strength on competing intuition in Nikola's case would be so startling low that he would never detect the conflict and engage in analytical thinking. The decision to move on would be made quickly and without thinking. Of course, this example is excessive simplification of reality, but it serves the purpose of illustrating how differential intuition strength encourages or discourages detection of the conflict between the intuitions and subsequent analytical thinking and deeper processing. Crucially, this also means that human cognition requires a conflict monitoring system in order to detect these issues so that they may be resolved with subsequent processing (Pennycook, 2023).

Finally, if the conflict between competing intuition is detected in second stage, this triggers analytic thinking in the final, third stage. This analytic thinking can come in two forms. A "good" form is cognitive decoupling, which means that an individual engaged analytic thinking to inhibit and override a prepotent intuitive answer and in this way came up with a better, more correct answer (or decision, behavior etc.). However, actively suppressing prepotent response and generating an alternative requires additional cognitive resources and given that people are cognitive misers (as indicated previously), we can expect that in some instances this process will fail. Alternative and easier process of analytical

thinking, the "bad" one, is called rationalization (although, this is also a simplified view; cf. Cushman, 2020; DeNeys, 2020). Rationalization occurs when System 2 is triggered, but the individual simply focuses on bolstering the initial intuitive response, and not questioning or overturning it.

The three-stage framework, in addition to being the most comprehensive account or reasoning currently, provides an insight into different ways a reasoning can go wrong and different traits that underpin success or failure in reasoning. Basically, reasoning can fail at any of the three stages. For example, strong but incorrect intuition can lead a person to commit reasoning or decision-making error without noticing any conflict or engaging in analytical thinking, which would represent a first-stage error. Additionally, if there are competing intuitions, conflict between them could exist, but metacognitive monitoring could be weak leaving the conflict undetected and, thus, failing to trigger additional deliberative thinking. This would be a second-stage error. Finally, there are reasoning errors that can happen in the final, third stage. Here, a person could engage deliberative thinking for purposes of rationalization instead of decoupling, i.e. to bolster initial incorrect intuition instead of questioning it. However, reasoning error is possible even if a person engages in decoupling and overturning initial incorrect intuition. This could happen if a person lacks ability, disposition or knowledge necessary for sustained reasoning and normative responding.

A tripartite theory of mind and the difference between rationality and intelligence

Crucially, what these different types of possible errors show is that, looking from the perspective of measurement and assessment, we currently lack instrument and measures capable of capturing abilities and dispositions important for successful reasoning and decision-making. In the last few decades, a number of papers, both empirical and conceptual, advocated for broadening of the study of cognitive abilities by including concepts and constructs from the domain of decision-making (Baron, 1985; Stankov, 2017; Stanovich, 2009a, 2009b, 2012; Stanovich & West, 1998, 2000, 2008). There have been some indications that tasks that measure different cognitive biases capture something other than "classical" intelligence, a construct that is labelled as rationality (e.g. Stanovich & West, 1998, 2000, 2008; Stanovich et al., 2016).

Theoretical reasons for distinguishing between intelligence and rationality are presented in Stanovich's (2009a, 2012) tripartite theory of mind. This theory differentiates between autonomous, algorithmic and reflective parts of the mind. It also explains why different tasks intended to capture susceptibility to

cognitive biases actually assess broader set of abilities and dispositions than classical IQ tasks. According to it, in order to successfully solve the majority of cognitive biases tasks, a person first has to overcome initial incorrect response generated by the autonomous mind. In other words, a person has to reflect on his/her response and recognize the need to engage in more deliberate processing (reflective mind's task) and also possess adequate ability and computational power to calculate or come up with a correct response (algorithmic mind's task). Conversely, success on classical intelligence tests do not depend so much on the reflective, but only on the algorithmic mind, constituting this as the crucial difference between the two. In this framework, the reflective mind refers more to different thinking dispositions, while algorithmic mind refers more to cognitive capacities or abilities in the narrower sense (e.g. fluid intelligence; Stanovich, 2012). In this conceptualization of intelligence (i.e. what the usual intelligence tests measure), intelligence is practically not dependent on dispositions but mostly on capacities. From this, it follows that rationality captured by cognitive biases tasks is a broader construct than intelligence, as it is more dependent on thinking dispositions, and therefore it makes sense to conceptually differentiate between the two (e.g. Stanovich, 2009b, 2012).

Given previous discussion about different ways that reasoning can go wrong, we can see that reasoning can fail not only because someone lacks the cognitive ability needed for knowledge acquisition, successful decoupling and normative responding, but also because a person can lack disposition and motivation to, for example, question intuition, be careful and reflective or engage in deeper processing. Thus, if we are going to properly capture individual differences in rationality and reasoning and decision-making quality, we have to move beyond merely measuring individual differences in cognitive capacities or IQ. There have been significant improvements in this regard in recent decade, however there are still important issues that need to be addressed, both in terms of validity and practical usefulness of such measures.

Decision-making skills and styles

Dalal and Brooks (2013) proposed that, when it comes to decision-making properties on the individual level, it is possible to differentiate between the decision-making skills and decision-making styles. By skills, they primarily mean skills to resist succumbing to different types of cognitive biases, while the styles would be more related to usual ways of thinking when making judgments or decisions. This division is not completely accurate, but useful. Inaccurate because success on essentially every objective task meant to capture some cognitive property will depend both on cognitive capacity and on thinking

dispositions (e.g. Baron, 1985), and this is especially true for cognitive biases tasks (as discussed above). Therefore, it is not possible to strictly differentiate between style or disposition and skill or capacity. However, it is useful because it makes it easier to think about different ways we can measure traits relevant for good reasoning and decision-making. Basically, we can use objective tasks that have normative correct and incorrect responses (i.e. decision-making skills) and we can use subjective items that ask participants to state their opinions about proper ways of thinking or to describe how they typically think or behave in decision-making situation (i.e. decision-making styles).

Decision-making skills

Here I will describe some of the tasks that are widely used to assess individual differences in the domain of reasoning and decision-making. Stanovich et al. (2016) have proposed that these tasks differ along the two dimensions: a) how much they depend on successful conflict detection and override, and b) how much they depend on domain knowledge. Crucially, even though there are substantial differences among the tasks in regard to domain knowledge that person must possess in order to correctly solve them, basically all of these tasks are moderately or highly dependent on successful conflict detection and override. This is precisely, as discussed previously, what makes them different from classical IQ tests and more suitable for measuring dispositions important for quality reasoning and decision-making.

Perhaps the most known of these tasks are the tasks from the Cognitive reflection test (CRT; Frederick, 2005). This test has been enormously popular in the literature and the paper that introduced it was cited 5892 times at the time of writing. One of the most famous problems from this test is the "bat and a ball" problem from that goes as follows: „A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. Ho w much does the ball cost?" Similarly as other problems from this test, this one automatically triggers relatively strong initial response (i.e., 10 cents). However, after a more careful reflection, it becomes clear that the right response is in fact 5 cents. Several things contributed to its popularity – it is short, consisting of only three items (although it has been extended since its introduction), and attractive because of intuitive lures. However, it also elegantly illustrates main points of the tripartite theory of mind. Namely, it is posited that in order to overcome the initial wrong response generated by the fast and automatic System 1 (10 cents) and arrive at the correct one (5 cents), one has to reflect on the answer and recognize the need to engage in a more deliberate processing (the reflective mind), but also to possess adequate computational power, knowledge and abilities to calculate the right answer (algorithmic mind). The fact that it requires detection of the conflict between intuitive, but incorrect, and correct response

and override of this incorrect response that made it a perfect candidate task for assessment of crucial dispositions for quality reasoning and decision-making, those missed by more classical IQ tasks.

Other candidate tasks for measurement of decision-making skills and dispositions are tasks that assess susceptibility to different cognitive biases. In one of the first attempts to measure decision-making quality, Parker and Fischoff (2005) came up with a measure of decision-making competence (DMC), consisting of seven different behavioral decision tasks (consistency in risk perceptions, recognizing social norms, resistance to sunk cost, resistance to framing, applying decision rules, path independence and overconfidence). DMC was able to predict important real-life outcomes such as an index of different life outcomes that are most likely result of person's flawed judgment (Bruine de Bruin et al. 2020), psycho-social difficulties (Weller et al., 2015) and childhood delinquency and the number of sexual partners (Parker et al., 2018). Similarly, Stanovich et al. (2016) created the Comprehensive Assessment of Rational Thinking (CART), which is basically a composite measure of large number of different cognitive bias tasks, accompanied with numeracy measure and some additional thinking disposition measures. Composite measures of tasks that access cognitive biases similar to CART were also shown to shown to predict a composite score of real-world outcomes across several different domains (electronic media use, secure computing, substance use, driving behavior, financial behavior and gambling; Toplak et al. 2017). These effects remained event after controlling for the effects of intelligence. These results suggest that CRT and different cognitive bias tasks could be ideal tasks for measurement of important reasoning and decision-making dispositions missed by classical IQ tasks.

Decision-making styles

Decision-making styles are generally defined as learned, habitual response patterns exhibited by an individual when confronted with decision situations (Scott and Bruce, 1995). One of open questions here is the question of number of styles that can adequately capture various ways in which people approach decision-making process. The propositions range from two (e.g., rational decision style and intuitive decision style; Hamilton et al., 2016) to seven styles (e.g., vigilant, intuitive, spontaneous, dependent, anxious, brooding, and avoidant decision-making style; Leykin and DeRubeis, 2010).

Dewberry et al. (2013b) proposed a framework that distinguished between differences in cognitive processes that people use to make decisions (captured by rational/vigilant and intuitive styles) and regulatory processes concerned with choice regulation (captured by avoidant, dependent, or anxious

style). Styles concerned with cognitive processes in decision making (rational/vigilant, intuitive and spontaneous) align with well-known dual-process theory (Kahneman, 2011) that differentiates between System 1 (intuitive, automatic, associative, fast) and System 2 (analytic, explicit, rule-based, relatively slow). Consistent with the distinction between the two systems, intuitive and spontaneous styles seemed to indicate the extent to which individuals rely on heuristic/System 1 processing, while rational/vigilant style pertains to deliberate/System 2 processing. However, the remaining styles are not related with the two systems of information processing but more with the regulation of choice - the extent to which decisions tend to be delayed or avoided (avoidant style), referred to others (dependent style) or followed by negative affect (anxious style). The main insight here is that the feeling of anxiety over making decisions underpins the three regulatory styles.

One of the most known models of decision-making styles seems to be the one proposed by Scott and Bruce (1995). Their model is broad enough to encompass both styles related with cognitive processes and those related with regulation processes. They proposed five different decision-making styles: rational (a tendency towards thorough search for and logical evaluations of alternatives), intuitive (an inclination to rely on hunches and feelings), dependent (a propensity to search for advice and direction from others), avoidant (a proclivity to avoid decision making) and spontaneous (a sense of urgency to finish decision-making process as soon as possible). According to this model, these five styles can adequately capture the breadth of approaches to decision-making among individuals.

However, there seems to be at least one another style or thinking disposition that is highly indicative of quality, rational reasoning and decision-making. It is called Actively Open-minded Thinking (AOT; Baron, 2000; 2019; Baron et al., 2015). In short, AOT can be defined as the disposition to actively search for and give fair treatment to possibilities other than the one we initially favor. Therefore, its main difference from rational thinking style is that it deals not only with the quantity of thinking (i.e. whether a person tends to think a lot before making important decisions), but also the quality, or the direction of thinking (i.e. whether person tends to look for counterevidence and reasons why he/she might be wrong which is a direct antidote to some prevalent reasoning errors). Empirically, this thinking disposition was found to be related to a range of indicators of rational thinking - it correlates negatively with a wide range of cognitive biases such as confirmation bias, sunk cost effect, outcome bias, belief bias and others (Stanovich & West, 1997; Stanovich et al., 2016; Toplak et al., 2014) and with endorsement of epistemically suspect beliefs such as conspiracy, superstitious or paranormal beliefs (Pennycook et al.,

2020; Svedholm & Lindeman, 2013; Svedholm-Häkkinen & Lindeman, 2018), while correlating positively with accuracy on a variety of judgments, such as forecasting world events (Mellers et al., 2015) or distinguishing between real and fake news (Bronstein, et al., 2019). Therefore, AOT seems to be one of the key decision-making styles or dispositions that make someone a better reasoner and decision-maker.

Gaps in the literature and the current line of research

There are multiple gaps in the literature related to the assessment of individual differences in reasoning and decision-making, both in terms of validity of the measures and in terms of practical usefulness. The current line of research tried to address some of those gaps. For example, although CRT is often referred to as a measure of reflection or analytical thinking, indicating a common stance among researchers that it primarily taps into dispositions related to conflict detection and override, it is also shown to correlate very highly with cognitive abilities, especially numeracy (e.g. Liberali et al., 2012; Campitelli & Gerrans, 2014; Finucane & Gullion, 2010; Primi et al., 2016; Thomson & Oppenheimer, 2016; Welsh et al., 2013). Some recent studies even indicated that CRT might be barely factorially distinguishable or even totally undistinguishable from the other cognitive ability measures (e.g. Attali & Bar-Hillel, 2020; Blacksmith et al., 2019). If this is true, that would mean that the lures, a defining characteristic of CRT tasks, do not in fact make it any different from other tasks without such specific items. This would also call into a question a conclusion that the CRT is a measure of reflection and analytical thinking engagement as the suppression and override of intuitive response that the person is lured into is precisely why CRT was thought to tap into these dispositions. Therefore, Chapter 2 of this thesis describes a study in which we investigated the convergent and divergent validity of two measures of cognitive reflection using the structural equation modelling (SEM) approach.

Another issue with the CRT, as well as other tasks from the cognitive biases tradition, is that there are different ways to solving them correctly. A traditional view is that a person typically solve these tasks through additional deliberation, after engaging resources to override the intuitive response and to come up with the correct one. However, it is also possible that some individuals can solve these tasks intuitively correctly. In other words, some individuals could, for whatever reasons, have correct intuitions and be correct on such tasks without ever engaging in deeper processing and deliberation. In fact, this is what some of the recent studies found. Not only that some individuals are capable of responding intuitively correctly, but it turns out that this way of responding is far more prevalent than the traditional way of

engaging in deeper processing and deliberation (e.g. Bago and De Neys, 2017, 2019). If majority of people solve CRT-like items correctly through intuition, than again the score on CRT cannot be indicative of someone's reflection or analytical thinking. Thus, in Chapter 3, we describe two studies in which we, using two different paradigms, set out to investigate the prevalence of different types of responding to CRT-like problems and the individual differences in abilities, dispositions and knowledge that underpin these different types of responses.

In addition to these open questions about the validity of the CRT tasks, there are also gaps in understanding the other types of tasks related to decision-making skills, namely tasks that tap into different cognitive biases. One of the issues here is related to the question of structure or dimensionality of such tasks. Just as different cognitive ability tasks share a common core called the g-factor, could it be that different tasks that measure resistance to cognitive biases also have something in common, something that could be called the rationality factor. Theoretically speaking, as these tasks are supposed to be dependent both on dispositions related to conflict detection and override of intuitive responses and to intellectual capabilities needed for normative responding, they should have something in common. So, a person that possesses these dispositions and intellectual capabilities should be better at solving all of these tasks compared to a person without such traits. However, previous literature on this matter gives little support for this position. Few studies that examined a factorial structure of different cognitive bias tasks mostly failed to establish a single factor underlying the performance on variety of such tasks. What is more, there is little consistency in terms of underlying structure of these tasks across the studies, with tasks being scattered across different dimensions in different studies (e.g. e.g. Aczel et al., 2015; Berthet, 2021; Berthet & de Gardelle, 2021; Ceschi et al., 2019; Teovanović et al., 2015; Weaver & Stewart, 2012). These results raise serious questions about the construct validity of cognitive bias tasks – what underlying constructs do these tasks tap into and is there any communality among them? In Chapter 4, I describe two studies that we conducted with the aim of investigating the existence of rationality factor(s) and examining construct and predictive validity of such factor(s).

In addition to these concerns regarding the validity of tasks assessing reasoning and decision-making skills, there is also a large gap in the literature regarding the practical value of decision-making skills and styles, especially in the context of workplaces and organizations. Despite the agreement about the importance of good judgments and decision making in organizations, the field of work and organizational psychology paid very little attention to the research on leadership judgments and decision-making (JDM).

Only recently, several researchers called for cross-fertilization between JDM and the fields of work and organizational psychology (e.g., Dalal et al., 2010; Moore & Flynn, 2008; Tapia & Gaddis, 2017). We responded to such calls with the last two studies of this research program. In Chapter 5, I describe three studies in which we set out to investigate the validity of five decision-making styles (Scott & Bruce, 1995) across three different contexts and samples (undergraduates, employees, and entrepreneurs). We were specifically interested in finding out whether the decision-making styles would be able to predict important work-related outcomes even after accounting for the effects of cognitive abilities and personality traits on these outcomes.

Finally, we turned our attention to organizational leaders and their decision-making skills and styles. Specifically, we examined the effects of managers' actively open-minded thinking on different individual, team and organizational level outcomes. Given that most of the competency-based models of managerial work put decision-making at the forefront of the managerial duties (e.g., Bartram, 2005; Dierdorff & Rubin, 2006; Tett et al., 2000), and that plethora of studies showed that managers are bad decision-makers (e.g. Lovallo & Sibony, 2010; Nutt, 2002), succumbing to a range of decision-making errors that this disposition would counteract (e.g. Ketchen & Craighead, 2022; Sibony, 2020), it is remarkable that the concept of actively open-minded thinking is practically non-existent in the managerial literature. Thus, in Chapter 6, I describe two studies in which we examined managers' tendency to think in actively open-minded way and related it to a range of positive organizational outcomes.

2. STUDY 1: A REFLECTION ON COGNITIVE REFLECTION – TESTING CONVERGENT/DIVERGENT VALIDITY OF TWO MEASURES OF COGNITIVE REFLECTION

This chapter was previously published as: Erceg, N., Galić, Z., & Ružojčić, M. (2020). A reflection on cognitive reflection–testing convergent/divergent validity of two measures of cognitive reflection. *Judgment and Decision making*, *15*(5), 741-755.

Introduction

To make a rational decision, frequently we need to take time to deliberate, question the idea that first comes to mind and reflect on the available information before deciding. This principle lead Frederick (2005) to construct a short three-item measure in which every question was designed in a way that triggers an intuitive, impulsive answer that is always incorrect. In order to resist reporting the (inaccurate) response that first comes to mind, it is presumed that a person needs to „reflect“ on it and engage in slower and more deliberate thinking that is required to realize the correct response. Because of this characteristic, the test was named the Cognitive Reflection Test (CRT). In his seminal paper, Frederick reported that for the majority of students the CRT was quite hard, in spite the fact that it requires only basic mathematical skills to be correctly solved. The CRT was also shown to be related to different measures of cognitive abilities and analytic cognitive style, but the correlations were low enough to allow the conclusion that the CRT and other used cognitive measures „likely reflect common factors, but may also measure distinct characteristics, as they purport to“ (Frederick, 2005, p. 35).

Since then, the CRT became popular among researchers because of its brevity and the fact that it was able to predict an incredibly wide range of cognitive and behavioral outcomes. Specifically, CRT has been found to predict performance on a range of tasks from the heuristics and biases (H&B) domain. For example, the CRT score was negatively correlated with susceptibility to the conjunction fallacy and conservatism in updating probabilities (Oechssler, Roider & Schmitz, 2009), the base rate fallacy (Hoppe & Kusterer, 2011), and positively correlated with a general indicator of resilience to using mental shortcuts, as indicated with a composite of 15 different H&B tasks, including sample size problem, gambler's fallacy, Bayesian reasoning, framing problem, sunk cost and others (Toplak, West & Stanovich, 2011). Moreover, the predictiveness of the CRT spans outside the cognitive domain. CRT was found to be a significant predictor of religious belief (Pennycook, Cheyne, Seli, Koehler, and Fugelsang 2012; Shenhav, Rand & Greene, 2012), political orientation (Deppe et al., 2015; Pennycook

& Rand, 2019), science understanding (Shtulman & McCallum, 2014, Gervais, 2015), moral reasoning (Paxton, Ungar & Greene, 2012; Royzman, Landy & Goodwin, 2014) and susceptibility to pseudo-profound bullshit statements (Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2015; see Pennycook, Fugelsang and Koehler (2015) and Pennycook and Ross (2016) for a detailed account of predictiveness of the CRT across different domains).

Such breadth of the CRT bears the following question: where does this predictivity of the CRT come from? On the one hand, the CRT might be such a potent predictor because, similarly to some other non-lure measures (e.g. numeracy), it assesses different cognitive capabilities (i.e., abilities in a narrow sense, as discussed in Baron, 1985) and thinking dispositions that substantially account for performance on different tasks that the CRT predicts. For example, CRT was found to be highly correlated with "general cognitive ability" (e.g. Blacksmith, Yang, Behrend and Ruark 2019; Frederick, 2005) as well as with numerical ability (Campitelli & Gerrans, 2014; Finucane & Gullion, 2010; Liberali, Reyna, Furlan, Stein & Pardo, 2012; Primi et al., 2016; Thomson & Oppenheimer, 2016; Welsh, Burns and Delfabbro, 2013). To a certain extent, the CRT also assesses thinking dispositions, broadly defined as the tendencies towards particular patterns of intellectual behavior (Tishman & Andrade, 1996). One example is reflection/impulsivity (R/I), disposition to be careful at the expense of speed so those that are reflective are willing to sacrifice the efficiency and speed in responding in order to be more accurate (Baron, 2018; Baron, Scott, Fincher and Metz (2015); Baron, Gürçay & Metz, 2017). This view also follows from the results that show positive correlation between response time and accuracy on the CRT (e.g. Frey, Johnson & De Neys, 2017; Stupple, Pitchford, Ball, Hunt & Steel, 2017) and, in this regard, CRT might not be especially different from other tasks in which slower and more careful responding can lead to more accurate responses. Therefore, the traits that influence performance on any cognitive task that asks for both ability and deliberation (either with or without lures), might account for the predictive potency of the CRT.

On the other hand, the CRT has a distinctive characteristic of luring participants into incorrect intuitive responses that, allegedly, need to be detected and overridden in order to come up with correct response responsible. Some authors believe that this characteristic of the test should be mostly responsible for predictive potency of the CRT. In this regard, it is said that the CRT measures some additional ability or disposition, not shared with non-lure measures, to resist reporting a first response that comes to mind (Frederick, 2005), something that might be termed cognitive miserliness (Stupple et al., 2017; Toplak et

al., 2011; Toplak, West & Stanovich, 2014). Thus, this additional ability or disposition could be responsible for CRT's correlation with various outcomes.

Therefore, the key question is whether the lures make the CRT "special" or can some other, non-lure tasks predict the same outcomes to a similar degree. Several recent studies argue that the lures or the disposition to reflect and correct the intuitive wrong response are not important for the predictive power of CRT. For example, Baron et al. (2015) concluded that there is no evidence that "intuitive lures" matter at all for reliability or predictive validity of the CRT. A final piece of evidence that the lures do not account for the predictive potency of CRT comes from a recent study by Attali and Bar-Hillel (2020). Across two studies, they showed that the latent CRT factor and numerical factor formed with items without lures were correlated so highly that they were practically factorially indistinguishable. Their data showed that the predictive power of the CRT items came from their quality as math items and not from their "lureness". This result goes against the usual interpretation of CRT as a measure of some additional dispositions uniquely assessed by lures and shows that the lures are not the reason why CRT predicts performance on different cognitive tasks as well as various real life outcomes. Thus, in our study we decided to constructively replicate (Lykken, 1968) these findings using different set of CRT and well as math problems.

Current study

In our study, we investigated are the lures responsible for the correlations that the CRT has with different outcomes. To strengthen our constructive replication of Attali and Bar-Hillel (2020) study, in addition to CRT, we also used syllogisms that assess belief bias (belief bias syllogisms, BBS) as additional measure of cognitive reflection. Similarly to the CRT, BBS also trigger intuitive but incorrect response that needs to be detected and overridden in order to give a correct response. In other words, BBS items have lures but, unlike CRT, do not require participants to know math to solve them. Baron et al. (2015) showed that BBS are valid cognitive reflection items and they have been shown to predict performance on H&B tasks similarly as the CRT (West, Toplak & Stanovich, 2008). As non-lure tasks we used numeracy tasks (Cokely, Galešić, Schulz, Ghazal, & Garcia-Retamero, 2012) and verbal reasoning items (Condon & Revelle, 2014).

In order to accomplish study aims we did three things. First, we correlated our lure and non-lure measures with different tasks from the H&B domain (base-rate neglect, four card selection, causal base rate,

gambler's fallacy and availability bias tasks) and a thinking disposition measure (AOT questionnaire). We chose these H&B tasks because the cognitive reflection measures should be uniquely suited for predicting them, better than the non-lure measures. This view follows from the tripartite theory of mind (Stanovich, 2012; Pennycook, Fugelsang, & Koehler, 2015a) that differentiates between autonomous, algorithmic and reflective parts of the mind. The bat-and-ball CRT problem elegantly illustrates this: „A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" This problem automatically triggers relatively strong initial response (i.e., 10 cents). However, after a more careful reflection, it is clear that this is an incorrect answer, and that the right response is in fact 5 cents. Thus, in order to overcome the initial wrong response (generated by the autonomous mind), and arrive to the correct one, one has to first reflect on the answer and recognize the need to engage in a more deliberate processing (the reflective mind), but also to possess adequate computational power, knowledge and abilities to calculate the right answer (algorithmic mind). Stanovich, West and Toplak (2016), in their categorization of rationality tasks according to their dependence on the conflict detection/knowledge, put both cognitive refection tasks and H&B tasks on the same high level of dependence on the conflict detection dimension. That means that both of these two types of tasks cue intuitive but incorrect responses that need to be detected and overridden (reflective mind) if the task is to be solved correctly.

Conversely, according to the tripartite theory, non-lure tasks, or the tasks that do not depend on the conflict detection (such as tests of fluid intelligence), should capture only algorithmic mind and not the dispositions towards analytic/reflective thinking that are unique to the tasks high on the conflict detection dependence (Stanovich, 2009, 2012; Pennycook, Fugelsang, & Koehler, 2015). Thus, because cognitive reflection and H&B tasks have this common characteristic of triggering intuitive incorrect response and non-lure tasks do not, correlations between these two types of tasks should be greater than correlations between the non-lure and H&B tasks.

Second, we aimed to replicate Attali and Bar-Hillel (2020) who showed that one-factor model that did not differentiate between CRT items and ordinary math problems showed excellent fit to their data. They concluded that CRT items are essentially high quality math items and that the CRT's predictive value stems from the fact that it captures, what they called, "mathematical ability" (p. 95). In other words, the CFA suggested that the fact that the CRT items have lures did not ensure that they capture different construct than the regular math problems. In the current study, we seek to constructively replicate their

results with different sets of CRT and math problems. As non-lure math problems we are using The Berlin numeracy test (BNT; Cokely et al, 2012). This measure of statistical numeracy is particularly good test of convergent/discriminant validity of the CRT because BNT successfully predicted similar outcomes as CRT such as the ability to evaluate and understand risks (Cokely et al., 2012), maximization of expected value on monetary lotteries (Sobkow, Olszewska, & Traczyk, 2020), financial literacy (Skagerlund, Lind, Strömbäck, Tinghög, & Västfjäll, 2018) and performance on some of the H&B tasks (e.g. sunk cost, framing, base rate neglect, gambler's fallacy, etc.; Allan, 2018; Ghazal, S., 2014). There is also evidence that both BNT and CRT assess similar thinking dispositions related to deliberation, reflectiveness and actively open-minded thinking (Baron et al., 2015; Cokely, Feltz, Ghazal, Allan, Petrova, & Garcia-Retamero, 2018; Cokely & Kelley, 2009; Ghazal, Cokely, & Garcia-Retamero, 2014). Therefore, it is not surprising that several previous studies that investigated both CRT and BNT reported very high correlations between the two (e.g. Cokely et al. (2012) reported the correlation of r = .56 (disattenuated r = .93), Skagerlund et al. (2018) reported correlation of r= .61 (disattenuated r = 1) and Sobkow et al. (2020) reported correlation of r = .59 (disattenuated r = .90)). Taken together these results indicate that BNT as a non-lure math measure is perfectly suited for a replication of Attalli and Bar-Hillel (2020) result that the CRT and non-lure math problems load on the same factor. This would be another evidence against the importance of lures in predicting various outcomes.

Finally, to make our conclusions about the importance of lures more robust and expand on Attali and Bar-Hillel findings, we tested the importance of lures for predictiveness of BBS tasks. If BBS and BNT predict H&B tasks for the same reasons (i.e. not because of lures), than the correlations between the BBS and the H&B tasks should be greatly diminished once we statistically account for the effect of BNT in these tasks.

Methods

Participants

506 undergraduate University of Zagreb students (67% Faculty of humanities and social sciences students, mostly psychology students, and the rest from various other University of Zagreb faculties), participated in the study (27% males). The mean age was 21.2 (min = 18, max = 31, SD = 2.13).

Instruments

a) Cognitive reflection tasks

We used two different measures of cognitive reflection, the numerical one that required certain levels of mathematical skills to come to the correct responses and the verbal one and BBS that do not require any mathematical knowledge.

We used an expanded, 10-item version of the CRT in order to increase reliability and response range of the total score. It consisted of three original CRT items (Frederick, 2005), but also additional items from previously reported alternative CRT measures (Primi et al., 2015; Thomson & Oppenheimer, 2016; Toplak et al., 2014). An example of an item is "*In an athletics team, tall members are three times more likely to win a medal than short members. This year, the team has won 60 medals so far. How many of these have been won by short athletes?*". Here, the intuitive incorrect answer is 20 and the correct one is 15. All the items are listed in the Appendix. Total score was calculated by summing the correct responses, thus one could score anywhere between 0 (if none of the responses were correct) and 10 (if all the responses were correct).

BBS tasks assess the cognitive reflection by examining the susceptibility to belief bias. An example task goes as follows: "Premise 1: All flowers have petals. Premise 2: Roses have petals. Conclusion: Roses are flowers." (Markovits & Nantel, 1989). According to this syllogism, it does not follow that only flowers have petals, so roses might as well be something other than flowers (e.g. children collage art). However, because the conclusion that roses are flowers conforms with our empirical reality, it is quite believable and many people accept it as valid. Thus, the false intuitive response is the product of believability of the conclusion, while strong conformity with logical principles is needed to come up with the right, logically valid response. In addition to the "Roses have petals" example we used three additional syllogisms whose conclusions were believable, albeit logically incorrect (see Appendix for all the tasks). We considered as correct the response where participants identified believable conclusion as logically incorrect. Participants' scores ranged between 0 and 4.

b) Non-lure cognitive ability tasks

We used The Berlin numeracy test (BNT; Cokely et al., 2012) as a measure of numeracy. The BNT is a four-question test for assessing numeracy and risk literacy. An example of a question is "Imagine we are

throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?". The questions are designed in a way that they gradually become harder and a total score is calculated by summing up the correct responses on the four questions (see Appendix for all the items).

Verbal Reasoning (VR) was measured with four items taken from the International Cognitive Ability Resource (ICAR; for details see icar-project.com and Condon and Revelle, 2014). VR items include different logic, vocabulary and general knowledge questions. All of the items are presented in the Appendix.

c) Thinking dispositions

In this study we used a 15-item AOT scale introduced by Campitelli and Gerrans (2014) as a measure of thinking disposition. It is a self-report scale where participants indicate their level of agreement with the items on a six-point scale (1 – strongly disagree to 6 – strongly agree). An example of an item is "It is OK to ignore evidence against your established beliefs" (all of the items are in the Appendix). The total score on this scale is calculated as a mean level of agreement with the items and can be anything between 1 and 6.

d) H&B tasks

Four-card selection problem

We used five different tasks that had the same structure (all of the items are presented in the Appendix). A rule was explicitly stated for each of the items and participants were informed that the rule may or may not be correct. Their task was to check the accuracy of the rule by turning two cards of their choice. For example, one of the items was: "Rule: If a card shows "5" on one side, the word "Excellent" is on the opposite side. Which two cards would you choose to turn to check the accuracy of this rule?". Participants were then showed four cards that had numbers 5 and 3 and words "Excellent" and "Good" written on the front side. The correct answer here would be to turn the card containing number 5 and word "Good" because turning only these two cards would allow one to conclude whether the rule is correct or false. However, because the card with word "Excellent" is present, participants could be lured to turn it instead of the card "Good", although for the rule to be correct it does not matter what is behind the "Excellent" and "3" cards (Nickerson, 1998). Picking the two accurate cards to turn would be scored as 1 so the minimum score on this task was 0 while the maximum was 5.

Base-rate neglect

Base-rate neglect task consisted of five similar problems where the description of a person was contrasted to the base-rate information. Specifically, there were two possible answers, a stereotypical one (based on the description of a person) and a base-rate consistent one. For example, one of the items was: "Among the 1000 people that participated in the study, there were 50 16-year-olds and 950 50-year-olds. Helen is randomly chosen participant in this research. Helen listens to hip hop and rap music. She likes to wear tight T-shirts and jeans. She loves to dance and has a small nose piercing. Which is more likely? a) Helen is 16 years old; or b) Helen is 50 years old."

Here, the description of Helen was stereotypical for a teenager. Thus, a person who heavily relies on this information would respond with an "a". However, a base-rate information indicated that there is much greater probability that randomly chosen participant is indeed a 50 years old. Thus, a response "b" was coded as a correct one. However, it has to be noted that technically this does not have to be a correct response and that this depends on the diagnosticity of the information in the task (e.g. the information could be that Helen is a minor which would render a base-rate based response incorrect[1]). Nevertheless, as the stereotypical response is intuitive response on these tasks and one needs to engage in correcting this intuitive response in order to accompany a base-rate information into a judgment (Barbey & Sloman, 2007; Pennycook, Fugelsang, & Koehler, 2012), we always coded a response based on base-rates as a correct one. The correct responses were scored as 1 and the theoretical range of scores was 0 to 5.

Causal base-rate

In the causal base-rate, participants are provided with two conflicting pieces of information: one is statistical and favors one decision while another is based on personal, case-based experience and favors another decision (Toplak et al., 2011; Stanovich et al., 2016). We present one of the items we used here, and report all three in the Appendix:

Professor Kellan, the director of a teacher preparation program, was designing a new course in human development and needed to select a textbook for the new course. She had narrowed her decision down to one of two textbooks: one published by Pearson and the other published by McGraw. Professor Kellan belonged to several professional organizations that provided Web-based forums for its members to share information about curricular issues. Each of the forums had a textbook evaluation section, and the

1 We thank a reviewer Guillermo Campitelli for this observation.

websites unanimously rated the McGraw textbook as the better choice in every category rated. Categories evaluated included quality of the writing, among others. Just before Professor Kellan was about to place the order for the McGraw book, however, she asked an experienced colleague for her opinion about the textbooks. Her colleague reported that she preferred the Pearson book. What do you think Professor Kellan should do?

a. She should definitely use the Pearson textbook; b. She should probably use the Pearson textbook; c. She should probably use the McGraw textbook; d. She should definitely use the McGraw textbook.

Here preference for the McGraw textbook indicates a tendency to rely on the large-sample information in spite of salient personal testimony. A preference for the Pearson textbook indicates reliance on the personal testimony over the large-sample information. Each item was scored one to four. In this case, one point is given if a participant thinks that a) She should definitely use the Pearson textbook while four points are given if participant thinks that d) She should definitely use the McGraw textbook.

Gambler's fallacy

Gambler's fallacy refers to the tendency for people to see links between events in the past and events in the future when the two are really independent (Stanovich et al., 2016). Consider the following problem which is one of the five we used (see Appendix for all the problems):

"When playing slot machines, people win something about 1 in every 10 times. Julie, however, has just won on her first three plays. What are her chances of winning the next time she plays?

____ out of ____."

Here the correct answer is 1 out of 10 (it was scored as 1, while all the other responses were scored as 0). However, people that are prone to gambler's fallacy would reason that, since Julia already won three times in a row, her probability of winning again would somehow need to be lower than 1 in 10. This does not make sense as slot machine does not remember Julia's previous outcomes and always presents outcomes with the same 1/10 probability. We measures gambler's fallacy with five items. We scored correct responses as 1 and incorrect as 0 and the theoretical range of results was 0 to 5.

Availability bias

Availability heuristic refers to assessing the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind (Tversky & Kahneman, 1973). Availability or the ease of retrieval certain instances of events is often influenced by the vividness or media exposure and does not necessarily correspond to the true frequency of such instances. For

example, people might think that homicide is much more common cause of death than the diabetes (it is the other way round; this was one of the questions we asked our participants) because homicides are often covered in media while diabetes complications and deaths are rarely discussed publicly. In this study, we followed a paradigm introduced by Lichtenstein, Slovic, Fischhoff, Layman and Combs (1978), by asking participants which of the four pairs of lethal events is more common. Choosing causes of death that are more vivid and more covered in media is a sign of over-reliance on easily available and retrievable information, i.e. availability heuristic (Pachur, Hertwig, & Steinmann, 2012; Stanovich et al., 2016). Thus, we refer to responses that follow from the availability heuristic even in situations when this does not correspond to reality as the availability bias. We scored the correct responses as 1 and incorrect (based on the availability heuristic) as 0. Thus, one could score between 0 (greatest availability bias) and 4 (lowest availability bias).

Procedure

Participants solved all the tasks as a part of a larger data collection effort in which they also solved a number of additional tasks that were not part of the current study. The regular and verbal CRT items were presented in four fixed, but different sequences and these sequences were randomly distributed across participants. All the other instruments were solved in fixed order. The students filled-in the tests and questionnaires on computers, in groups of 20 to 25 participants under the supervision of the investigators. Participants were reimbursed with course credits and/or cinema card vouchers. The whole testing session lasted up to two hours with a break of 10 to 15 minutes in the middle of a session. Upon reaching half of our planned sample (N = 253) we changed some of the measures and added some additional measures, mostly related to H&B tasks. This is why all the analyses involving H&B tasks are done on the remaining half of the sample (N = 253).

Results

To answer our first question whether the tasks with lures exhibit greater correlations with H&B and thinking disposition tasks than our non-lure tasks, we calculated correlation coefficients among all our variables. We report these correlations along with descriptive statistics and G6 reliability coefficient in Table 1.

Table 1. Descriptive statistics and correlations among all the variables. The G6 reliabilities are shown in the diagonal, bivariate correlations are below the diagonal, correlations between the latent factors are above the diagonal.

	M	SD	Min	Max	CRT	BBS	BNT	VR	AOT	BRN	FCS	CBR	GF	AV
CRT	5.59	2.91	0	10	.92	.66	.93	.77	.26	.35	.34	.35	.04	.19
BBS	2.10	1.62	0	4	.55**	.93	.68	.54	.32	.33	.25	.30	-.04	.13
BNT	1.56	1.12	0	4	.58**	.42**	.61	.80	.26	.41	.20	.41	-.03	.28
VR	3.50	0.81	0	4	.46**	.33**	.36**	.64	.22	.27	.30	.40	.17	.45
AOT	4.51	0.65	1.87	6	.22**	.28**	.17**	.14**	.83	.26	.21	.23	.25	.18
BRN	2.71	1.86	0	5	.30**	.30**	.26**	.15*	.25**	.92	.25	.40	.27	.20
FCS	1.53	1.48	0	5	.28**	.22**	.14*	.19**	.19**	.22**	.86	.22	-.08	.14
CBR	8.88	1.52	4	12	.22**	.20**	.19**	.17**	.16*	.28**	.12*	.45	-.01	.40
GF	4.09	1.04	0	5	.05	-.03	-.01	.11	.20**	.20**	-.04	-.01	.76	.16
AV	2.72	1.19	0	4	.11	.11	.14*	.13*	.11	.19**	.12	.23**	.03	.79

Note.* p< .05, p < .01; CRT – Cognitive reflection test; BBS – belief bias syllogisms; BNT – Berlin numeracy testy; VR – verbal reasoning; AOT – actively open-minded thinking; BRN = base-rate neglect; FCS – four cards selection task; CBR – causal base-rate; GF – gambler's fallacy; AV – availability bias.

In order to estimate the relationships among the variables while accounting for the measurement error, we calculated the correlations between the latent factors and reported them in the upper part of the Table 1, above the diagonal. Prior to that, we made sure that a one-factor structure fits each of our instruments well and that all of the items load sufficiently on their respective factors. We report the details of the analyses and fit indices for each of the factors in the Table A1 in Appendix. In short, for each of the factors, a one-factor solution proved to be a very good fit. Most of the loadings were much higher than .30, in fact only three of the total number of loadings did not pass this cut-off: a) on VR factor, the first item had loading lower than .30; b) on GF factor, first variable had loading lower than .30; c) on AV factor, first item had loading lower than .30. Thus, we can conclude that majority of our items are appropriate manifest indicators of their respective latent factors and that it is appropriate to do further analyses on these factors.

By looking at the upper part of the correlation table, two things are apparent. First, CRT and BNT factors correlate so highly (r = .93) that it is appears that these two factors are empirically indistinguishable. Second, both our lure (CRT and BBS) and non-lure measures (BNT and VR) show moderate to high correlations with thinking disposition and most of the H&B measures. In fact, these correlations are remarkably similar and it does not appear that our data support the expectation that the lure measures are related more with H&B tasks than the non-lure measures. In fact, BNT factor correlated more strongly with three H&B factors (BRN, CBR and AV factors) than either CRT (test for differences in correlations: z = 2.75; p = .00 for BRN and CBR; z = 5.56, p = .00 for AV) or BBS factor (z = 1.73, p = .04 for BRN; z = 2.36, p = .01 for CBR; z = 4.32, p = .00 for AV). CRT factor did not even correlate higher than BNT with the other measure of cognitive reflection (i.e., BBS), even though the two are allegedly measuring the same ability/disposition to resist reporting initial, intuitive responses. The only case that a lure measure correlated more than BNT with an outcome was of the CRT - FCS correlation (z = 6.17; p = .00). However, even here this correlation did not surpass the correlation between VR factor (another non-lure measure) and FCS (z = 0.99, p = .16). Thus, judging from the correlation matrix, it does not seem that the lures gave either CRT or BBS additional predictive power over the non-lure measures.

In the next two analyses, we investigated whether the CRT and BNT are factorially indistinguishable and whether the lures are responsible for the correlations between BBS and H&B tasks. Specifically, if BBS predicts H&B tasks for the same reason BNT predicts them (i.e. because the abilities and thinking dispositions not related to lures that are important for all three types of tasks and the lures are not so important), then the correlation between the BBS and the H&B tasks should be greatly diminished once we statistically account for the effect of BNT in these tasks. To assess these parameters free from error and to control for the Type 1 errors, we used CFA and SEM methods (Westfall & Yarkoni, 2016).

To test whether the CRT and BNT are factorially distinguishable, we compared a model where the correlation between the latent CRT factor and latent BNT factor was freely estimated with the one where the correlation was fixed at 1 (meaning that both CRT and BNT items loaded on a single factor). Both models showed excellent fit to the data ($\chi^2(76) = 57.07$, p = .95; *CFI* = 1; *TLI* = 1; *RMSEA* = .00 for the two correlated factors model and $\chi^2(77) = 58.61$, p = .04; *CFI* = .1; *TLI* = .1; *RMSEA* = .00 for the one factor model) and there was no significant differences in the fit between the models ($\Delta\chi^2(1) = 1.54$, p = .22) indicating that the latent factor of cognitive reflection is practically indistinguishable from the latent

factor of numeracy. To check whether the CRT items factor loadings on this single factor are related with lureness of the items, we calculated the correlation between the loadings and the lureness index. We calculated the lureness index for each of the items as a proportion of intuitive responses in all incorrect responses on that specific item (we report the lureness of each of CRT items in Table A2 in Appendix). The relationship between the loadings and the lureness is pictured in the Figure 1 from which it is clear that the lures are not the reason why the items loaded on single CRT – BNT factor that fitted the data best (r = -.08, p = .82).

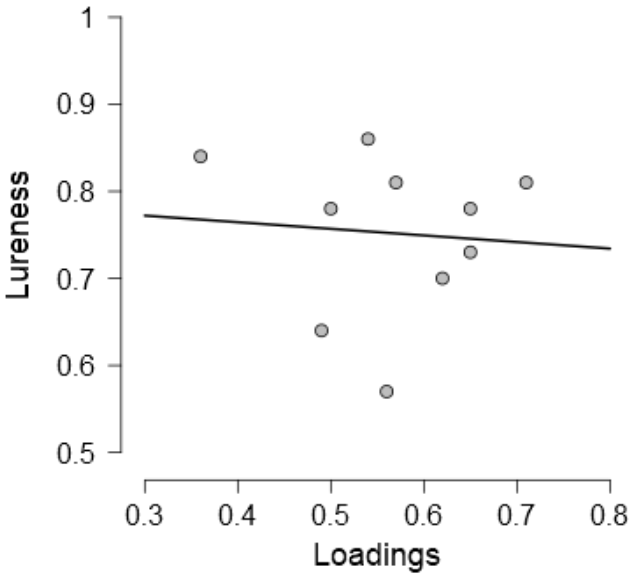


Figure 1. Relationship between the CRT items loadings on a single CRT - BNT factor and the lureness index of items

To further strengthen our findings, we explored how mathematical models developed by Campitelli and Gerrans (2014) to assess CRTs construct validity fitted to our data. In short, Campitelli and Gerrans developed three models which they called mathematical ability model (MATH), rational thinking model (RAT) and thinking disposition model (DISP). The first MATH model assumes that the CRT measures only mathematical ability and is equivalent to a regression analysis in which CRT performance is predicted only by the score in the numeracy test. The RAT and DISP models assume that the CRT, in addition to mathematical ability, also measures rational thinking (assessed by BBS) and the thinking disposition of AOT. Campitelli and Gerrans (2014) concluded that the "analyses provided very strong

evidence (BIC difference > 10) in favor of RAT and DISP over MATH and that, "therefore, CRT is not just another numeracy test" (p. 441). On the contrary, and in accordance with our findings that the CRT and BNT are factorially indistinguishable, our analyses showed that the MATH model fitted our data better than the RAT and DISP models (*BIC* (math) = 3993.49; *BIC* (rat) = 4001.26; *BIC* (disp) = 4345.35). Therefore, it seems that the CRT scores are best explained by the BNT scores alone.

The finding that the traits that the CRT shares with non-lure BNT tasks explain all the variance in the CRT tasks indicates that the lures are not essential for the predictive power of the CRT. These results replicate the results of Attali and Bar-Hillel, although their explanation that CRT measures "numerical ability" seems too narrow, as we believe that both CRT and BNT also capture different thinking dispositions that might even be more important for their predictive power than the "pure" mathematical ability.

As BBS are not math tasks, it did not make sense repeating the same analysis that we did on CRT, i.e. checking whether BNT and BBS are factorially indistinguishable. Therefore, we conducted a different analysis that helped us answer the question of lure importance for (supposedly) cognitive reflection measures. We wanted to see to what degree will accounting for the effects on BNT in BBS and H&B tasks using SEM affect the correlations between the BBS and H&B factors. In order to do that, we specified a model in which a BNT factor was regressed on each of the BBS and H&B factors, and left residual variance in the factors free to co-vary. The results showed that, when the effects of BNT were accounted for in this way, all of the correlations between BBS factor, H&B factors and AOT factors substantially decreased and ceased to be significant (for BRN from r = .33 to r = .01; for FCS from r = .25 to r = .09; for CBR from r = .30 to r = .03; for AV from r = .13 to r = -.03; for AOT from r = .32 to r = .14). Judging from these results, it seems that the BBS correlates with different outcomes mostly for the same reasons that the non-lure BNT correlates with these same outcomes. Again, as for CRT, the most plausible conclusion seems to be that the lures are not crucial for the predictiveness of the BBS.

Discussion

Our study represents a test of convergent/discriminant validity of CRT and BBS, two types of tasks that are supposed to capture the cognitive reflection construct. More specifically, we wanted to explore whether their unique characteristic of cuing a strong intuitively appealing, but wrong, response is responsible for their correlations with different H&B tasks and thinking dispositions. We did this in

several different ways. First, we compared the correlation coefficients between our two cognitive reflection measures with lures (CRT and BBS) and H&B tasks with the correlations between our non-lure tasks and H&B tasks. These correlations were either the same or our non-lure BNT task was correlated more strongly with H&B tasks. Second, we tested whether the CRT and BNT are factorially indistinguishable by comparing a two-factor model (CRT and BNT items load on separate factors that are allowed to correlated) with a one-factor model (CRT and BNT items load on the same factor). The two-factor model did not show better fit than the one-factor model, meaning that the same underlying trait probably affected both CRT and BNT performance. Third, using Campitelli and Gerrans (2014) formula, we tested a model that presumes that the CRT responses depend only on numeracy against the models that they, in addition to numeracy, also depend on rational thinking skills and thinking dispositions. The first model described out data the best. Numeracy was the only relevant predictor of the CRT responses, rational thinking (operationalized as BBS result) and thinking dispositions (operationalized as AOT result) did not contribute over numeracy. Fourth, in order to see whether the lures are making the CRT items "good" items, we correlated the lureness index of the CRT items with their respective loadings on a one CRT – BNT factor. These were not correlated meaning that whatever traits the CRT and BNT have in common, the lures are not responsible for it. Finally, we checked whether the correlations between the BBS and outcomes (H&B tasks and AOT) would be diminished when we statistically account for the effects of BNT on BBS, H&B tasks and AOT. All of the correlations were substantially smaller meaning that the BBS correlate with H&B tasks and AOT mostly for the same reasons that the BNT correlates with them. This represented another piece of evidence that the correlations between BBS and outcomes largely do not depend on the lures.

Our findings showed that all the valid variance in the CRT was explained by the numeracy factor as the same traits that influence performance on the non-lure numerical problems also influence performance on the CRT tasks with lures. Thus, for whatever reasons CRT predicts a wide range of outcomes described in the introduction, it has probably little to do with the lures. The characteristic that made the CRT items famous, ability to trigger false intuitive responses, seems not to be the test's characteristic responsible for its predictive validity. Performance on the CRT tasks predicts outcomes because these are good math tasks, not because these tasks require suppression of the initial wrong response. One implication of these results is that different studies that utilized regression analysis to conclude that the incremental validity of CRT over numeracy stems from lures (e.g. Barr, Pennycook, Stolz, & Fugelsang, 2015a,b; Liberali et al., 2012; Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2014; Trippas,

Pennycook, Verde, & Handley, 2015) might be due to a) narrow measures of numeracy that did not capture complete range of the disposition (at least not to the extent that BNT does), b) low reliability of numeracy and CRT measures making both measures imperfect and incomplete measure of the numeracy construct (see Baron, et al., 2017 for a discussion about statistical control), or c) Type 1 error characteristic of this kind of regression analysis (Westfall & Yarkoni, 2016).

However, the key question is which abilities and/or dispositions account for performance on math tasks, whether the lure or the non-lure ones. Attali and Bar-Hillel (2020) call these traits "mathematical ability". Although we are sure that they do not imply that the traits affecting CRT and BNT responses are abilities in a narrow sense of capabilities free from certain thinking dispositions, nevertheless this does sound a bit narrow. Therefore, we would argue (along with a lot of other authors, i.e. Baron et al., 2015; Cokely and Kelley, 2009; Ghazal et al., 2014) that, in addition to mathematical ability in a narrow sense, some thinking dispositions must play role in the CRT and BNT performance and account for their correlations with different outcomes. Our finding that non-math task (BBS) correlates with different outcomes for the same reasons as the math task (BNT) implies that BNT (and consequently CRT) does not correlate with these outcomes only because it assesses mathematical ability that might account for these correlations. Instead, at least one disposition could account for BBS and BNT correlations with different outcomes. This disposition might be reflective and careful approach to cognitive tasks that includes taking more time in order to be more accurate, a disposition referred to as R/I (Baron, 2018; Baron et al., 2015; Baron et al., 2017). In their protocol analysis of decision making under risk, Cokely and Kelley (2009) found that both CRT and numeracy predicted higher number of verbalized considerations on risk decision-making tasks and number of considerations was further related both to the number of normative correct responses and to the response times. The authors concluded that CRT and numeracy are associated with more careful, thorough, and elaborate cognition. In line with this are the findings that there is sometimes a positive correlation between CRT score and CRT response time (e.g. Baron et al., 2015; Stupple et al., 2017), as well as that participants that scored higher on BNT performed better on various tasks (lotteries, intertemporal choice, denominator neglect, and confidence judgments) because they deliberated more during decision making and, in that way, more accurately evaluated their judgments (Ghazal et al., 2014).

In sum, we can conclude that our results thus replicate Attali and Bar-Hillel (2020) findings that all the systematic variance in the numerical CRT can be explained by "the math factor" where this factor is

influenced both by math ability and thinking dispositions (such as R/I). What seems to be clear from this, as well as several previous studies (Attali & Bar-Hillel, 2020; Baron et al., 2015) is that the lures are not essential for the predictive validity of cognitive reflection measures. In other words, our findings indicate that what supposed to be a cognitive reflection test does not capture the ability or disposition to resist reporting the response that first comes to mind (Frederick, 2005) but rather a stable characteristic to be careful and reflective from the start. In this regard it is similar to many of the others cognitive tests that allow participants to sacrifice speed for accuracy. We also tried to expand on Attali and Bar-Hillel results by examining BBS as another measure of cognitive reflection. Similarly as for the CRT, our results indicate that the lures do not play important role in correlations between BBS and other tasks. Thus, we doubt that either of cognitive reflection measures actually measure cognitive reflection as defined by Frederick (2005).

The conclusions of the current study are qualified by several facts. First, as mentioned before, our sample consisted of college students that are on average more intelligent, numerate and open-minded than the general public. In this particular case, this fact can be relevant. Namely, at least some of the college students could have ample experience with basic mathematical operations that are required to successfully solve CRT items and through their education they could have lots of opportunities to train their skills. This means that some of the college students might have developed good mathematical intuitions that allow them to do basic mathematical operations swiftly and almost intuitively. It is also in line with the „hybrid" dual-process model that posits that not only incorrect but also correct responses can be intuitively cued and with greater probability among those more experienced in particular task (De Neys, 2017). However, this could in turn mean that the effect of deliberation and reflection on accuracy in solving CRT tasks would be diminished in our sample. The other significant drawback of the study is the fact that the sample on which we calculated our correlations between our (non)lure tasks and H&B tasks was halved. This could mean that the parameters are estimated with lesser precision.

Conclusion

CRT is deemed to be a specific measure of cognitive reflection defined as the ability or disposition to resist reporting first response that comes to mind because of its ability to cue intuitive but incorrect responses that need to be detected and overturned in order to produce a correct response. However, it seems that neither the CRT nor BBS as another cognitive reflection measure capture cognitive reflection conceptualized in this way. This conclusion follows from the fact that, in our study, the same traits that

accounted for performance on the non-lure cognitive task (those that do not cue intuitive incorrect response) completely accounted for performance on the CRT. This means that the lures do not capture any additional disposition not captured by numerical non-lure tasks and, thus, that they do not account for the broad predictive ability of the CRT. Similarly to the CRT, the lures do not appear to be especially important for the predictive ability of BBS as its correlations with various outcomes were substantially diminished once the effect of non-lure task (BNT) was statistically accounted for in a SEM regression. We believe that cognitive reflection measures capture some basic cognitive capabilities and thinking dispositions that allow them to correlate with such a wide variety of tasks as well as real life outcomes.

3. STUDY 2: WHO DETECTS AND WHY – HOW DO INDIVIDUAL DIFFERENCES IN COGNITIVE CHARACTERISTICS UNDERPIN DIFFERENT TYPES OF RESPONSES TO REASONING TASKS?

This chapter was previously published as: Erceg, N., Galić, Z., Bubić, A., & Jelić, D. (2022). Who detects and why: how do individual differences in cognitive characteristics underpin different types of responses to reasoning tasks? *Thinking & Reasoning*, 1-49.

Introduction

One of the most famous problems in the decision-making literature is the "bat and a ball" problem from the cognitive reflection test (CRT; Frederick, 2005). The problem goes as follows: „A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" Similarly to other problems from this test, this one automatically triggers a relatively strong initial response (i.e., 10 cents). However, after a more careful reflection, it becomes clear that the right response is 5 cents. The test became widely popular because it elegantly illustrates the main points of the dual-process theories of reasoning.

The dual-process theories

Dual-process theories characterize human thinking and reasoning as an interplay of fast, automatic, autonomous, and non-conscious System 1 and slower, rule-based, effortful deliberate System 2 (De Neys, 2012; 2015; Evans & Stanovich, 2013; Kahneman, 2011). A more recent expansion of the dual-process model is the tripartite theory (Stanovich, 2009; Stanovich, West & Toplak, 2016) which further differentiates between two aspects of System 2 processing, the reflective and the algorithmic mind. It is this interplay between the two systems that is elegantly captured by the CRT items. To overcome the initial wrong response generated by the fast and automatic System 1 (10 cents) and arrive at the correct one (5 cents), one has to reflect on the answer and recognize the need to engage in more deliberate processing (the reflective mind), but also to possess adequate computational power, knowledge, and abilities to calculate the right answer (algorithmic mind).

There are two basic models of the dual-process theories. According to the first, default-interventionist model, the two systems operate serially. Automatic and fast System 1 processes are activated first and they produce an intuitive, heuristic response (Evans, 2006; Kahneman & Frederick, 2005). To arrive at

the correct conclusion, System 2 must "intervene" and change this default, heuristic response. Therefore, when one gives a heuristic response to the problems such as the CRT problem above, it is because one failed to engage in effortful and deliberate System 2 processing. That can be either because one did not recognize the need to engage in more deliberate processing (lack of the reflective mind) or because (s)he could not simply calculate the right response (lack of the algorithmic mind). When the correct response is given it has to be because System 2 intervened and inhibited the invalid, heuristic response (Travers, Rolison & Feeney, 2016). According to the second, parallel activation model, the two systems do not act serially one after the other but are both active at the same time, simultaneously computing a problem solution from the start and competing for control of response and behavior (Handley & Trippas, 2015; Sloman, 1996; 2014; Trippas & Handley, 2018). Both of these models face a problem when trying to explain how and when a reasoner detects that the output from the intuitive System 1 is wrong and in conflict with the correct response. On the one hand, the default-interventionist model has a hard time explaining how a reasoner can ever detect the conflict if System 2 is not already engaged from the start. On the other hand, the parallel activation model unrealistically expects that the harder and more cognitively demanding System 2 processing will be engaged from the very start of a reasoning process (De Neys, 2015).

A new, hybrid model of logical intuitions was recently proposed (De Neys, 2012, 2014, 2015; Pennycook, Fugelsang & Koehler, 2015). The logical intuitions model represents a modification of traditional dual-process theories as it takes into account evidence indicating that people are generally implicitly aware of the conflict between the heuristic and normative response, even in the cases when they select the incorrect response. The typical experimental paradigm for studying conflict detection contrasts the original problem that cues the wrong heuristic response (such as the bat-and-ball problem) with a similar no-conflict problem where the heuristic response is aligned with the normative response and where the participants are not being lured into giving a wrong response. For example, a no-conflict version of the bat-and-ball problem would be: "A bat and a ball cost $1.10 in total. The bat costs $1. How much does the ball cost?" (De Neys, Rossi, & Houdé, 2013). Unlike the original version of the problem, this version does not cue the wrong response and should be much easier to solve for participants. After answering both the conflict and no-conflict questions, participants are asked to indicate their confidence in the answers. Naturally, since the no-conflict version of the problem is trivially easy and mostly solved correctly, participants are quite confident in their responses. The logic behind asking for confidence ratings is that, if people do not notice that they are being lured into a wrong response on the original

conflict task, then they should perceive the two problems to be essentially the same, i.e. trivial. Therefore, they should be similarly confident in their responses.

However, people seem to be less confident in their responses on the original conflict task than on the parallel no-conflict task also in the cases when they end up giving the wrong heuristic response (De Neys, Cromheeke, & Osman, 2011; De Neys et al., 2013). As this drop of confidence does not come from deliberate and careful thinking, the authors have proposed that the intuitive System 1 is also sensitive to logical principles and mathematical rules and that, at least under certain conditions, it is capable of producing both the incorrect heuristic and correct normative response. These System 1 produced responses that are following logical and mathematical rules are called logical intuitions (but see Ghasemi, Handley, Howarth, Newman & Thompson, 2021). These findings are the basis of the previously described hybrid dual-process model. Given that intuitions are activated at some level, people detect the conflict and are aware that something is happening even if not quite sure what. In the case when the heuristic and normative responses conflict, this conflict creates a sense of arousal that signals people to doubt their heuristic response and lower their confidence (De Neys, 2015) which sometimes leads to better accuracy on reasoning problems (e.g. Šrol & De Neys, 2021).

Conflict detection indicators

In addition to the decreased confidence ratings (Bago & De Neys, 2017; De Neys, et al., 2011, 2013; Frey, Johnson, & De Neys, 2018; Mevel et al., 2015; Stupple, Ball, & Ellis, 2013), other indicators of conflict detection on reasoning tasks have been studied. For example, Stupple, Pitchford, Ball, Hunt, and Steel (2017) assert that conflict detection should also be observable from the response times (RT) on such tasks. If people detect the conflict between the heuristically wrong and normative right answer, they should be more careful when solving the problem resulting in longer RTs compared to those that do not detect the conflict, and this increase in time should be related to greater accuracy on the task. In this regard, studies are equivocal as some found a positive correlation between response accuracy and RTs (although the effect sizes are quite modest, e.g. $r = .18$ in Stupple et al., 2017) while others failed to find such correlation (e.g. Damnjanović, Novković, Pavlović, Ilić and Pantelić, 2019). However, it must be noted that response time can only be a noisy and imperfect proxy for conflict detection as prolonged time can, for example, just mean that a person has a preference for a slower and more careful approach when solving problems (Bago & De Neys, 2017). To control for this overall preference, similarly to the confidence ratings, it is possible to compare the RTs between conflict and no-conflict tasks. These studies

generally show that people spend more time solving a conflict task in comparison with its no-conflict counterpart (e.g. De Neys & Glumicic, 2008; Frey et al., 2018; Johnson, Tubau, & De Neys, 2016), although this prolonged time failed to result in higher accuracy in several recent studies (Swan, Calvillo, and Revlin, 2018; Šrol & De Neys, 2021; Teovanović, 2019). Taken together, the confidence and response time differences between the conflict and no-conflict task, as well as the simple response time that one takes to complete the reasoning task could be taken as indicators of conflict detection. As Šrol and De Neys (2021) point out, any single detection indicator is imperfect and could therefore point to different conclusions, so it is useful to use several indicators simultaneously in a study.

The universality of conflict detection

Considering that many conflict detection studies seem to show that conflict detection is ubiquitous and present even among the people that fail to solve the reasoning conflict tasks correctly, some authors speculate that logical intuitions are universal. For example, De Neys (2015) asserted that the "lack of individual differences in conflict detection efficiency further suggests that the necessary normative knowledge activation is indeed effortless" (pp. 31). However, studies investigating individual differences in conflict detection suggest that conflict detection, as well as logical intuitions, might not be that universal after all. For example, in Mevel et al. (2015) study, only 56% of biased respondents showed a confidence decrease on the ratio bias reasoning task (e.g. assessing the probability of drawing a red marble from a tray of red and white marbles based on the relative proportion and not a total number of red marbles), indicating that only about half of the participants that failed to correctly solve the problem detected the conflict. Similarly, Frey et al. (2018) reported that 66% of biased respondents showed signs of conflict detection on the base-rate problem, 57% on the conjunction problem, and only 38% on the bat-and-ball problem. Therefore, it seems that although a relatively large proportion of participants detects the conflict, there is also a substantial proportion of those "happy fools" (De Neys et al., 2013) that do not detect it and proceed with giving the wrong response without ever doubting it.

Furthermore, some people seem to be able to instantly give the right response on the reasoning tasks with showing very little signs of conflict detection. The two-response paradigm is especially suitable for studying this. In this paradigm, the participants give their responses to the same task two times: the first response is given under a strict deadline to ensure that it is a product of an intuitive System 1 processing, while the other is given without constraints, and respondents can change their original response (Thompson & Johnson, 2014; Thompson, Turner & Pennycook, 2011). Using this paradigm, Bago and

De Neys (2017) showed that approximately 30% to 40% of participants intuitively gave the correct responses on base-rate problems and belief-bias syllogisms, in comparison to only 6% to 10% of those that gave intuitively incorrect response, but corrected it later through deliberation, as the default-interventionist dual-process view would predict. The results were similar for CRT problems too – although a majority of participants respond incorrectly to these problems, when a correct response is given, in around two-thirds of the cases it is generated intuitively, from the start, and only rarely through additional deliberation (Bago & De Neys, 2019). Similarly, using the protocol analysis, Szaszi, Szollosi, Palfi, and Aczel (2017) demonstrated that the majority of participants who correctly solved CRT items (77%), immediately started their response with a correct answer or with a line of reasoning leading to a correct answer. Only the minority (23%) started their response with an incorrect answer and then, through deliberative thinking, concluded that it was wrong and came up with the right answer. These results strongly indicate that correct responses to reasoning problems in the majority of cases are generated intuitively. This represents a big problem for the classical dual-process views that posits that the correct responses should be reached through the activation of slow and deliberate System 2 processes. So, how can these findings be explained and reconciled with the dual-process position?

Differential intuition strength

Recently Bago and De Neys (2017, 2019, 2020) extended the hybrid, logical intuition view with a notion of differential intuition strength. They built this model on Pennycook et al. (2015) three-stage model of analytic engagement. In short, this model posits that in the first stage, the autonomous Type 1 processes generate several intuitive responses and that some of those intuitive responses come to mind faster and more fluently than others. If one response substantially dominates the others in terms of this fluency, then no conflict between the responses will be detected and this intuitive response would be the final one. If, however, no initial response significantly dominated the others in terms of ease of generation, the person might detect that two or more of the initial responses conflict with another. In this case of conflict detection, according to Pennycook et al. (2015), there are two possibilities. First, a person can focus on justifying and elaborating the first intuitive response without giving serious thought to the competing response(s), presumably because the first one came to mind somewhat easier than the others. The authors call this process rationalization. Alternatively, after detecting the conflict, a person can engage in what is typically seen as analytical thinking, conclude that the initial response is not the best one after all, and opt for another response that seems better after more careful deliberation.

Although the initial Pennycook et al. (2015) model does not explicitly state that the dominant intuitive response will be an incorrect one, it nevertheless seems to imply it. For example, they say that the process of rationalization "leads to a response in line with what would typically be considered bias (i.e., one's strongest intuition, which will often be personally relevant), but that has been bolstered by analytic reasoning (an ''effortful'' belief-based response)" (p. 40). Thus, it seems that, at least implicitly, the model presumes that the dominant intuition will in most cases be incorrect.

However, as it is clear from the two-response studies described before, this does not have to be the case and in many cases - it is not. Bago and De Neys (2017, 2019) model explicitly accounts for this by presuming that the initial and final responses will depend on the absolute and relative strengths of different intuitive responses that are generated, of which one will typically be the normatively correct response, the product of logical intuition. Specifically, according to their differential intuition strength view, the initial, intuitive responses in a two-response paradigm will depend on the absolute strength of the intuition. If the logical intuition is stronger than the incorrect one, then the initial response will be correct, otherwise, it will be incorrect. The subsequent conflict detection will depend on the relative strengths of competing intuitions. If logical and incorrect intuitions are roughly similar in strength, i.e. no response is particularly more salient or fluent, the person will then probably detect the conflict between those responses. The more similar the strength of intuitions, the greater the probability of conflict detection. Conversely, the greater the difference in the strength of intuitions, the lower the probability of conflict detection. These differences in the intuition strengths can elegantly account for the recent findings that contradicted the traditional dual-process views. For example, an intuitive correct response when respondents show very little signs of conflict detection is probably due to logical intuition being substantially stronger than the incorrect one, while incorrect responses even after a period of deliberation point to the opposite (i.e., logical intuition being substantially weaker than the incorrect one). What was previously thought to be a predominant way of solving these reasoning tasks, by detecting and overriding a conflict through deliberation, would only be instances where the intuitions are similar in strength.

Determinants of intuition strength

In his recent work, Stanovich (2018) proposed that the mindware, i.e. specific knowledge and skills one gains through experience, is a key variable that affects the strength of logical intuitions for a given task. Recently, Purcell, Wastell, and Sweller (2020) nicely elaborated on how this idea of differential

mindware instantiation aligns with a hybrid dual-process model. As a person's mindware (i.e. domain-specific experiences and skills) becomes more advanced, he/she relies less on working memory and Type 2 processes. In the beginning, with no experience, a person does not possess any kind of relevant mindware, thus having no or very weak logical intuitions related to the problem at hand. With such an underdeveloped mindware, the potential for conflict detection is very weak as there is no logical intuition strong enough to conflict with the incorrect one. Thus, the intuitive incorrect response given with little thought is the most probable in the first phase. However, as a person progresses with learning and practicing, his/her mindware becomes more developed, increasing the strength of logical intuitions. When a mindware becomes learned to a sufficient degree, a logical intuition can become as strong as any other. It is at this stage of mindware instantiation that conflict detection and override become possible through the engagement of Type 2 processes. If the conflict is strong enough, a person can engage in deliberate, analytical thinking and, drawing from the acquired mindware, reject an incorrect intuitive response in favor of the correct one. Further down the road, with sufficient experience and practice, a mindware can become overlearned to a degree that the correct response becomes automatic. In other words, with such developed mindware (rich knowledge and experience related to a specific task at hand), logical intuitions can become so strong and come to mind so easily as to completely dominate over the incorrect one. When this happens, a person can give a correct response instantly with little to no conflict detected.

In sum, it can be said that there are three key phases regarding the development and instantiation of mindware and the corresponding reliance on different types of thinking. In the first phase, with underdeveloped mindware, a person employs Type 1 processing, resulting in generally incorrect responses with little conflict detected. In the second phase, a person employs Type 2 thinking, drawing on the acquired mindware and increasing the chance of correct responses. Finally, in the third phase, a person again employs Type 1 thinking, only this time giving mostly correct responses with little conflict detection due to overlearned and highly developed and automatized mindware (Stanovich 2018; Purcell et al., 2020). Purcell et al. (2020) have only partially confirmed these hypotheses, namely in the case when mindware quality was operationalized by real-life mathematical expertise and experience (undergraduate psychology students as a low-experience group, undergraduate science and engineering students as intermediate experience group, and postgraduate mathematical students as high experience group), but not when it was experimentally manipulated. In the former case, the intermediate experience group hypothesized to rely on Type 2 processes when solving CRT tasks indeed scored significantly

lower when their deliberative abilities were constrained compared to an unconstrained situation. This indicates that this group was indeed able to reach the correct response if allowed enough time to do so, i.e. they were able to come up with the correct response through Type 2 processes. Conversely, the low and high experience group did not show diminished performance in constrained vs. unconstrained conditions. The low experience group was similarly bad in both conditions, indicating that they accepted their incorrect initial response with little questioning of the response, presumably due to underdeveloped mindware that would allow them to detect the conflict between the correct and incorrect intuition. The high experience group responded similarly – by responding intuitively – only their intuition was correct from the start, indicating that their overlearned mindware could have produced very strong correct, logical intuition, much stronger than the incorrect one.

Several recent studies confirmed the quintessential role of mindware in successfully responding to reasoning tasks. The most direct test was reported by Šrol ad De Neys (2021) who showed that mindware instantiation was the single best predictor of both conflict detection efficiency and overall accuracy on reasoning tasks, even after accounting for the effects of several important individual differences measures (cognitive ability, numeracy and need for cognition). Apart from this study, at least two recent studies showed how the development of mindware boosts intuitive correct responding. Specifically, Boissin, Caparos, Raoelison, and De Neys (2021) demonstrated how training participants on CRT-like tasks not only significantly increased their final responses in the two-response paradigm, but also their initial, intuitive responses, and these effects were sustained over two months. The authors speculated that the short training boosted participants' mindware by reminding them how to use the knowledge that they already possessed, i.e. by making that knowledge and its usage more available in participants' minds. Finally, an elegant test of the "automatized mindware" idea was conducted by Raoelison, Boissin, Borst, and De Neys (2021). They demonstrated how the development and automatization of the relevant mindware in children between 7^{th} and 12^{th} grade dramatically impacted the way they responded to reasoning problems (base-rate neglect and belief-bias syllogisms). Older children were not only more likely to deliberately correct an erroneous initial response but also to generate a correct response from the start, confirming the crucial role of mindware instantiation in developing strong logical intuitions.

The role of individual differences in abilities and dispositions

The question is how do the individual differences in cognitive abilities and thinking dispositions fit with this view of the crucial role of mindware. The hybrid dual-process theory throws an interesting and new

perspective on the role of cognitive abilities and dispositions in the success on reasoning tasks. Currently, the dominant view in the literature is the "smart deliberator" view that assumes that people with higher cognitive abilities are better at reasoning tasks because they are better at correcting erroneous intuitions (cf. Raoelison, Thompson, & De Neys, 2020). Similar can be said for thinking dispositions such as the disposition to engage in analytical thinking that has been dubbed to influence the motivation to engage in the deliberative correction of intuitive incorrect responses. This view, for example, follows from the tripartite theory of Stanovich and colleagues (e.g. Stanovich, 2009; Stanovich et al., 2016) which posits that cognitive abilities and thinking dispositions are jointly responsible for recognizing, overturning, and correcting the intuitive incorrect responses.

However, the hybrid dual-process model suggests that abilities and dispositions could affect the performance of tasks not because they help one to overcome intuitive response but because they are crucial for developing strong logical intuitions. As Evans (2019) put it, high-ability people are better at solving different reasoning tasks because they are more practiced in reasoning and have thus automated some of the skills required for successfully responding to certain types of tasks, such as processing numerical information. In other words, high abilities and dispositions toward analytical thinking could predispose some people to search for situations in which they can gain specific knowledge and skills (i.e., some people can show a preference for complex problems that require hard thinking, such as difficult mathematical or logical problems) and also allow them to learn more and more quickly from such situations. With practice, these people will develop relevant mindware and automatize it to the degree that it will allow them to respond correctly to reasoning tasks by following their logical intuitions. From this view, it would follow that, for example, people who in two-response studies respond correctly from the start (what is usually coded as "11" responses) should have higher cognitive abilities and disposition to engage in analytical thinking than those who would respond correctly through deliberation (typically coded as "01" responses). Conversely, the "smart deliberator" view would predict no differences between those responses, or even "01" responders scoring higher on cognitive ability and thinking disposition measures than "11" responders.

There were only a few studies so far that aimed at investigating whether cognitive abilities matter more for intuitive or deliberative responses, but the evidence seems to tip in favor of the "smart intuitor" and not "smart deliberator" view. More specifically, Raoelison et al. (2020) showed that cognitive capacity was positively correlated with the probability of giving "01" responses (i.e. responding correctly through

deliberate correction of erroneous intuition), but that correlation was significantly stronger for "11" responses, meaning that the cognitive capacity was more important for intuitively correct responding than for correct responding through corrective deliberation. Several other recent findings align with this conclusion. For example, Thompson, Pennycook, Trippas, and Evans (2018) showed that for high-capacity reasoners, statistical intuitions were stronger than the incorrect, stereotypical ones, interfering with the ability of high-capacity participants to produce stereotypical (in case of base-rate neglect tasks) or believable (in case of belief-bias syllogisms) responses. The opposite was true for low-capacity responders. This indicates again that cognitive ability matters more for developing strong logical intuitions and responding in line with them than for correcting incorrect initial responses. More recently, Schubert, Ferreira, Mata, and Riemenschneider (2021) replicated these findings by again showing that high-ability participants performed worse when asked to assess the believability of a conclusion, rather than its logical validity. Again, strong logical intuition in high-ability responders seems to have interfered with the ability to produce a response based on erroneous intuition. However, this study has also investigated the role of thinking dispositions in addition to cognitive abilities. Unlike the cognitive abilities that were important foremost for intuitively correct responding, the need for cognition as a disposition towards analytical thinking was related to successful conflict resolution, i.e. responding correctly through detecting the conflict and deliberately correcting incorrect intuitive responding. This is more in line with the "usual" view of the role of thinking dispositions as a motivation to engage in analytical thinking, and less in line with a view that thinking dispositions motivate people to expose themselves to situations that facilitate the development of relevant mindware and strong logical intuitions.

Current study

As it is clear from the introduction so far, the hybrid model makes somewhat different assumptions about the role of cognitive abilities and thinking dispositions in success on reasoning tasks than the "classical" default-interventionist or parallel dual-process model, and there are only few studies that have tested these assumptions. Our goal is to contribute to this body of evidence by broadening the range of individual difference variables and testing them by using two different methodological approaches. Specifically, the goal of the current study is to investigate the individual differences in cognitive abilities (intelligence and numeracy), thinking dispositions (actively open-minded thinking and need for cognition), and knowledge (high-school math knowledge) that underpin different ways of solving two reasoning tasks, cognitive reflection tasks and belief-bias syllogisms. Specifically, as it follows from

Stanovich's (2018) discussion and Purcell et al. (2020) elaboration, it is theoretically possible to differentiate between four different types of responding to reasoning tasks depending on the probability of conflict detection and accuracy. First, a person with underdeveloped mindware relevant for the task will probably give wrong responses with little signs of conflict detection as his/her logical intuitions will presumably be very weak. That should reflect in incorrect non-detection trials. Second, for those whose mindware became more developed, the strength of logical intuition will increase and they will probably start detecting the conflict between the intuitions. However, they can still end up giving an incorrect response which will reflect in incorrect detection trials. Third, for some reason (e.g. somewhat more developed mindware, higher abilities, or dispositions to engage in analytical thinking), after detecting the conflict between the competing intuitions, a person can overturn his/her initial intuitive response in favor of the correct one (correct detection trials). Finally, a person can have highly developed and overlearned mindware so that his/her logical, correct intuition becomes so strong and overcomes an incorrect one to the degree that the person does not even experience a conflict between the intuitions, but intuitively responds in a correct way (correct non-detection trials).

In line with the literature review, we advanced the following hypotheses.

a) When participants show little signs of conflict detection yet give the correct responses (correct non-detections), this is a sign of strong logical intuitions that arise due to overlearned mindware. In this regard, we expect this type of response to be associated with higher cognitive ability, numeracy, math skills, and thinking dispositions towards analytical thinking than the other types of responses.

b) When participants show little signs of conflict detection and fail to respond correctly (incorrect non-detections), this implies very weak logical intuitions, probably due to the underdeveloped mindware. In this regard, we expect this type of response to be associated with the lowest scores on cognitive ability, numeracy, math skills, and thinking dispositions towards analytical thinking of all the four response types. By referring to this and previous trials as "non-detections", we do not imply that those that respond in this way did not detect the conflict at all, but only that they show substantially weaker signs of conflict detection than the others (Bago & De Neys, 2017; 2019).

c) When participants detect the conflict and give correct responses (correct detections), they have more developed mindware compared to those that detect the conflict but give the incorrect response (incorrect

detections). In this regard we expect the latter response type to be associated with lower scores on cognitive abilities, numeracy, math skills, and thinking dispositions towards analytical thinking than the former response type.

We tried to answer our research questions by conducting two different studies. In the first one, we employed a single-response format and tried to capture conflict detection by combining three indicators, response time on reasoning tasks, response time differences, and confidence differences between the tasks with and without lures. These differences in response times and confidences should indicate that a person detected the conflict between a logical and incorrect intuition that is evoked in tasks with lures, but not in those without lures. In the second study, we employed a two-response paradigm where participants first responded under a strict time deadline and cognitive load, followed by responding without any time limit or cognitive strain. This approach allowed us to separate those that respond intuitively correctly from those that give correct responses through more pronounced conflict detection and corrective deliberation and those that do not manage to respond correctly even after deliberation. However, it must be noted that in Study 2, due to time constraints, we did not use the tasks without lures. Therefore, we were not able to differentiate between incorrect responses on which the conflict was detected versus non-detected. Thus, in this case, we can only compare the three response types: correct non-detections, correct detections, and incorrect responses, regardless of the conflict detection.

Study 1

Methods

Participants

506 university students participated in our study (26.5% males). The mean age was 21.15 (min = 18, max = 31, SD = 2.13).

Instruments

CRT (Frederick, 2005) is an instrument that was designed to measure individuals' ability to resist reporting an intuitive incorrect answer (i.e., cognitive reflection). The original version consists of three items (item example presented in the introduction) with each cuing intuitive but wrong response that

needs to be detected and corrected to arrive at a correct response. In this study, to increase the reliability and response range, we used five different CRT tasks, three from the original Frederick (2005) version and two additional taken from Toplak, West, and Stanovich (2014) expansion and Thomson and Oppenheimer (2016) alternate form of the CRT (all items are in Appendix). Thus, on the CRT one could score anywhere between 0 (if none of the responses were correct) and 5 (if all the responses were correct).

CRT control tasks. Besides the five original tasks, the participants solved five control tasks that differed from the original ones only in the fact that they did not cue intuitive wrong heuristic answers (see Appendix). These tasks were used for calculating the conflict detection indices. Specifically, considering that we timed all the responses, as well as assessed degrees of confidence in the responses, by subtracting the response times and confidences of the control tasks from the ones of the original tasks, we were able to calculate two conflict detection indices, namely the response time difference and the confidence difference.

BBS tasks. BBS tasks assess the susceptibility to belief bias. They pit the believability of a conclusion against its logical validity. In that regard they are like the CRT tasks in that, to arrive at the correct conclusion, a person must first notice that the believable conclusion, although intuitively receptive and believable, is false. In the present study, we assessed the belief bias using four different syllogisms taken from Markovits and Nantel (1989; see Appendix for all the items). Thus, on BBS, a participant could score anywhere between 0 and 4.

The International Cognitive Ability Resource (ICAR) is a broad cognitive ability assessment tool consisting of four different types of tasks: letters and numbers series, matrix reasoning items, verbal reasoning items, and three-dimensional rotation items. In this study, we administered a 16 items version consisting of four items of each type. The validation of this measure is reported in Condon and Revelle (2014). The total score was calculated as the sum of correct responses on the 16 items.

The Berlin numeracy test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) is a four-question test for assessing numeracy and risk literacy. The questions are designed in a way that they gradually become harder and a person could score anywhere between 0 and 4 on this test. An example of an item is the following: “Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in a choir, 100 are men. Out of the 500 inhabitants that are not in a choir 300 are men. What is

the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent." The total score was calculated as a sum of correct responses.

Matura score. To assess mathematical skills, we asked our participants to report their scores on the state matura that most of them had to take in the last few years. Similar to the SAT, the matura test is an objective standardized test that assesses the knowledge of high school mathematics. There are two different levels of matura that students can take, the easier and the harder one. To account for this, we assigned a higher ponder (1.5) to a harder level and computed a total matura score by multiplying the matura grade with the ponder. For example, if a student took a harder level of matura and received a grade of 5, his/her final score would be 7.5 (5 x 1.5). If a student took an easier level and scored 5, his/her total score would be 5 (5 x 1).

Actively open-minded thinking (AOT) refers to adherence to the standard of thinking that includes a thorough search relative to the importance of a question, confidence according to the amount and quality of thinking carried out, and consideration of alternatives different from the one we initially favor. In this study, we used a 15-item AOT scale that was used in Campitelli and Gerrans (2014) paper (e.g. "Changing your mind is a sign of weakness"). It is a self-report scale where participants indicate their level of agreement with items on a six-point scale (1 – strongly disagree to 6 – strongly agree). The total score on this scale is calculated as a mean level of agreement with the items and can be anything between 1 and 6.

Indices of conflict detection. We measured and calculated three different indices of conflict detection on reasoning tasks.

a) Response time. We measured response time in seconds for each of the reasoning tasks' items, from the moment an item would appear on the screen to the moment a participant would write an answer and press "Next" to move to the next question.

b) Response time difference. We calculated a response time difference as a difference between a mean response time for the original reasoning tasks' items and the mean response time for the control items. A positive value means that a person spent more time solving an original task than the control one, indicating conflict detection.

c) Confidence difference. After each of the original and control items, we asked participants how confident they are in their answers on a scale up to 100 %. We calculated a confidence difference as a

difference between mean confidence on the control tasks and mean confidence on the original tasks. A positive value means that the participants were more confident in their answers when solving control items than the original ones.

An important thing to emphasize is that, as we did not have BBS control tasks (tasks without the lures), we could not calculate response time and confidence differences for BBS. Thus, for BBS we report calculations based solely on response time as a conflict detection indicator.

Procedure

The data were collected in two sessions with approximately six months' time intervals between them, with half of the participants participating in each of the sessions. In both sessions, participants solved questionnaires on a computer in groups of 20 to 25 participants in the same room under the supervision of the experimenters. In the first session, participants first solved the CRT and CRT control tasks which were presented in a quasi-random order, meaning there were several fixed orders of task presentation, and participants were randomly assigned to one of the orders. This was followed by solving the BBS tasks and four items numeracy test before taking a 10-minute break. After the break, participants solved 16 ICAR tasks and a self-report AOT scale. The tasks were presented as a part of a bigger battery of tasks from the judgment and decision-making domain as this study was part of a bigger data collection effort for several different projects and not all collected data are reported here. In the second session, the order of the tasks was somewhat different. Specifically, participants first solved 16 ICAR tasks, followed by the BBS tasks and four numeracy tasks before taking a 10-minute break. After the break participants solved the CRT and CRT control items in quasi-random order followed by a self-report scale of actively open-minded thinking. Similarly as in the first wave, in this wave we also collected additional data that is not reported in this study.

Results

We are starting this section by presenting the basic descriptive statistics of all the variables as well as the correlations among all the variables. After this, we will first present the results related to the CRT followed by the results related to the BBS. The basic descriptive statistics are given in Table 2.

Table 2. Descriptive statistics and the reliabilities of the measures used in the study

	Min	Max	Mean	SD	Cronbach α
CRT	.00	5.00	3.16	1.15	.65
CRT control	2.00	5.00	4.66	0.61	.20
BBS	.00	4.00	2.10	1.62	.83
ICAR	1.00	16.00	10.29	3.09	.73
Numeracy	.00	4.00	1.56	1.12	.46
Matura score	1	7.5	4.97	1.59	/
AOT score	1.87	6.00	4.48	0.65	.82
CRT time	6.21	104.20	38.46	15.89	.47
CRT control time	10.47	64.06	30.79	9.69	.51
CRT time difference	-32.09	61.93	7.67	13.61	/
CRT confidence	33.80	100	88.00	12.52	.49
CRT control confidence	38.40	100	90.47	10.67	.54
CRT confidence difference	-45.00	46.00	2.47	11.43	/
BBS time	7.11	51.66	18.19	6.83	.58

Note. CRT = Cognitive reflection test; BBS = Belief bias syllogisms; ICAR = International cognitive ability resource; AOT = Actively open-minded thinking. CRT and BBS time = average response time in seconds for the CRT and BBS items; CRT time difference = average response time difference in seconds between the original and control CRT items; CRT confidence difference = average confidence difference between the control and original CRT items.

Several things are apparent from Table 2. First, participants found control CRT items to be easier than the original ones (t (505) = 22.60, p = .00, d = 1.00). A great majority of the participants had perfect or near-perfect scores on these items but not on the original ones, indicating that they had little problems solving tasks that did not cue wrong responses. Second, the difference in difficulty between the two versions of the CRT tasks was reflected in average response times for these tasks where it took participants significantly more time to solve the original than the control items (t (505) = 12.68, p = .00,

d = 0.56), as well as confidence in their responses where respondents were more confident when solving control items than the original ones (*t* (505) = 4.86, *p* = .00, d = 0.22). However, as it is evident from the negative minimum values on the response time difference and confidence difference variables, there were large individual differences where some participants unexpectedly took more time/were less confident solving control tasks than the original tasks while others took less time/were more confident solving control tasks than the original ones.

In sum, it seems that participants on average managed to detect the conflict on the CRT items, although wide ranges of values on the conflict detection indicators show that substantial individual differences exist. Before moving on to the main analyses, we present the correlations among the individual difference variables and conflict indicators in Table 3.

Table 3. Correlations among the study variables

	BBS	ICAR	Numeracy	Matura	AOT	CRT Time	CRT time diff.	CRT conf. diff.	BBS time
CRT	.49**	.51**	.50**	.25**	.21**	.10*	-.01	-.34**	.10*
BBS	1	.39**	.42**	.25**	.28**	-.01	-.05	-.16**	.16**
ICAR		1	.42**	.17**	.22**	.12**	.03	-.10*	.15**
Numeracy			1	.23**	.18**	-.04	-.03	-.12**	.04
Matura				1	.17**	-.07	-.07	-.02	-.07
AOT					1	.10*	.06	-.01	.05
CRT time						1	.80**	.12**	.30**
CRT time diff.							1	.29**	.16*
CRT conf. diff.								1	.00
BBS time									1

Note. CRT = Cognitive reflection test; BBS = Belief bias syllogisms; ICAR = International cognitive ability resource; AOT = Actively open-minded thinking; CRT time = Cognitive reflection test response time; CRT time diff. = Cognitive reflection test response time difference; CRT conf. diff. = Cognitive reflection test confidence difference; BBS time= Belief bias syllogisms response time.

* p < .05; ** p < .01

As can be seen from Table 3, the correlations between the two time-based indicators of conflict detection are quite high (r = .80). This gives additional support to our assumption that the response time captures conflict detection, however imperfectly. This is important as a prolonged response time by itself does not mean that a person has detected a conflict and engaged in corrective deliberation, but could just mean that he/she was careful on each item, meaning that the response time does not necessarily imply conflict detection (e.g. Bago & De Neys, 2017). However, a high correlation between response time and response time difference, which is a purer measure of conflict detection, testifies that response time, at least in this study, can be taken as a proxy of conflict detection. On the other hand, the correlations between the confidence difference indicator and the two time-based indicators are substantially lower (r = .12 and r = .29). This shows that the confidence difference indicator is somewhat distinct from the ones based on response times and confirms the importance of using different indicators of conflict detection for a more complete view.

Most of the conflict detection indices were poorly related to accuracy both for the CRT and the BBS. It seems that conflict detection is a relatively poor predictor of accuracy. However, this is not that surprising from the point of view of the hybrid dual-process model and logical intuitions which posits that these relationships will be moderated by the mindware instantiation, or the strength of logical intuitions. For example, for those with very strong logical intuitions (evidenced by the high scores on knowledge, abilities, and dispositions measures), we would expect that conflict detection is a poor predictor of accuracy, whereas for those with somewhat weaker logical intuitions (e.g. comparable in strength with incorrect intuition cued by the task characteristics) we would expect a positive correlation between conflict detection and accuracy. We will explore these interactions, therefore, in our main analyses.

Although these were not the focus of the present study, it is worth noting that the correlations among the CRT, intelligence, and numeracy are positive and of medium magnitude as it is often the case in the literature (e.g. Campitelli & Gerrans, 2014; Frederick, 2005; Welsh, Burns, & Delfabbro, 2013), and the one between the CRT and math skills is somewhat smaller but still significant. A small and positive correlation was also recorded between the CRT and AOT, which is also in line with previous findings (e.g. Campitelli & Gerrans, 2014; Toplak, West & Stanovich, 2011). The pattern of BBS correlations with these variables closely mirrors the CRT pattern of correlations. Furthermore, very low correlations were observed between the conflict detection indices and intelligence, numeracy, matura score, and AOT.

However, it is interesting to note the positive correlation between ICAR and response time both for CRT and BBS, showing that more intelligent people took somewhat more time to solve these tasks. Finally, numeracy was somewhat negatively correlated with confidence differences on CRT tasks. This suggests that the skepticism in one's responses was indicative of lower numeracy when solving the CRT tasks.

The CRT analyses

We conducted our main analyses at the level of the individual trials, meaning that the trials, and not participants, were the main unit of analysis. This created two-level nested data in which individual trials (level one) were nested under participants (level two). In this way, we effectively increased the "sample size", thereby increasing the statistical power, and ensured that the effects of conflict detection can be analyzed for correct and incorrect trials separately. To answer our main research question (how the four types of responding – correct non-detections, correct detections, incorrect non-detections, incorrect detections - reflect abilities, knowledge, and dispositions), we had to divide our trials based on the accuracy and probability of conflict detection. As we explicated before, these different types of responses should be indicative of the mindware development and the corresponding strengths of logical intuitions. To differentiate between probable conflict detections from non-detections, we combined the response times with confidence differences between the control and original item. The logic we followed here is that if a person responded relatively quickly on the reasoning task AND showed no confidence difference between the response on the control vs. original item, then that person probably did not detect the conflict on that trial. This would be equivalent to participants whose initial, fast response in a two-response paradigm is correct and who show very little confidence difference in initial responses between control and original items. These participants should have the most developed mindware and the strongest logical intuitions. However, to define our categories, we still had to decide on the cut-off point between relatively fast and other responses. The logic should be the following: the lower the response time, the lower the chance of conflict detection, given the no confidence difference. However, as this is always an arbitrary decision, it would be useful to report results for at least two different cut-off points to explore whether this change in arbitrary decision substantially affected the conclusions. If the conclusions remain similar across cut-off points, we can be more confident that our results are reliable and not much affected by the arbitrariness of our decisions.

We have decided to make two cut-off points for response times that separate a) the 10% of the fastest, and b) the 20% of the fastest respondents from the rest. As we said earlier, the lower we set the threshold, the greater chance that the response was given with very little or no conflict detection[2]. Therefore, we classified as non-detection those trials on which the response times were among 10% and 20% of the fastest respectively, and on which there was no confidence difference in responses between the control and original item. Those trials where the confidence difference between the control and original item was negative (meaning that the person was more confident in his/her response on the original item than on the easy, control item) were discarded from further analyses as these cases are very hard to logically explain. The frequencies of the trials according to accuracy/detection are shown in Table 4, both for the 10% and for the 20% threshold.

Table 4. Frequencies of the CRT trials based on accuracy and conflict detection for two different cut-off points for conflict detection, 10% and 20% of the fastest responses.

	N (10% fastest)	N (20% fastest)
Correct non-detection	85	173
Correct detection	1132	1044
Incorrect non-detection	79	152
Incorrect detection	710	637
Total	2006	2006

What is the most apparent from this table is that the intuitive correct responding (correct non-detection) is substantially rarer than deliberate correct responding (correct detection). This differs from most of the findings from the two-response studies that show that, when the correct responses are given, they are

[2] In this case, our 10% cut-off point is only slightly higher than the average reading time for CRT items that we obtained in a small pre-study (N = 18) that we conducted prior to our Study 2. In this pre-study, four of the five CRT items were the same as in this study (all but the fourth item) so we could make the comparisons. The smallest difference between the 10% threshold and average reading time was for CRT 2 item (0.54 seconds), and the largest was for CRT 3 item (3.19 seconds). Therefore, this 10% threshold obviously did not allow our participants to do much thinking, especially if we take into account that they were not instructed to read without pauses (as they were in the pre-study). Given this, it can be argued that the 10% threshold is overly conservative – as our participants did not have the instruction to read the items quickly, there is high chance that those that would be intuitively correct in two-response paradigm would end up being classified as conflict detectors here. Therefore, if we find, for example, that the correct non-detectors were significantly smarter and better at math than the correct detectors, these differences would probably be even more expressed have we expanded the threshold. It also gives us additional argument to use the 20% threshold in addition to the 10% threshold: not only we will see how this decision affects the results, but by expanding the threshold, we will probably categorize those that would respond intuitively correctly a bit more precisely.

mostly given intuitively. This either means that our cut-off points were quite strict, leaving plenty of non-detection trials categorized as detection ones, or that majority of people do not respond intuitively if they are not explicitly asked to. However, even if our cut-off points were too strict, we would not argue that this is not a problem for our purposes. As we said earlier, by making this cut-off stricter, we increase the chances that most of the correct non-detection responses are correctly classified, and that is what matters the most if we want to see the trends in individual differences between those with quite strong logical intuitions and the rest.

To examine whether individual differences in cognitive ability, numeracy, math skills, and thinking disposition underpin the four response types, we conducted a multilevel logistic regression predicting a response type from individual-differences variables. Specifically, we created three dichotomous dependent variables corresponding to the differences between the response types of interest (correct non-detections vs. correct detections, correct detections vs. incorrect detections, and incorrect detections vs. incorrect non-detections) and predicted them with individual-differences variables (abilities, dispositions, and knowledge). We did not conduct a multiple regression analysis including all the predictors in a single model, but rather a separate analysis for each predictor as we were not interested here in incremental validity of our measures. To account for the multilevel nature of the data, for each model we estimated the random variation of the intercepts across participants and the fixed effects of individual-differences variables (these are level two variables and, thus, their slopes cannot vary across the participants).

We repeated the same analyses for 10% and 20% fastest participants as the cut-off for detection. In addition to this, we repeated the same analytical approach with time-difference as the conflict detection indicator. As the results from all these analyses largely pointed to similar conclusions, to aid the readability of the paper we only report results from the analyses using response time as a conflict-detection indicator and 20% of the fastest participants as the cut-off for conflict detection. The results are presented in Table 5. All the other analyses are reported in the Appendix B. To aid the interpretability of the results, we have plotted the means and confidence intervals of our individual-difference variables for each of the response types in Figure 2.

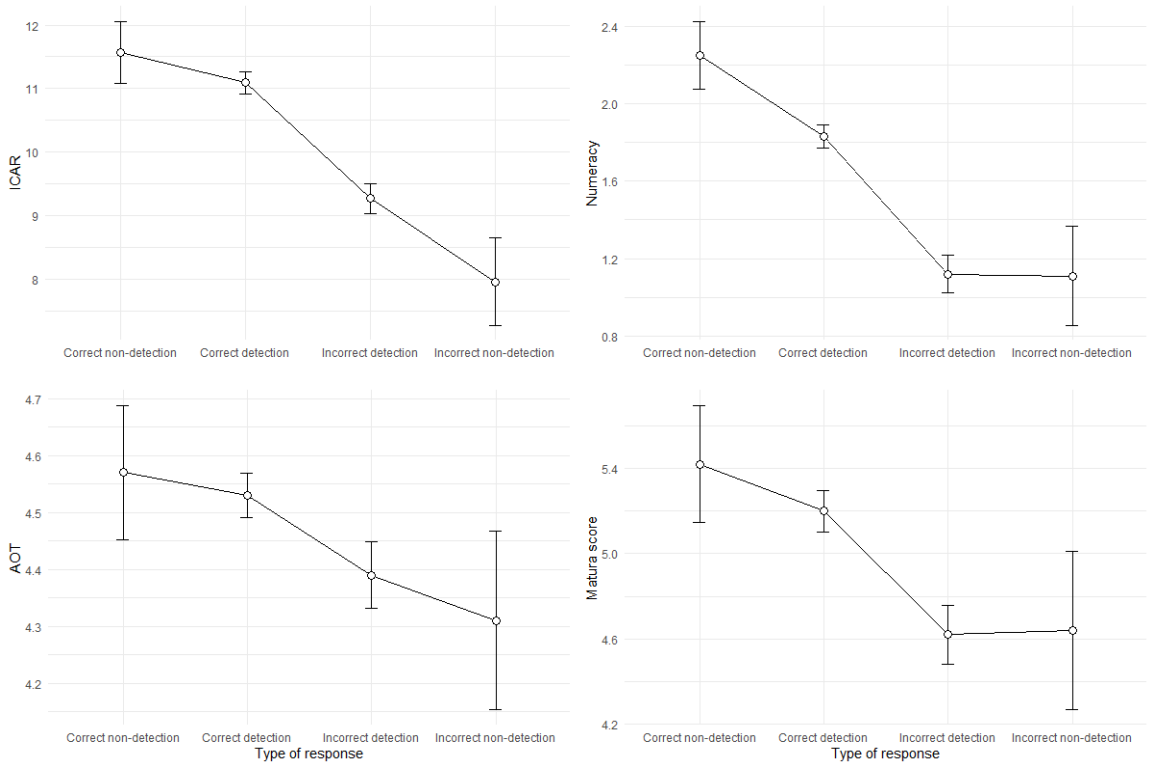


Figure 2. Differences in cognitive abilities, numeracy, actively open-minded thinking, and matura score between the four response types on CRT tasks based on response time and confidence differences as conflict detection indicators

Table 5. Results of the multilevel logistic regression analyses for the Cognitive reflection test tasks

	Correct non-detection vs. correct detections			Correct detections vs. incorrect detections			Incorrect detections vs. incorrect non-detections		
	B	SD	OR	B	SD	OR	B	SD	OR
ICAR	0.06 .	0.04	1.06	0.31***	0.04	1.36	0.09**	0.03	1.09
NUM	0.44***	0.09	1.55	0.71***	0.12	2.03	0.03	0.11	1.03
AOT	0.10	0.16	1.11	0.52**	0.19	1.68	0.13	0.16	1.14
Matura	0.07	0.07	1.07	0.30***	0.08	1.35	0.12	0.21	1.13

Note. Outcome variables are coded such that the first category (e.g. correct non-detections) is coded as 1 and the second category (e.g. correct detections) is coded as 0.

SD = Standard deviation; OR = Odds ratio; ICAR = International Cognitive Ability Resource; NUM = Numeracy; AOT = Actively open-minded thinking. p < .10; * p < .05; ** p < .01; *** p > .001

First, Table 5 offers one unsurprising result – the effects were largest when the outcome variable was "correct non-detections vs. incorrect detections", meaning that our individual-differences variables were the strongest predictors of whether a response would be correct or incorrect. Specifically, looking at the

odd ratios, a one-point increase in ICAR, Numeracy, AOT, and Matura score leads to a 1.36, 2.03, 1.68, and 1.35 increase in odds that the response would be classified as "correct non-detection" instead of "incorrect detection." What is more interesting is whether there were differences between probable conflict detections and non-detections on correct and incorrect trials. In this regard, among the correct trials, the effects were generally positive, meaning that an increase in abilities, dispositions, and knowledge increases the odds of responding in a "correct non-detection" way. However, the effect was significant only for numeracy, where a one-point increase in numeracy score increased the odds of "correct non-detections" compared to "correct detections" 1.55 times. Among the incorrect trials, although the trend was again present, the only significant, a rather small effect, was for ICAR: a one-point increase in ICAR score increased the odds of "incorrect detections" compared to "incorrect non-detections" 1.09 times.

BBS analysis

As we did not have appropriate control items for the BBS items, we could only rely on the raw response times as conflict detection indicators. As explained before, raw response times have several limitations as a conflict detection indicator. However, by using stricter thresholds for conflict detection, these limitations are somewhat mitigated[3]. As with CRT response time analyses, we again used the two cut-off points: 10% and 20% of the fastest[4]. We are showing the frequencies of BBS trials across our four categories for both cut-off points in Table 6.

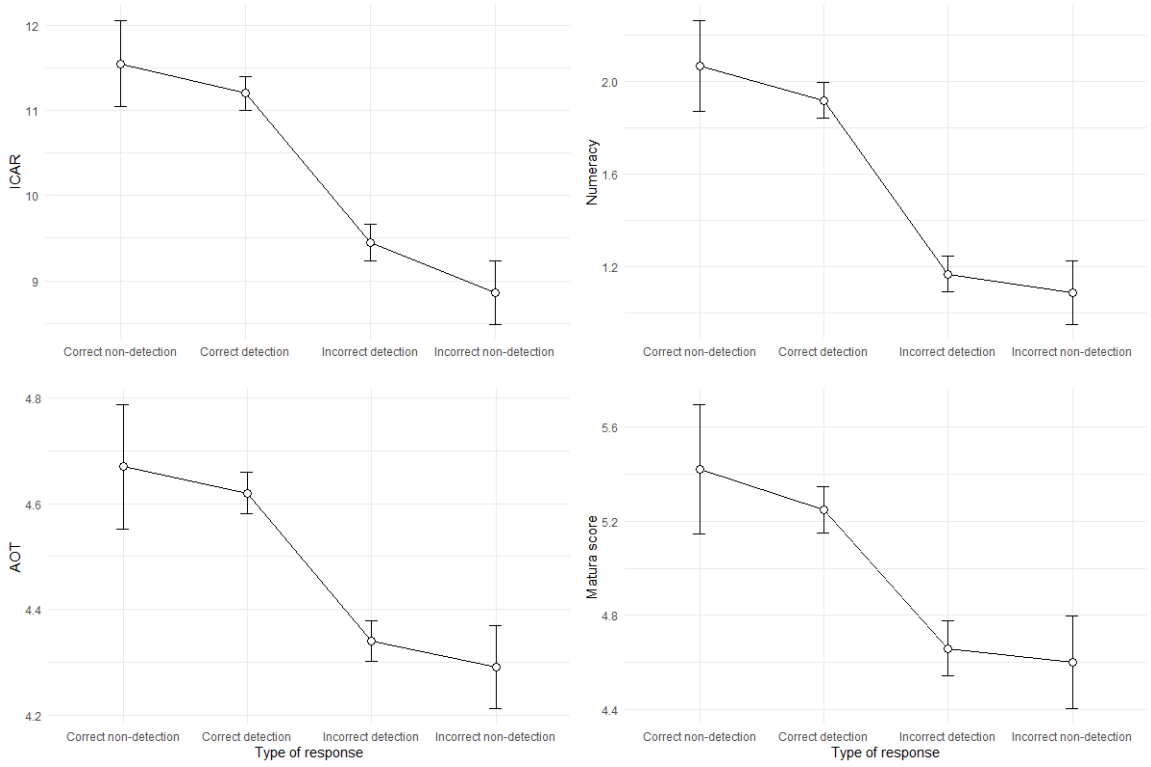
[3] Here, we could not combine response times with confidence differences, but we compared the classifications for CRT responses when only response times were used to those when they were used in conjunction with confidence differences. If there is a significant overlap between the categories, we can conclude that classifying trials based solely on (very fast) response times can be a satisfactory, although imperfect, proxy for conflict detection. For the CRT trials, the overlap was substantial for both cut-off points. For example, with the 10% cut-off point, out of 88 trials categorized as correct non-detections based on response time, 85 were also classified in this category when response time was used in conjunction with confidence difference, with only 3 being classified as correct detectors. The overlap was somewhat lower, but still relatively high, for the incorrect non-detection trials. Out of 115 incorrect non-detections based on response times, 79 were in that same category based on response times AND no confidence difference criterion. Therefore, low response times by themselves seem to be relatively good, although not perfect, indicators of non-detection.

[4] When compared with BBS reading times from the pre-study to Study 2, the 10% cut-off point was only slightly higher than the average reading time (ranging from only 0.01 seconds for item 3 to 1.72 seconds for item 1). Therefore, the fastest 10% did not have much time, if any, for conflict detection and deliberate correction of erroneous first responses. Thus, in this group, if there were instances of conflict detection, it was probably very

Table 6. Frequencies of the BBS trials based on accuracy and conflict detection for two different cut-off points for conflict detection, 10% and 20% of the fastest responses.

	N (10% fastest)	N (20% fastest)
Correct non-detection	78	155
Correct detection	983	906
Incorrect non-detection	126	248
Incorrect detection	837	715
Total	2024	2024

As the results of analyses for both cut-off values point to a similar conclusion, we again report only the results when 20% of the fastest was used as the cut-off for conflict detections. The 10% results are reported in the Appendix B. We repeated the same analyses as before – multilevel logistic regression analyses with three dichotomous outcome variables. The results of these analyses are shown in Table 7. Again, to ease the interpretation of the results for BBS, we have plotted the means and confidence intervals of our individual-difference variables for each of the response types in Figure 3.



weak, meaning that these responses should have predominately been given by those with very strong (if they responded correctly) or very weak (if they responded incorrectly) logical intuitions.

Figure 3. Differences in cognitive abilities, numeracy, actively open-minded thinking, and matura score between the four response types on BBS tasks based on response time as conflict detection indicator

Table 7. Results of the multilevel logistic regression analyses for the Belief bias syllogism tasks

	Correct non-detection vs. correct detections			Correct detections vs. incorrect detections			Incorrect detections vs. incorrect non-detections		
	B	SD	OR	B	SD	OR	B	SD	OR
ICAR	0.01	0.04	1.01	0.70***	0.11	2.01	0.08**	0.03	1.08
NUM	0.10	0.11	1.11	1.95***	0.32	7.03	0.05	0.09	1.05
AOT	0.08	0.21	1.08	1.29*	0.57	3.63	0.13	0.13	1.14
Matura	0.10	0.08	1.11	0.42 .	0.22	1.52	0.06	0.06	1.06

Note. Outcome variables are coded such that the first category (e.g. correct non-detections) is coded as 1 and the second category (e.g. correct detections) is coded as 0.

SD = Standard deviation; OR = Odds ratio; ICAR = International Cognitive Ability Resource; NUM = Numeracy; AOT = Actively open-minded thinking. $p < .10$; * $p < .05$; ** $p < .01$; *** $p > .001$

Again, the effects were the strongest when predicting "correct non-detections" vs. "incorrect detections", i.e. when differentiating between correct and incorrect BBS responses. In this regard, ICAR, Numeracy, and AOT significantly predicted the outcome: an increase of one point on these measures increased the odds of "correct detections" responses 2.01, 7.03, and 3.63 times respectively. However, this time neither of our individual difference variables managed to significantly differentiate correct non-detections from correct detections. Finally, similarly to the CRT, ICAR was the only significant predictor of the "incorrect detection" vs. "incorrect non-detections", but the effect was again rather modest. Namely, a one-point increase in ICAR increased the odds that the person who gave an incorrect response detected (vs. did not detect) the conflict 1.08 times.

Study 1 discussion

In Study 1, we tried to classify responses on CRT and BBS items based on their accuracy and probability of conflict detection. If the trial had a low probability of conflict detection and was accurate, we classified it as correct non-detection. Conversely, if it was inaccurate, we classified it as incorrect non-detection. If there were signs of conflict detection, based on their accuracy, trials were categorized as either correct or incorrect detections. We obtained some interesting differences in cognitive ability, numeracy,

disposition toward actively open-minded thinking, and mathematical knowledge score between these trials. These effects are shown in Figures 2 and 3. The categories on the x-axis are arranged so that they reflect a mindware instantiation continuum – the first category on the left (correct non-detection) should reflect the most developed mindware and the strongest logical intuition, while the last category (incorrect non-detection) reflects the least developed mindware and the weakest logical intuition. The general pattern seen in the figures, both for CRT and BBS items, is that as the supposed mindware quality/strength of logical intuition decreases, so do most of the individual differences variables. In doing so, the biggest differences are always between the correct and incorrect detections, indicating perhaps that these individual differences are important both for conflict detection, but also for deliberate correction through doing explicit calculations that draw on cognitive capacities and math knowledge. This is basically what a classical dual-process view or Stanovich et al.'s (2016) tripartite theory would suggest.

However, there were also some interesting differences regarding conflict detection/non-detection within incorrect and correct trials. Although our individual-difference variables did not particularly well differentiate between detections and non-detections on incorrect trials, the effects of intelligence were nevertheless significant both for CRT and BBS tasks. Intelligence was related to the ability to detect the conflict even among those that did not manage to solve the problems correctly. This perhaps means that there exists some minimal threshold of intelligence below which a person did not manage to develop the mindware relevant for these kinds of tasks even to be able to "feel" that something was wrong with their responses. This is perhaps due to lower capacity people not having interest and drive to engage in these types of tasks, therefore not being experienced in it. Or it could be due to these people not being able to drive meaningful conclusions and insights from these types of tasks and thus not being able to develop relevant intuitions that could help them solve these and similar tasks.

Among the correct trials, although the trend of the non-detection trials being related to higher scores than the detection trials was present, neither intelligence nor thinking dispositions were that important. Actually, for BBS, none of the measured individual difference variables were much important. However, for CRT trials, numeracy was the only variable that managed to significantly differentiate between the correct detections and non-detections, and these effects were non-negligible. The conclusion that could be drawn from these results is that, while cognitive abilities are important for mindware instantiation, they are not sufficient for developing strong and quality mindware and expert intuitions. To make this

"extra step", some kind of specialized knowledge is needed. Cognitive abilities and dispositions, although perhaps necessary for gaining this type of knowledge, are not sufficient. It probably takes a mix of abilities, motivation, and opportunity to engage with these types of tasks to obtain knowledge and experience rich enough to allow for very strong and correct intuitions. For BBS tasks, our individual differences variables did not capture this knowledge, but for CRT it seems that the numeracy as we measured it was exactly the type of knowledge and skills that reflects expertise needed for success. This aligns nicely with a number of studies showing very large correlations between the CRT and numeracy scores (e.g. Erceg, Galić & Ružojčić, 2020; Finucane & Gullion, 2010; Primi, Morsany, Chiei & Donati, 2016; Welsh, Burns & Delfabbro, 2013). In sum, each of our individual differences variables was important in developing quality mindware and strong logical intuitions, with intelligence perhaps being a prerequisite for developing minimal mindware, and strong numerical abilities a necessity for developing very strong mindware and logical intuitions for CRT problems.

However, as we already suggested, a single-response paradigm has substantial drawbacks in terms of the ability to capture conflict detection and deliberate correction processes. The main problem was that the participants were not instructed to read the items quickly and respond quickly and intuitively. This could have resulted in participants mostly taking their time to read the problems and responding carefully, but also in inconsistencies in reading times on different items (i.e. randomly taking more or less time depending on the concentration, boredom, other thoughts that might have occurred during reading, etc.). This first drawback significantly invalidates raw response times as a conflict detection indicator. We tried to mitigate this problem by substantially lowering the threshold to capture mostly those that did respond quickly and intuitively. However, this introduced further problem – we probably classified many of those that would respond quickly and intuitively correctly (had they been told to do so) as conflict detectors, possibly raising the average scores on our individual difference variables in the correct detection group and thereby blurring the differences between the correct non-detection and detection group. The second drawback is more related to the time difference as a conflict detection indicator. The inconsistencies in response times due to the lack of instruction to read the items and respond to them in a consistent way probably resulted in a high number of misclassifications due to luck/randomness. This again might have blurred the differences between our groups.

In response to these problems, we have conducted a second study, but this time based on the two-response paradigm. In this paradigm, participants responded two times on the same items: first with substantially

limited time and cognitive resources, followed by classical responding without any kind of limitations. This way it is possible to differentiate between correct non-detections and detections in a much more precise way.

Study 2

The goal of the second study was to investigate individual differences in abilities, knowledge, and dispositions that underpin different types of responses to reasoning tasks using the two-response paradigm. This paradigm allows us to more precisely differentiate between those that manage to respond correctly even when time and cognitive resources are severely limited from those that manage to respond correctly only when given enough time to think more carefully about the problem. The former type of response is usually coded as "11", meaning that the participant responded correctly both in fast and slow conditions. Participants who respond in this way probably have highly developed mindware and very strong logical intuitions (substantially stronger than incorrect intuitions) that allow them to respond intuitively correctly and in doing so detect a very little conflict. The latter type of response is usually coded as "01" as the first response was incorrect and the second was correct. Participants who respond in this way probably have somewhat less developed mindware and weaker logical intuitions, resulting in them first giving an incorrect response, but then detecting the conflict and correcting the erroneous intuition. Finally, the two-response paradigm also allows for detecting the third group of those that do not manage to respond correctly even after they spend time and cognitive resources on the problem (coded as "00"). This group probably has the least developed mindware and weakest logical intuitions, much weaker than the incorrect intuition. These differences in types of responding should be underpinned by differences in cognitive abilities, knowledge, and thinking dispositions and these differences are the focus of this study.

Methods

Participants

A total of 83 subjects participated in Study 2. Participants were university students from different University of Zagreb faculties, mostly females (N = 56), with a mean age of M = 22.8 (SD = 3.63).

Instruments

In this section, we will describe the instruments that we used in Study 2 that were either new or different from the ones we used in Study 1.

Cognitive reflection items (CRT). In Study 2, we used seven CRT-like items (see Appendix): three from the original CRT (Frederick, 2005), two from the Thomson and Oppenheimer (2016) version, one from the Primi et al. (2016) version, and one related to the illusion of linearity from Putarek and Vlahović-Štetić (2019). There were four response options for each item, containing a correct response, an intuitive incorrect response, and two other incorrect responses. Participants' task was to choose the response they think is the correct one. The total score was calculated as the average number of correct responses on the seven items.

Belief-bias syllogisms (BBS). One of the drawbacks of Study 1 was also the fact that we used only the BBS items with believable, but logically incorrect conclusions. In this study, we balanced the item pool by adding the three BBS items with unbelievable, but logically correct conclusions to the four believable, but logically incorrect items, totaling seven BBS items (see Appendix). All items were taken from Markovits and Nantel's (1989) study. The total score was calculated as the average number of correct responses on the seven items.

Actively open-minded thinking (AOT) scale. AOT scale was also different from the one we used in Study 1. This time, we opted for the, at the time of writing, recommended 10-item AOT scale (http://www.sjdm.org/dmidi/Actively_Open-Minded_Thinking_Beliefs.html). Participants rated their level of agreement with these items on a five-point scale (the items are available in the Appendix), and the total score was calculated as the average of ratings on 10 items after some of the items were recoded such that higher scores indicate higher AOT.

Need for cognition (NFC) scale. To measure NFC, we used a short, five-item scale (see Appendix for items; Cacioppo & Petty, 1982). Participants rated their level of agreement on a five-point scale, and the total score was calculated as the average of the ratings after some items were recoded such that higher ratings indicate higher NFC.

Apart from these, we also measured intelligence, numeracy, and matura score, but these instruments were the same as in Study 1.

Procedure

The study was conducted online using the Limesurvey software. Before solving the questionnaire, participants were gathered in small groups for a live online meeting in which an experimenter guided participants through all the instructions. The instructions were also written within the survey, but we insisted on a live meeting to make sure that all the participants read and properly understood them. The experimenter and participants went through several survey pages that contained the instruction together, with the experimenter reading the instructions aloud, and answering the participants' questions.

The exact wording for general and specific CRT and BBS instruction can be found in the Appendix B. In short, participants were told that they are about to solve some reasoning tasks, but that they will solve them twice. Both for CRT tasks and BBS tasks, they were told that they will first solve the tasks with a severe time limit and burdened capacity for deliberation (by a memory task), and immediately after that, in the slow condition, they will solve the same task again, only this time without any constraints. Following the instruction, participants solved two training tasks to see what the whole process looks like and to get the feeling of how little time they had to read and respond in a fast situation. There were several steps in this solving process. Participants were first presented with a 3 by 3 matrix that contained four black dots for five seconds (see Appendix B for an example of this matrix). They were instructed to memorize the pattern as they will be asked to recognize it among four different matrices later. Immediately after five seconds ran out, participants were automatically transferred to the next page that contained a CRT item with a time limit. Time limits were set based on average reading time for these items obtained through a small pre-study (N = 18)[5]. The limits were between 8 and 15 seconds for CRT items and between 7 and 9 seconds for BBS items, meaning that participants had that much time to read the item, read the response alternatives and respond. CRT tasks varied more in length than BBS tasks,

[5] As the feedback from the several "testing" participants was that the time limits were too short and that there was too little time to read the majority of the items, we increased the time limits to be slightly higher than the pre-study reading average. This was particularly the case for CRT items, and the reason for this mismatch is probably because in the pre-study we did not show participants four response options, but only the item stems, asking them to click "next" once they read the item. In the "real" survey, there were four response options, so participants had to read all the options before responding and this required some additional time. On average, across all the items, we increased time limits for 0.77 seconds.

which is why there is a greater time limit range for CRT tasks (e.g. the shortest item had 15 words whereas the longest item had 40 words). If we take response times in Study 1 where we did not employ time pressure as a reference, roughly only about 10% of participants in Study 1 managed to read and solve the items within these time limits, attesting that these limits were quite challenging for our Study 2 participants. Once the time was up, participants were again automatically transferred to the next page where we asked them how confident they were in their first response (0% - 100%). This was not timed. After indicating the degree of confidence in their results, participants had to recognize the matrix they memorized among the four options. Once they responded to this memory task, they moved on to the next page where they responded to the same CRT item, only this time without the time limit. Finally, once they chose their response, they were again asked to indicate their degree of confidence in their second response. This sequence was the same for CRT and BBS items. Participants always solved CRT items first and BBS items after. After solving seven CRT and seven BBS tasks in fast and slow conditions, participants solved 16 ICAR items, four numeracy items, a 10-item AOT scale, and a five-item NFC scale. The survey ended with several demographic questions.

Results

We will first present descriptive statistics and correlations between our focal variables at the participant level. The main analysis of differences between the three types of responses ("11", "01", "00") was done at the item level, analogous to the analyses from Study 1. We note here that one more response type is theoretically possible in the two-response paradigm and that is the "10" response (intuitive correct and then deliberate incorrect response). However, these types of responses are nonsensical and, expectedly, were extremely rare in our case. Among CRT trials, there were only six out of 499 valid trials (meaning the trials that were given within the time limit) that were responded to in this way, while among the BBS trials there were none of the "10" cases. Therefore, we discarded the six CRT trials and conducted further analyses on the three remaining categories. Descriptive statistics for our focal variables are shown in Table 8, and the correlations among our focal variables in Table 9.

Table 8. Descriptive statistics of Study 2 focal variables

	M	SD	Min	Max	Cronbach α
CRT fast	0.40	0.26	0	1	.66
CRT slow	0.66	0.29	0	1	.77
BBS fast	0.59	0.26	0	1	.76
BBS slow	0.65	0.29	0	1	.83
ICAR	10.37	3.04	0	16	.75
NUM	1.60	1.23	0	4	.56
AOT	3.98	0.41	2.70	5	.50
NFC	3.41	0.81	1.40	5	.83
Matura	5.01	1.25	2	7.50	/
CRT fast conf.	59.16	19.73	2.86	95.71	.58
CRT slow conf.	90.78	13.14	10.00	100	.79
CRT slow time	29.37	18.54	9.95	104.20	.68
BBS fast conf	75.15	20.24	1.43	100	.81
BBS slow conf.	92.01	13.67	9.29	100	.91
BBS slow time	14.33	6.40	5.60	35.27	.57

Note. CRT fast = Cognitive reflection test average score in the fast-responding situation; CRT slow = Cognitive reflection test average score in the slow-responding situation; BBS fast = Belief bias syllogisms average score in the fast-responding situation; BBS slow = Belief bias syllogisms average score in the slow-responding situation; ICAR = International cognitive ability resource; NUM = numeracy; AOT = Actively open-minded thinking; NFC = Need for cognition; CRT fast conf. = Average confidence in responses on cognitive reflection tasks in the fast-responding situation; CRT slow conf. = Average confidence in responses on cognitive reflection tasks in the slow-responding situation; CRT slow time = Average response time in seconds on cognitive reflection tasks in the slow-responding situation; BBS fast conf. = Average confidence in responses on belief bias syllogisms in the fast-responding situation; BBS slow conf. = Average confidence in responses on belief bias syllogisms in the slow-responding situation; BBS slow time = Average response time in seconds on belief bias syllogisms in the slow-responding situation.

Table 8 shows that the pattern of the results generally aligns with the expectations. The participants were better at reasoning tasks in the slow situation, when they had time to think about their responses, than in the fast situation, both for CRT items ($t = 9.07$, $p < .001$, $d = 1.00$) and BBS items ($t = 3.97$, $p < .001$, $d = .37$). This was accompanied by generally lower confidences in responses when the time was limited compared to the "slow" condition ($t = 17.94$, $p < .001$, $d = 1.98$ for CRT and $t = 11.61$, $p < .001$, $d = 1.29$).

Table 9. Correlation among the Study 2 focal variables

	CRT s.	BBS f.	BBS s.	ICAR	NUM	AOT	NFC	Matura	CRT f. c.	CRT s. c.	CRT s. t.	BBS f. c.	BBS s. c.	BBS s. t.
CRT f.	.56**	.25*	.31**	.33**	.54**	.05	.30**	.28**	.33**	.20	-.18	.08	.14	.12
CRT s.		.42**	.48**	.60**	.60**	.13	.20	.49**	.17	.28*	.27*	.05	.13	.38**
BBS f.			.83**	.43**	.40**	.40**	.20	.39**	-.02	-.01	.17	-.08	.00	.10
BBS s.				.41**	.50**	.33**	.18	.43**	.00	-.02	.19	-.10	.01	.30**
ICAR					.48**	.22*	.24*	.42**	.22	.32**	.22	.22*	.30**	.22*
NUM						.31**	.22*	.36**	.22*	.26*	.13	.09	.22	.33**
AOT							.26*	.05	.20	.16	.11	.06	.06	.10
NFC								.14	.20	.31*	-.09	.16	.21	-.06
Matura									.22	.22	.02	.06	.09	.19
CRT f. c.										.59**	-.14	.50**	.39**	-.03
CRT s. c.											-.02	.57**	.69**	.13
CRT s. t.												-.02	.07	.37**
BBS f. c.													.77**	-.29**
BBS s. c.														-.06

Note. CRT f. = Cognitive reflection scores in the fast-responding situation; CRT s. = Cognitive reflection scores in the slow-responding situation; BBS f. = Belief bias syllogisms scores in the fast-responding situation; BBS s. = Belief bias syllogisms scores in the slow-responding situation; ICAR = International cognitive ability resource; NUM = Numeracy; AOT = Actively open-minded thinking; NFC = Need for cognition; CRT c. f. = Average confidence in responses on cognitive reflection tasks in the fast-responding situation; CRT c. s. = Average confidence in responses on cognitive reflection tasks in the slow-responding situation; CRT t. s. = Average response time on cognitive reflection tasks in the slow-responding situation; BBS c. f. = Average confidence in responses on belief bias syllogisms in the fast-responding situation; BBS c. s. = Average confidence in responses on belief bias syllogisms in the slow-responding situation; BBS t. s. = Average response time on belief bias syllogisms in the slow-responding situation.

There are several interesting findings apparent in Table 9. First, there were very high correlations between the first and second responses, especially for BBS items. This indicates that for both tasks, when the correct response was given, it was very often given already in the fast condition. Second, not surprisingly, CRT and BBS scores in the slow condition were positively correlated with all of our individual difference variables, with some of these correlations being quite high (e.g. relationships of CRT/BBS with ICAR and numeracy). These high correlations between cognitive abilities and CRT/BBS are not surprising as similar correlations are often reported in the literature (e.g., Blacksmith, Yang,

Behrend and Ruark 2019; Toplak, West & Stanovich, 2011; Thomson & Oppenheimer, 2016). However, what is interesting are moderate to high correlations between our individual difference variables and success on CRT/BBS in the fast condition. This means that cognitive abilities and to a degree thinking dispositions are important not only in detecting the conflict and correcting erroneous intuitive responses but also in generating correct intuitive responses. Third, cognitive abilities and thinking dispositions were also to some degree predictive of confidence in responses and response times. However, there were also some interesting insights here. For example, numeracy was a significant predictor of confidence in CRT responses in fast condition, but not of confidence in BBS fast responses. This possibly reflects the importance of numeracy for strong CRT logical intuitions, but not so much for strong BBS logical intuitions. Finally, smarter and more numerate participants on average tended to take more time when responding in the slow condition, and this was more pronounced when solving BBS tasks.

In a bid to answer our main research questions, before repeating the analyses that we did in Study 1, we first decided to replicate the analysis reported by Raoelison et al. (2020). They correlated cognitive abilities with proportions of "11", "01" and "00" responses at the participant level (i.e., for each participant the number of his/her responses in each category was divided by the total number of his/her responses). We followed the same approach here. Therefore, in Table 10 we are showing the correlations between our focal variables and the proportion of "11", "01", and "00" responses.

Table 10. Correlations among the focal variables and proportions of response categories

	ICAR	NUM	AOT	NFC	MAT	CRT f. c.	CRT s. c.	CRT s. t.	BBS f. c.	BBS s. c.	BBS s. t.	BBS 00	BBS 01	BBS 11
CRT 00	-.53**	-.57**	-.07	-.21	-.45**	-.09	-.18	-.23*	.06	-.05	-.35**	.44**	-.13	-.40**
CRT 01	.28*	.08	.02	-.09	.25*	-.28*	.00	.49**	-.17	-.11	.30**	-.21	.07	.18
CRT 11	.38**	.56**	.06	.31**	.31**	.34**	.22	-.16	.12	.19	.13	-.31**	.05	.30**
BBS 00	-.38**	-.48**	-.32**	-.13	-.40**	.09	.11	-.19	.23*	.08	-.29**		-.37**	-.86**
BBS 01	-.07	.17	-.12	-.17	.03	-.19	-.24*	.04	-.39**	-.23*	.40**			-.15
BBS 11	.44**	.42**	.41**	.23*	.41**	.00	.02	.18	-.03	.04	.09			

Note. CRT 00 = Proportion of cognitive reflection "00" responses; CRT 01 = Proportion of cognitive reflection "01" responses; CRT 11 = Proportion of cognitive reflection "11" responses; BBS 00 = Proportion of belief bias syllogisms "00" responses; BBS 01 = Proportion of belief bias syllogisms "01" responses; BBS 11 = Proportion of belief bias syllogisms "11" responses; ICAR = International cognitive ability resource; NUM = Numeracy; AOT = Actively open-minded thinking; NFC = Need for cognition; CRT c. f. = Average confidence in responses on cognitive reflection tasks in the fast-responding situation; CRT c. s. = Average confidence in responses on cognitive reflection tasks in the slow-responding situation; CRT t. s. = Average response time on cognitive reflection tasks in the slow-responding situation; BBS c. f. = Average confidence in responses on belief bias syllogisms in the fast-responding situation; BBS c. s. = Average confidence in responses on belief bias syllogisms in the slow-responding situation; BBS t. s. = Average response time on belief bias syllogisms in the slow-responding situation.

Table 10 reveals what Table 9 hinted at: cognitive abilities, knowledge, and thinking dispositions are primarily responsible for forming strong logical intuitions that allow people to respond quickly and intuitively correctly on BBS and CRT tasks. Of course, in addition to this, they were also highly predictive of serious mindware deficiencies reflected in "00" responses, with thinking dispositions playing a somewhat smaller role here. This seems logical to us: without a sufficient degree of cognitive abilities and knowledge, no amount of disposition to think hard and carefully will be enough to come up with a correct response to these tasks. Another interesting thing to notice is that the confidence in fast responses was positively correlated with the proportion of "11" responses and negatively with the proportion of "01" responses both for CRT and BBS tasks. This means that participants with more "11" responses were more confident in their initial responses, indicating that they experienced lower conflict

between logical and erroneous intuition. Conversely, lower confidence in initial responses predicted the probability of correcting erroneous intuition, indicating a conflict detection in "01" responses. This was also accompanied by longer response times meaning that the participants who detected the conflict and corrected their erroneous initial response took more time to think about the task and respond to it. Finally, it was interesting to see that there exists a moderate correlation between the proportion of "00" and "11" responses across the two tasks. Participants with more "11" responses on CRT tasks also tended to have more "11" responses on BBS while those with more "00" CRT responses tended to also have more "00" BBS responses. Interestingly, there was no correlation in the proportion of "01" responses between the tasks.

Finally, we conducted our main analyses to investigate the differences in abilities, dispositions, and knowledge that underpin the three response types, "11", "01" and "00". Here it must be noted that around 13% of the CRT trials and 7% of the BBS trials were discarded before analyses due to missing the response deadline. [6] The two tasks somewhat differed in the frequency of each of the three categories. For CRT, there were 189 "11" trials, 145 "01" trials, and 159 "00" trials. For BBS, there were 302 "11" trials, 56 "01" trials, and 171 "00" trials. Therefore, while 43% of the CRT correct trials were answered through deliberation (57% of correct responses were given from the start, intuitively), only 16% of the BBS correct trials were given after deliberation. As in Study 1, we again conducted multilevel logistic regression analyses, but this time with two dichotomous outcome variables ("11" vs. "01" and "01" vs. "00"). The results of the analyses are shown in Table 11. To foster the interpretation of results, we plotted the means and confidence intervals for each of the individual-difference variables for each response type in Figure 4.

[6] In an additional 6% of CRT trials and 4% of BBS trials participants failed the memorization task, i.e. failed to recognize the matrix they needed to memorize. However, as the results were virtually the same with or without these failed memorization trials, we have decided to keep them in the analyses. Therefore, the only trial we discarded were the ones where the deadline was missed.

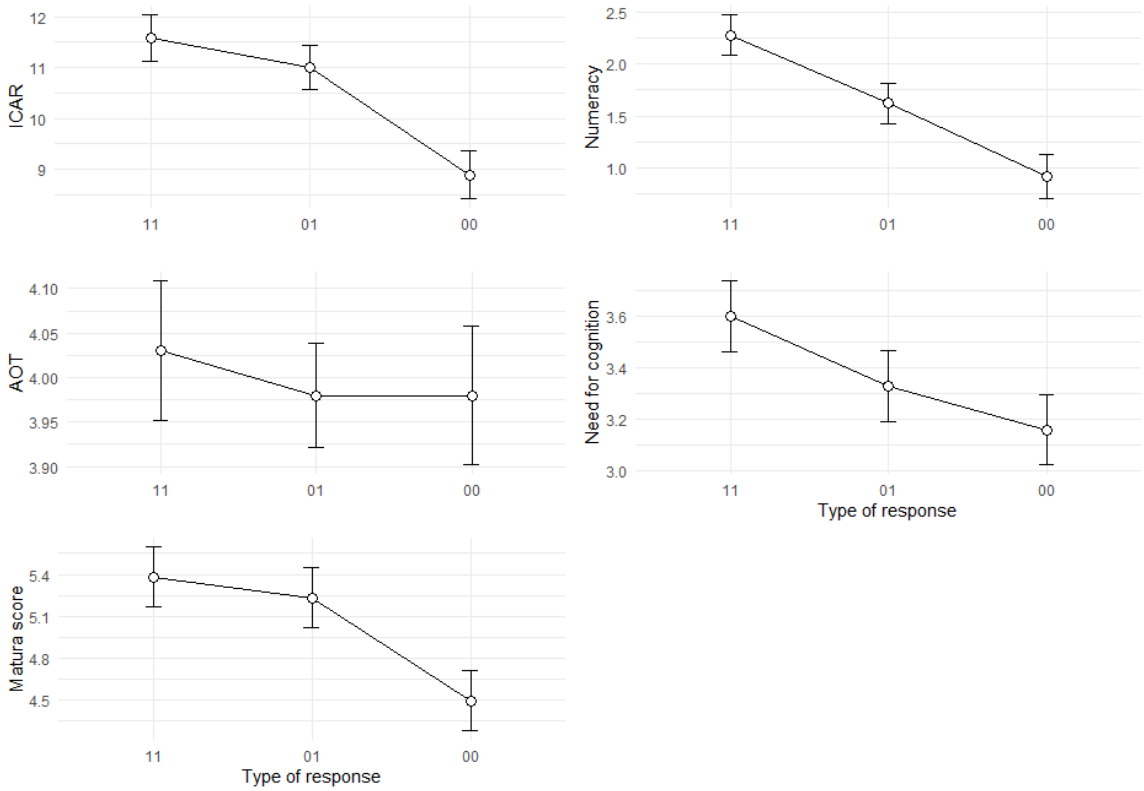


Figure 4. Differences in ICAR, numeracy, need for cognition, and matura score between different response types for CRT tasks

Table 11. Results of the multilevel logistic regression analyses for Cognitive reflection and Belief-bias syllogism tasks

	CRT						BBS					
	"11" vs. "01"			"01" vs. "00"			"11" vs. "01"			"01" vs. "00"		
	B	SD	OR	B	SD	OR	B	SD	OR	B	SD	OR
ICAR	0.05	0.05	1.05	0.30***	0.06	1.35	0.12 .	0.08	1.13	0.23	0.15	1.26
NUM	0.35***	0.10	1.42	0.73***	0.17	2.08	-0.05	0.15	0.95	1.22**	0.37	3.39
AOT	0.27	0.35	1.31	0.30	0.48	1.35	0.95*	0.43	2.59	1.18	1.01	3.25
NFC	0.39*	0.17	1.48	0.22	0.23	1.25	0.38 .	0.22	1.46	-0.13	0.46	0.88
Matura	0.02	0.11	1.02	0.68***	0.17	1.97	0.12	0.15	1.13	0.64*	0.31	1.90
Confidence	0.05***	0.01	1.05	-0.02***	0.004	0.98	0.04***	0.01	1.04	-0.08***	0.02	0.92
Time	-0.06***	0.01	0.94	0.02***	0.003	1.02	-0.08***	0.01	0.92	0.11***	0.02	1.12

Note. Outcome variables are coded such that the first category (e.g. “11”) is coded as 1 and the second category (e.g. “01”) is coded as 0.
SD = Standard deviation; OR = Odds ratio; CRT = Cognitive reflection test; BBS = Belief bias syllogisms; ICAR = International Cognitive Ability Resource; NUM = Numeracy; AOT = Actively open-minded thinking; NFC = Need for cognition; Confidence = Initial confidence in fast responding condition; Time = Response time in slow responding condition.
. $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

A general trend was that intuitive correct responses (“11”) were associated with the highest abilities, dispositions, and knowledge, followed by “01” responses and lastly by “00” responses. However, not all the differences between the categories were significant. For example, only numeracy and NFC managed to significantly differentiate between “11” and “01” responses on CRT, with a one-point increase on these measures leading to a 1.42 and 1.48 times increase in odds of “11” responses compared to “01” responses. Regarding the “01” and “00” categories for CRT, the effects of ICAR (1.35 times increase in odds of “01” compared to “00” responses), numeracy (2.08 times increase in odds of “01” compared to “00” responses) and matura score (1.97 times increase in odds of “01” compared to “00” responses) were significant, while the effects of dispositions (AOT and NFC) were not.

For the BBS tasks, “11” and “01” responses were significantly differentiated only by AOT (a one-point increase on the AOT scale leading to 2.59 times increase in odds of “01” compared to “00” responses). The problem here was that “01” responses on BBS were heavily underrepresented which increased the standard errors and negatively affected the statistical power to detect the effects. This is the reason why two more effects, although not negligible, were only marginally significant (the effects of ICAR and especially NFC). “00” and “01” responses were significantly predicted only by numeracy (OR = 3.39) and matura score (OR = 1.90). However, the problem of low power is even more pronounced here as even relatively strong effects (e.g. for AOT, OR = 3.25) were non-significant due to large standard errors. Again, to ease the interpretation, we have plotted the means and confidence intervals for each of the individual-difference variables for each response type on BBS tasks in Figure 5.

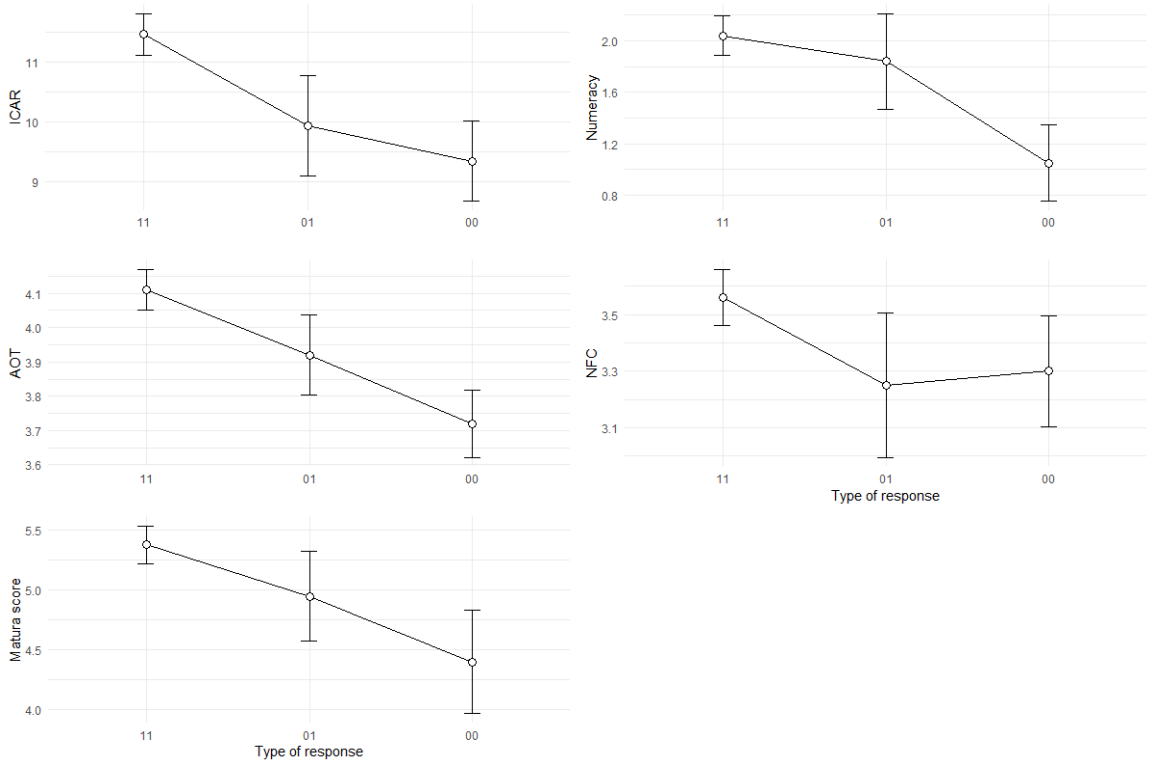


Figure 5. Significant differences in ICAR, numeracy, actively open-minded thinking, need for cognition, and matura score between different response types for BBS tasks

In addition to these analyses, we wanted to see how initial confidences (in the fast condition) and final response times (in the slow condition) differed across the three types of responses as these variables are indicative of conflict detection and deliberate correction of erroneous intuitions. To do this, we again conducted multilevel logistic regressions, this time with initial confidences and final response times as predictors. In short, both confidence and time were significant predictors of "11" vs. "01" and of "01" vs. "00" response types. These results are shown in the last two rows of Table 11 and plotted in Figure 6.

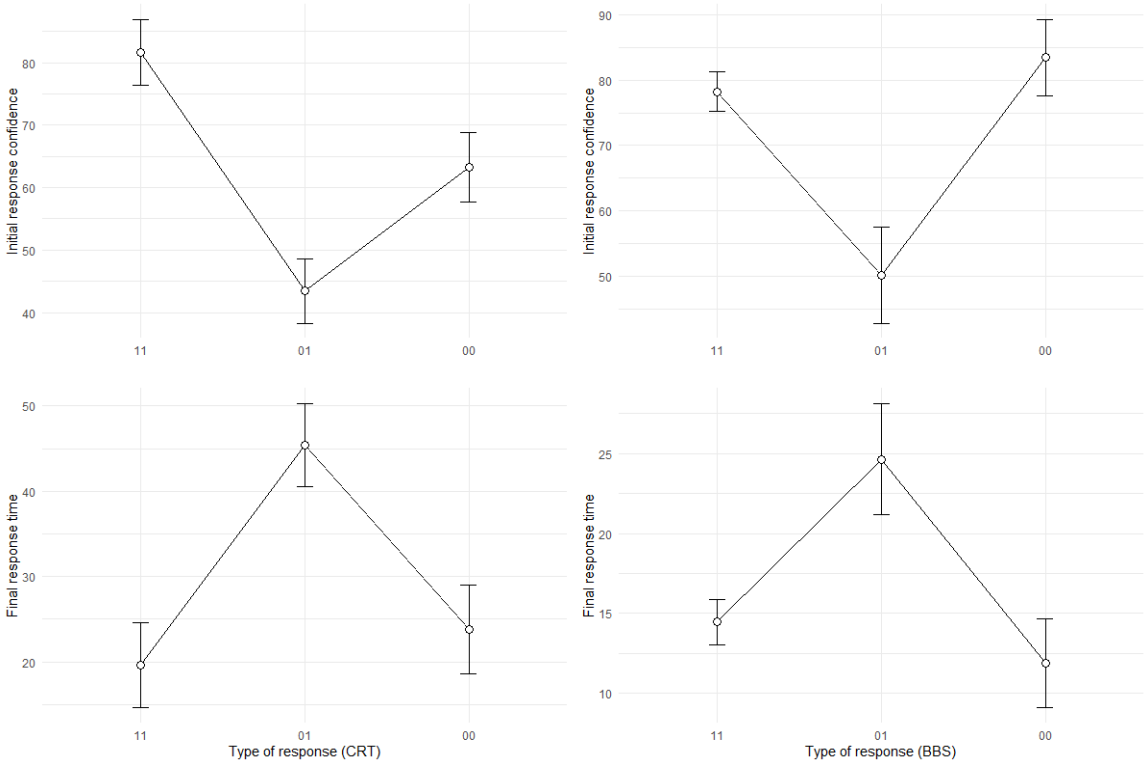


Figure 6. Average confidences in responses in the fast condition and average response times in the slow conditions for three different response types and the two reasoning tasks

As can be seen from Figure 6, the pattern of confidence and response time differences is similar for both tasks. Specifically, responses where participants managed to correct initially incorrect responses were related to lowest initial confidence and highest final response time, indicating that initial skepticism in one's response, possibly due to high conflict between the intuitions, predicted taking more time in the slow condition to correct the erroneous intuition and calculate the right response. These results basically replicate previous findings on the relationships between response confidence, rethinking times, and response change (e.g. Bago & De Neys, 2017; Thompson et al., 2011). On the other hand, responses where participants did not manage to respond correctly even after deliberation almost mimic the responses where participants gave a correct response from the start. This indicates that in "00" responses participants had very little idea or feeling that something was wrong and that they should slow down and think about the problem more carefully. This is indicative of underdeveloped mindware and weak correct, logical intuition. Conversely, this pattern of responding (high initial confidence, low final response time) in "11" responses is indicative of highly developed mindware and very strong logical intuitions. Participants who responded in this way either did not bother much to carefully check their responses in slow condition as they knew they were right from the start or their high abilities and knowledge allowed

them to run appropriate calculations and verify that they are right quickly, much quicker than participants with "01" responses could.

Study 2 discussion

The two-response study showed that substantial differences in cognitive abilities, thinking dispositions, and math knowledge underpin different types of responses to the two reasoning tasks. What is interesting is that these psychological traits not only differentiate between correct and incorrect responses but also between different types of correct responses. Specifically, correct intuitive responses ("11") on CRT items were given by responders that were substantially more numerate and predisposed towards analytical thinking than those that gave deliberate correct responses. As we already noted, this suggests that numeracy is one of the crucial ingredients of strong mindware relevant for CRT tasks, but also that disposition towards analytical thinking might be one of the crucial dispositions that help in the development of such mindware. For BBS tasks, these two types of correct responses did not differ in numeracy which makes sense as BBS tasks do not require math knowledge. However, thinking dispositions again substantially differed between participants with "11" and "01" responses. Thus, although we did not capture specific knowledge and skills (i.e. mindware) relevant for BBS tasks in our studies, our findings suggest that dispositions toward analytical thinking and seeking alternative evidence and point of view (as indicated by the AOT score) help in the development of such mindware. How these results overlap with the results of Study 1 and what are their implications for the role of abilities and dispositions in reasoning tasks, as well as for the validity of these reasoning tasks, will be discussed next.

General discussion

Across two studies we examined the differences in cognitive abilities, numeracy, math skills, and thinking dispositions (actively open-minded thinking and need for cognition) associated with different ways of solving two types of reasoning tasks, cognitive reflection tasks and belief-bias syllogisms. In Study 1, based on the accuracy of responses and indicators of conflict detection, we classified responses in four categories (correct non-detections, correct detections, incorrect non-detections, and incorrect detections), while in Study 2 we classified them in three categories (correct intuitive responses, correct deliberate responses, and incorrect responses).

Our two studies showed a certain degree of overlap in their conclusions, but there were also some differences. We will try to explain both. Notably, Study 2 replicated the results of Study 1 regarding the importance of numeracy for CRT tasks. Numeracy was the key individual difference that differentiated between the types of responses on CRT tasks in both studies, not only between correct and incorrect responses but also between the intuitive and deliberate correct responses. However, this was not the case for BBS tasks where it did not differ between intuitive and deliberate correct responses. This points to the conclusion that numeracy as we measured it is particularly indicative of the quality of mindware for CRT tasks. Not only does it capture cognitive abilities and dispositions important for success on these tasks, but it also assesses relevant math knowledge and experience indicative of highly developed mindware (Cokeley et al., 2012; Erceg et al., 2020; Skagerlund, Lind, Strömbäck, Tinghög, & Västfjäll, 2018; Sobkow, Olszewska, & Traczyk, 2020). As Ghazal, Cokely, and Garcia-Retamero (2014) noted, numeracy is a potent predictor because it simultaneously assesses important metacognitive skills and mathematical competency. These results also work well with Peters et al.'s (2006) findings that highly numerate people had more clear feelings about what response option was correct on a ratio-bias task, a task that assesses judgments of probabilities. Thus, highly numerate people seem to be more comfortable with numbers and to literary have "feelings" for numbers that might allow them to intuitively perceive the degree of the rightness of response in tasks that depend on math knowledge. Therefore, they can tell the right response from the wrong, even if they are not able to explain their decision without more careful deliberation (Bago & De Neys, 2019). As numerical/math knowledge is not particularly important for BBS tasks, numeracy was not indicative of highly developed mindware that allows for instant/intuitive correct responding, but given that it also captures cognitive abilities and dispositions relevant for success on BBS tasks, it was still able to differentiate between generally correct and incorrect responses.

The results are also in line with Reyna and colleagues (Reyna, Nelson, Han & Dieckmann, 2009; Reyna, Rahimi-Golkhandan, Garavito & Helm, 2017) fuzzy-trace theory that distinguishes between verbatim and gist representations of information. Verbatim representations of information are similar to information as presented, while gist representations of information refer to deriving the essential meaning of that same information (Reyna et al., 2017). Research has shown that people encode verbatim representations as well as multiple gist representations of the same information, with the gist-based intuition, not the verbatim representations, being an advanced model of reasoning (Reyna et al., 2009). This gist-based intuition is developed over time by processing information in a meaningful way. These meaningful insights shape gist representations and enable the transfer of knowledge to similar, but new

stimuli. Therefore, as people gain experience at specific tasks, they tend to rely more on gist rather than verbatim representations, and these gist representations often result in better judgments and decisions (Reyna et al., 2017). We, of course, cannot say much about underlying representations of CRT and BBS tasks given our methodological approach, but this is something that would merit future research.

Regarding the BBS tasks, there are some differences between the conclusions of Study 1 and Study 2. While in Study 1 neither of the individual differences we captured differed between correct detection and correct non-detection responses (i.e. intuitive correct and deliberate correct responses), in Study 2 AOT was the only measure that differed between these types of responses, with the effects of intelligence and NFC being "marginally" significant (we only mention this because the sample of "01" responses on BBS was quite low, which diminished statistical power). As we said previously, we believe that this means that the cognitive abilities and thinking dispositions that we measured in this study are important for the acquisition of adequate mindware for BBS tasks. Thus, we did not capture the relevant mindware (as we perhaps did for CRT by our numeracy measure), but we did capture traits conducive to its acquisition. We will argue later that this has substantial implications for the role of cognitive abilities and thinking dispositions in the success on popular reasoning tasks.

Our study offers additional evidence for a reconceptualization of the role of cognitive abilities and thinking dispositions in success in judgment and decision-making tasks. As Raoelison et al. (2020) note, the classical "smart deliberator" view posits that those smarter and prone to analytical thinking are better at CRT, BBS, and other famous reasoning tasks because they are a) more careful and tend to think deeper about the problems which allows them to detect that they are being lured into incorrect response, and b) smarter, which allows them to easily arrive at the right response once they have detected these lures. This is the view also favored by the tripartite theory (Stanovich et al., 2016) where both the reflective mind (i.e. thinking dispositions) and algorithmic mind (i.e. cognitive capacities) are at work when an individual correctly solves reasoning tasks.

Results from several recent studies (e.g. Raoelison et al., 2020; Schubert et al., 2021; Thompson et al., 2018), as well as ours, shed additional light on the role of abilities and dispositions and question this "smart deliberator" view. Specifically, from the perspective of the hybrid, logical intuition model, the main role of abilities and dispositions is in acquiring the relevant and quality mindware for solving these reasoning tasks. Thus, instead of helping people to recognize they are lured into wrong responses and to

calculate/identify the right response, these traits act as building blocks of relevant mindware that is acquired through repeated exercise and exposure. Dispositions towards analytical thinking (e.g. need for cognition, actively open-minded thinking) ensure that a person enjoys engaging with tasks that require relevant knowledge and/or hard thinking over a prolonged period and cognitive abilities ensure that he/she can extract relevant and transferrable knowledge and meaning from these tasks. Therefore, over time, people that are higher in thinking dispositions and cognitive abilities manage to develop adequate, relevant, and strong mindware and intuitions, for these types of tasks, allowing them to rely more on their feelings and intuitions than on deliberate processing of specific information. This, and not careful thinking and deliberate calculation, is what drives the correlation between cognitive abilities/thinking dispositions and success on these reasoning tasks.

However, it is possible to go a step further here and argue that this relationship also depends on the opportunity to engage with such tasks. For example, the educational system provides ample opportunity for a person to engage with various math tasks and even basic logic. Thus, if one has the disposition to engage with these tasks on a deeper level and the capacity to draw adequate lessons from tasks, the prerequisites for developing strong mindware over time are set. However, when such an opportunity is not readily available, then the repeated engagement is much harder and much more dependent on the individual. This latter situation is less conductive of the development of sufficient mindware and strong logical intuitions. A comparison between recent findings based on think-aloud protocols of classic numerical CRT and verbal CRT support this view. Szaszi et al. (2017) showed that, when solving numerical CRT, the majority of responders who responded correctly immediately gave a correct response, or at least started with a line of reasoning that leads to a correct response, without mentioning the incorrect intuitive response. Conversely, using the think-aloud protocol in their study on verbal CRT, Byrd, Joseph, Gongora, and Sirota (2021) found that the majority of the correct responses were not given intuitively, but instead involved reflection. An example of the verbal CRT item is “How many of each animal did Moses put on the ark?” (Moses did not put any animals on the ark, it was Noah; Sirota, Dewberry, Juanchich, Valuš & Marshall, 2021). Following our previous explanation, we would argue that this discrepancy can be explained by differences in opportunity to engage in these two types of tasks and, thus, differences in mindware instantiation. Unlike CRT and BBS tasks for which the relevant skills can be obtained throughout schooling, experience and skills needed for verbal CRT are not obtained through formal schooling (but perhaps through being fooled multiple times on the playground or similar

places). Therefore, many people can have strong intuitions for CRT, BBS, ratio-bias, or base-rate neglect tasks, but not for verbal CRT tasks (and perhaps for other stumpers and riddles).

Related to this, our final point is that our results have also implications for the validity of (at least) CRT and BBS tasks as measures of reflection, analytical thinking engagement, or miserly processing (e.g. Böckenholt, 2012; Frederick, 2005; Pennycook, Cheyne, Koehler & Fugelsang, 2016; Pennycook, Fugelsang & Koehler, 2015; Pennycook & Ross, 2016; Toplak et al., 2014). Given the plethora of findings from the two-response paradigm that show that, when solved correctly, CRT, BBS, and some other tasks are solved in the majority of cases intuitively correctly, without the need to engage in deliberate thinking, it is questionable to what degree a correct response on such tasks is indicative of reflective or analytical thinking. In fact, it seems that only a minority of people respond to these tasks through careful deliberation and engagement in analytical thinking. One obvious repercussion of this is that these tasks are not particularly good measures of analytical thinking as they are mostly solved intuitively.

However, given our results, another possibility is that these tasks are still able to capture disposition towards analytical thinking, only not through the mechanisms previously thought. As explained earlier, we believe that the correlations between thinking dispositions and reasoning tasks are not the result of high disposition people being more likely to detect and override the conflict, but the result of thinking dispositions being conducive to more developed mindware over time. Therefore, even though these tasks are mostly solved by relying on intuition, they are still able to capture the propensity to generally be a careful, more deliberate, and analytic thinker. Some other tasks (for example, abstract reasoning tasks with which a person has little experience or had little opportunity to engage with) might be better indicators of analytical and reflective thinking in the sense that they require deeper engagement and propensity toward analytical thinking to be solved correctly. One last implication of this is that, for the task to be a good indicator of reflective/analytical thinking, the lures are probably not crucial. This is exactly what several recent findings showed (e.g. Attali & Bar-Hillel, 2020; Baron et al., 2015; Erceg, Galić & Ružojčić, 2020). To be solved correctly every cognitive task with which one does not have ample experience or skills will draw both on cognitive capacities and on thinking dispositions to be careful and reflective. It is impossible or at least very hard to develop a task that would be free from the effects of thinking dispositions. This is also at the core of Baron's (1985) definition of intelligence as something

that includes capacities (more stable and less prone to change) and dispositions (more prone to change and teaching attempts).

Conclusion

In sum, it seems that different problem solvers solve reasoning problems in qualitatively substantially different ways and that the differences in intelligence, numeracy, math skills, and thinking dispositions underpin these different approaches. Those lowest on these attributes have underdeveloped mindware and weak logical intuitions, sometimes not even strong enough to make them question their first response. The higher these attributes are, the better the mindware and stronger the correct intuitions, resulting in first detecting the conflict and eventually even overturning the initial wrong response. Finally, among those highest in cognitive abilities, numeracy, math skills, and actively open-minded thinking, mindware is so developed and correct intuitions are so strong that a correct response is given instantly, with very little conflict detection. This sequence is in line with the logical intuitions model and mindware instantiation continuum (Stanovich, 2018; Purcell et al., 2020), but does not fit well with the classic dual-processes narrative that posits that correct responses must come from deliberate and time-consuming Type 2 processing. These results have implications both for the role of cognitive abilities and thinking dispositions in task performance and for the validity of CRT and BBS as measures of analytical thinking or reflection. Smarter and those more prone towards analytical thinking are good CRT/BBS solvers not because these traits allow them to detect the conflict and correct the erroneous intuition, but because they are conductive of attaining better mindware and stronger logical intuitions. As these intuitions enable correct intuitive responding in most cases, CRT and BBS are hardly good measures of reflection/analytical thinking.

4. STUDY 3: NORMATIVE RESPONDING ON COGNITIVE BIAS TASKS - SOME EVIDENCE FOR A WEAK RATIONALITY FACTOR THAT IS MOSTLY EXPLAINED BY NUMERACY AND ACTIVELY OPEN-MINDED THINKING

This chapter was previously published as: Erceg, N., Galić, Z., & Bubić, A. (2022). Normative responding on cognitive bias tasks: Some evidence for a weak rationality factor that is mostly explained by numeracy and actively open-minded thinking. *Intelligence, 90,* 101619.

Introduction

In the last few decades, a number of papers, both empirical and conceptual, advocated for broadening of the study of cognitive abilities by including concepts and constructs from the domain of decision-making (Baron, 1985; Stankov, 2017; Stanovich, 2009a, 2009b, 2012; Stanovich & West, 1998, 2000, 2008). There have been some indications that tasks that measure different cognitive biases (CBs) capture something other than fluid intelligence, a construct that is labelled as rationality (e.g. Stanovich & West, 1998, 2000, 2008; Stanovich, West & Toplak, 2016) or decision-making competence (DMC, e.g., Bruine de Bruin, Parker & Fischoff, 2007; Parker & Fischoff, 2005), and as such could enrich our understanding of individual differences in cognitive processing.

In line with this, the first goal of our study was to investigate a factorial structure of a set of CB tasks in a search for existence of a rationality factor. Many of the previous studies showed that the correlations between different CBs are generally very low and that rationality assessed using CB tasks has a complex factorial structure with a minimum of two to three factors needed to sufficiently account for the common variance among the tasks (e.g. Aczel, Bago, Szollosi, Foldes & Lukacs, 2015; Berthet, 2021; Berthet & de Gardelle, 2021; Ceschi, Constantini, Sartoti, Weller & Di Fabio, 2019; Slugoski, Shields & Dawson, 1993; Teovanović, Knežević & Stankov, 2015; Weaver & Stewart, 2012). Given these results, it was reasonable to expect that a multifactorial solution will be needed to appropriately account for the relations between individual CBs included in the study. Alternatively, if a single-factor solution turns out to be the most appropriate, which is possible given that previous studies failed to find any systematicity regarding CBs factorial structure, this factor would probably be weak and account for a modest amount of variance among the individual tasks.

Our second goal was to investigate the relationships between individual CBs and rationality factor(s) with different cognitive abilities such as fluid intelligence and numerical ability. Although Stanovich and

West (2008) showed that some CB tasks are independent from cognitive abilities, the majority of studies demonstrated that the correlations between CBs and cognitive abilities are low to moderate (e.g. Bruine de Bruin et al., 2007; Erceg, Galić & Bubić, 2019; Parker & Fischoff, 2005; Teovanović et al., 2015; Toplak, West & Stanovich, 2011). Still, a recent study by Blacksmith, Behrend, Dalal & Hayes (2019) even found a correlation between decision-making competence (which is a combination of CB tasks and other tasks that are not generally considered to assess cognitive biases) and general mental ability as high as to declare them to be empirically redundant. The conflicting findings pointed to a need for additional studies using different measures and additional research contexts.

Finally, our third goal was to validate extracted rationality factor(s) by correlating them with variables from their nomological network (Cronbach & Meehl, 1955), such as superstitious and conspiracy beliefs, thinking dispositions and personality traits (convergent validity), as well as with potential real-life consequences of decisions (criterion validity). Previous studies showed that better performances on CB tasks are related to lower susceptibility towards epistemically suspect beliefs (superstitious/ paranormal/conspiracy beliefs; Čavojova, Šrol & Jurkovič, 2020; Erceg et al., 2019; Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2015; Šrol, 2020, Toplak et al., 2017) but greater orientation towards actively open-minded thinking (Stanovich & West, 1997, 1998; Sá, West, & Stanovich, 1999; Toplak, West & Stanovich, 2014a). A few studies looking at the relationship between personality traits and CB performance found that the ability to resist framing errors was positively related with emotional stability, agreeableness and conscientiousness (Soane & Chmiel, 2005) and that a composite score on different CB tasks was positively correlated with conscientiousness, openness and honesty/humility (Weller, Ceschi, Hirsch, Sartori, & Costantini, 2018). Finally, performance on CBs tasks seems to be predictive of more positive real-life outcomes (Bruine de Bruin et al., 2007; Toplak et al., 2017).

Previous work on the dimensionality of cognitive bias tasks

There were several empirical attempts so far to establish the structure and dimensionality of CBs. In short, the results of these attempts mainly do not align particularly well with theoretical taxonomies of cognitive biases (e.g. Stanovich, Toplak & West, 2008; Oreg & Bayazit, 2009) and also fail to show a great level of consistency among themselves. In one of the earlier attempts at analyzing the structure of CBs, Weaver and Stewart (2012) concluded that a two-factor solution best describes the relationships among nine different tasks from judgment and decision-making domain. The tasks that are traditionally used in the heuristics and bias research loaded on the first, coherence factor (e.g. probability combination

tasks, conjunction fallacy task, framing task, base-rate task and four-card selection task). The other factor, correspondence factor, accounted for the variance in tasks such as judging the prices of cars and apartments, or the quality of teams based on different features. Teovanović et al. (2015) conducted a factor analysis with oblimin rotation on seven CB tasks (anchoring, belief bias, overconfidence bias, hindsight bias, base-rate neglect, sunk cost and outcome bias) and also found that the two-factor solution was the most appropriate for their data. They labeled the first factor that was mostly defined by the belief bias and outcome bias as the "normative rationality" factor, as higher score on this factor indicated higher rates of predictably irrational responses. The other factor was defined by positive loadings on anchoring and hindsight bias and negative loadings on overconfidence. They called this factor the "ecological rationality" factor as the biases that defined this factor indicated responsiveness to feedback and well calibrated confidence judgments, the characteristics of an ecologically rational agents.

Unlike the previous two studies, Aczel et al. (2015) showed that the four-factor solution was the best in both of their studies for the eight CBs that they investigated (gambler's fallacy, sunk cost, base-rate neglect, Monty Hall problem, insensitivity to sample size, relativity bias, outcome bias, anchoring). However, although the four-factor solution was the most appropriate in both of their studies, these factors were not consistent and the factor structures greatly varied. For example, outcome bias task formed a separate factor in one study, but loaded on the same factor with gambler's fallacy, sunk cost and relativity bias in the other study. However, it has to be noted that in this study each of the biases was measured with only one item.

Recently, additional indication of multidimensionality of CB tasks came from research by Ceschi et al. (2019). This study investigated the greatest number of CB tasks thus far (17). A three-component solution fitted the data best. The first dimension was defined by biases that indicate reliance on both availability heuristic (availability bias, imaginability bias) and representativeness heuristic (base-rate neglect, conjunction fallacy, gambler's fallacy). The second dimension was defined by biases that indicate overvaluation of costs and overestimation of losses (endowment effect, sunk cost) as well as those reflecting an overly optimistic view of the world (optimism bias). Finally, the third dimension was defined by biases that depend on the reference point (anchoring and regression towards the mean). Finally, two most recent studies (Berthet, 2021; Berthet & de Gardelle, 2021) showed that the correlations among eight and six CBs respectively were very low, to the degree that the datasets were not even suited for a factor analysis.

All of these studies indicate that CB tasks are very heterogeneous and share very little common variance among themselves. Here, we must take a small detour and mention somewhat related body of research on the so-called "decision-making competence". Motivated by Stanovich and West (1998, 2000) observation about existence of positive manifold among CB tasks, Parker and Fischoff (2005) factor-analyzed seven different behavioral decision tasks using principal components analysis (consistency in risk perceptions, recognizing social norms, resistance to sunk cost, resistance to framing, applying decision rules, path independence and overconfidence) and showed that, although a three-factor solution fitted data the best, one factor was able to explain 25.1% of the variance in these tasks. They concluded that judgmental biases are not just random errors and that the DMC construct can explain why some people are better and others worse at solving these types of tasks. Bruine de Bruin et al. (2007) tried to replicate and extend these findings. This time they found that the same tasks were best described by two factors, but that one factor was again able to explain a substantial portion of variance among tasks (30.1%). However, it must be noted that only five of the seven DMC tasks can be viewed as classical CB tasks (i.e. applying decision rules and recognizing social norms are not typical CB tasks). Relatively high correlations that these two tasks exhibit with other DMC tasks could be the reason why Parker and Fischoff (2005) and Bruine de Bruin et al. (2007) found enough communality among their tasks, while the authors examining exclusively CB tasks generally do not find these levels of communality.

The difference between rationality and intelligence

There are both empirical and theoretical reasons to treat rationality as a distinct construct from the fluid intelligence. In their book, Stanovich et al. (2016) systematized a large body of their own research on the validity of their rationality measure - Comprehensive Assessment of Rational Thinking (CART). The CART is, basically, a composite measure of large number of different CBs, numeracy and some thinking dispositions such as disposition towards superstitious or conspiracy thinking. The authors showed that the correlation between CART and fluid intelligence scores are moderate. Similar results were obtained when correlating DMC and fluid intelligence. In their recent work, Bruine de Bruin, Parker and Fischoff (2020) described a lot of DMC validation studies and concluded that the correlation between DMC and fluid intelligence seems to be generally moderate and positive. Building on these non-perfect correlations between rationality and fluid intelligence and additional theoretical work, Stanovich et al. (2016) claim that rationality assessed with CB tasks is broader and conceptually distinct from fluid intelligence.

Theoretically, the conceptual difference between fluid intelligence and rationality follows from the tripartite theory of mind (Stanovich, 2009a, 2012). This theory differentiates between autonomous, algorithmic and reflective parts of the mind. According to it, in order to successfully solve the majority of CBs tasks, a person first has to overcome initial incorrect response generated by the autonomous mind. In other words, a person has to reflect on his/her response and recognize the need to engage in more deliberate processing (reflective mind's task) and also possess adequate ability and computational power to calculate or come up with a correct response (algorithmic mind's task). Conversely, success on classical intelligence tests do not depend so much on the reflective, but only on the algorithmic mind, constituting this as the crucial difference between the two. In this framework, the reflective mind refers to different thinking dispositions that are in principle malleable and to a degree teachable (e.g. reflection/impulsivity (R/I), the disposition to be careful at the expense of speed when solving tasks [Baron, 2018]) and that help a person effectively solve a task, while algorithmic mind refers more to cognitive capacities or abilities in the narrower sense that are less prone to change (e.g. fluid intelligence; Stanovich, 2012). In this conceptualization of intelligence (i.e. what the usual intelligence tests measure), intelligence is practically not dependent on dispositions but mostly on capacities. From this, it follows that rationality captured by CB tasks is a broader construct than intelligence, as it is more dependent on thinking dispositions, and therefore it makes sense to conceptually differentiate between the two (e.g. Stanovich, 2009b, 2012).

Baron (1985) also holds that rational thinking is mainly about thinking dispositions. Specifically, it refers to the way people form beliefs based on which they make decisions, what rules they follow and what methods they use in the process. In other words, rationality can be seen as a disposition to adequately search for goals, possibilities, and evidence, trying to find even those that are against our current ones and giving them a fair treatment. However, as opposed to Stanovich who defines intelligence in narrower terms, Baron (1985) holds that thinking dispositions are integral part of intelligence. Hence, rational thinking and behavior is integral part of intelligent thinking/behavior. However, apart from rational thinking that is mostly about "how" people go about forming beliefs (i.e. dispositions), intelligence also includes additional properties such as cognitive capacities and knowledge. From this argument follows that intelligence is a broader concept than the rationality, where rationality represents a part of intelligence that is more malleable and teachable. This is also the main reason why it could be useful to make some sort of distinction between the two – as rationality is all about "how" to think, we can teach

people to be rational (and therefore more intelligent) by teaching them better, or more rational ways of thinking.

Convergent and predictive validity of rationality

Bruine de Bruin et al. (2020) showed that DMC factor predict real-life outcomes after statistically adjusting for individual differences in fluid intelligence. For example, DMC predicted an index of different life outcomes that could have come about due to person's bad decisions, over and above cognitive ability. Weller, Moholy, Bossard and Levin (2015) similarly showed that lower DMC obtained at ages 10-11 predicted greater psycho-social difficulties two years later, even after statistically accounting for the effects of numeracy and inhibitory control. Finally, DMC was shown to be negatively correlated with childhood delinquency and the number of sexual partners after statistically adjusting for cognitive ability (Parker, Bruine de Bruin, Fischoff & Weller, 2018). However, as we previously noted, DMC is a composite of both CB and non-CB tasks, so it would be instructive to see the predictive validity of rationality captured solely by CB tasks. In their study, Toplak et al. (2017) arrived at similar conclusion as DMC researchers: a composite score based on five CB tasks (ratio bias, belief bias in syllogistic reasoning, cognitive reflection, probabilistic and statistical reasoning, and rational temporal discounting) predicted a composite score of real-world outcomes across several different domains (electronic media use, secure computing, substance use, driving behavior, financial behavior and gambling) even when the effects of education and gender were taken into account.

Taken together, it seems that there is a general consensus among researchers that rationality taps into additional constructs besides fluid intelligence. In their review, Bruine de Bruin et al. (2020) cite studies that show that DMC, in addition to fluid intelligence, correlates with motivation to think (i.e. need for cognition), experience (i.e., crystallized intelligence), executive cognitive functioning (i.e., inhibition, monitoring and shifting; Del Missier, Mäntylä, & De Bruin, 2012) and numeracy, and conclude that "decision-making competence may reflect a combination of intellectual, motivational, emotional, and experience-based skills" (p. 188). This is similar to one of two possible interpretations of CBs put forward by Stankov (2017). According to this interpretation, CBs could lie on the cross-section of personality and abilities, being an amalgam of cognitive and non-cognitive processes (the other possibility is that CBs are domain specific and capture very specific processes). In addition to the previously mentioned non-cognitive variables relevant for DMC/rationality, additional thinking disposition of actively open-minded thinking (AOT) seems to be of particular importance. AOT is a disposition to be open to and actively

search for new information and evidence that counteract current beliefs as well as the willingness to revise beliefs if new evidence deems it necessary (Baron, 2019; Baron, Scott, Fincher & Metz, 2015). Again, drawing from the tripartite theory (Stanovich, 2012), this disposition could be related to the reflective mind - the ability and/or disposition to reflect on one's current beliefs/position and correct them. The relationship between the CBs and AOT has also been demonstrated empirically. For example, in Stanovich et al. (2016) work on validation of CART, AOT has consistently come up as one of the strongest correlates of the rationality score with the majority of the CART subtests showing moderate to high correlations with it.

In sum, the aim of our research was to explore the dimensionality of a relatively large number of CB tasks and to investigate the validity of rationality factor(s) by correlating it/them with different variables from its nomological network (i.e. fluid intelligence, numeracy, cognitive reflection, AOT, superstitious and conspiracy thinking and personality traits) as well as several real-life outcomes (DOI, life and career satisfaction). In comparison to previous studies that investigated the validity of rationality measures, ours has several advantages. For example, in comparison to Aczel et al. (2015), we use several tasks per bias and thus have more reliable CB measures. Next, we tested more CB tasks than Teovanović et al. (2015) and Bruine de Bruin (2007) and did it over two different samples (students and community sample). Although Ceschi et al. (2019) measured greater number of tasks, they did not include variables that could be used for validating their decision-making factors which was accomplished in this work. Therefore, our study extends previous in several ways: we investigate a great number of CB tasks, have two different samples and a large set of variables useful for investigating convergent and predictive validity of rationality measure.

Study 1

Methods

Participants

A total of 253 undergraduate University of Zagreb students participated in this study (214 from humanities and social sciences – mostly psychology students, 34 from other disciplines and five undeclared). There were 187 females, 62 males and four participants refused to report their gender. The mean participants' age was 21.47 (SD = 1.89; Min = 18, Max = 29).

Procedure

Participants solved our focal tasks as a part of larger battery of tasks not all of which are reported in this study. Relevant for the current study, the participants solved 10 different CB tasks (each CB was measured by several items), fluid intelligence test, numeracy test, cognitive reflection test, three different questionnaires, actively open-minded thinking, superstitious thinking and conspiracy thinking questionnaire and, finally, the decision-outcome inventory (DOI). Some of the tasks were pseudo-randomized (meaning that there were three different sequences of tasks randomly given to participants), while the others were solved in fixed order. The students filled-in the tests and questionnaires on computers, in groups of 20 to 25 participants under the supervision of the investigators. The whole testing lasted up to two hours and was organized in two parts divided by a 15 minute break. In the first part, participants first solved a group of tasks that were pseudo-randomized and that consisted of cognitive reflection, base-rate neglect, causal base-rate, outcome bias, sunk cost, gambler's fallacy, attribute and risk framing tasks. This was followed by the belief-bias syllogism tasks, numeracy and DOI that were presented in fixed order. After this, there was a 15 minute break followed by a second part of testing. In the second part, all of the tasks were presented in fixed order. Participants first solved the fluid intelligence test, followed by the conspiracy thinking questionnaire, actively open-minded thinking questionnaire, second part of CBs that have to be measured with two related questions in order to determine whether the bias is present (Type-B tasks in Aczel et al. [2015] taxonomy of CB tasks, such as framing effects and outcome bias), four-card selection tasks, availability bias tasks and, finally, the superstitious thinking questionnaire.

Instruments

Here we describe all of the measures used in the Study 1 and for each one provide one item/task as an example. We provide all the items we used in the Appendix A.

a) Cognitive biases tasks

Belief-bias syllogisms. Belief-bias syllogisms tasks pit the believability of a conclusion against its logical validity. An example task goes as follows: "Premise 1: All flowers have petals. Premise 2: Roses have petals. Conclusion: Roses are flowers." (Markovits & Nantel, 1989). This conclusion is logically incorrect because it does not follow from these premises that *only* flowers have petals, so roses might as well be something other than flowers (e.g. children collage art). However, because the conclusion that

roses are flowers conforms with our empirical reality, it is quite believable and many people accept it as valid. Thus, the false intuitive response is the product of believability of the conclusion, while strong conformity with logical principles is needed to come up with the right, logically valid response. In addition to this example, we used three additional syllogisms whose conclusions were believable, but logically incorrect and four more syllogisms whose conclusions were unbelievable, but logically correct, having eight items in total. The responses where participants identified believable and illogical conclusions as logically incorrect or unbelievable and logical conclusions as logically correct were scored as correct. We calculated the total score as a proportion of correct responses.

Four-card selection task. We had four different tasks based on the ones introduced by Wason (1966, 1968) that all had the same structure. A rule was explicitly stated for each of the items and the participants were informed that the rule may or may not be correct. Their task was to check the accuracy of the rule by turning two cards of their choice. For example, one of the items was: “Rule: If a card shows “5” on one side, the word "Excellent" is on the opposite side. Which two cards would you choose to turn to check the accuracy of this rule?”. Participants were then showed four cards that had numbers 5 and 3 and words “Excellent” and “Good” written on the front side. The correct answer here would be to turn the cards containing number 5 and the word “Good” because turning only these would allow one to conclude whether the rule is correct or false. However, often, the participants are lured to turn the card containing the word “Excellent” instead of the card “Good”, although for the rule to be correct it does not matter what is behind the “Excellent” and “3” cards (Nickerson, 1998). Besides the two non-deontic tasks, such as the one described, we also had two deontic tasks whose content was related to a socially relevant, not just arbitrary, rule (e.g. If a person drinks beer, he/she must be over 18 years old). Picking the two accurate cards to turn was scored as 1 while all other combinations were scored as 0. The total score was calculated as the average of responses on four tasks.

Base-rate neglect. The participants were presented with four different problems where the description of a person was contrasted to the base-rate information. These problems were modeled after Kahneman and Tversky (1973) items and the one we used four our study were from De Neys and Glumicic (2008) study. Specifically, there were two possible answers, a stereotypical one (based on the description of a person) and one that was consistent with the base-rate. For example, one of the items was: “Among the 1000 people that participated in the study, there were 50 16-year-olds and 950 50-year-olds. Helen is randomly chosen participant in this research. Helen listens to hip hop and rap music. She likes to wear tight T-shirts

and jeans. She loves to dance and has a small nose piercing. Which is more likely? a) Helen is 16 years old; or b) Helen is 50 years old."

Here, the description of Helen was stereotypical for a teenager. Thus, a person who heavily relies on this information would respond with an "a". However, base-rate information indicated a much greater probability that a randomly chosen participant would be 50 years old. Thus, response "b" was coded as a correct one. However, it has to be noted that technically this does not have to be a correct response and that this depends on the diagnosticity of the information in the task (e.g. the information could be that Helen is a minor which would render a base-rate based response incorrect). Nevertheless, as the stereotypical response is the intuitive one on these tasks and the participants need to engage in correcting this response in order to accompany base rate information into a judgment (Barbey & Sloman, 2007; Pennycook, Fugelsang, & Koehler, 2012), we always coded a response based on base-rates as a correct one. The correct responses were scored as 1 and total score was the average of four responses.

Causal base-rate. Here, participants are provided with two conflicting pieces of information: one is statistical and favors one decision while another is based on personal, case-based experience and favors another decision. These items were based on the classic Volvo vs. Saab items from Fong, Krantz and Nisbett (1986) and were used in previous studies (e.g. Toplak et al., 2011; Stanovich et al., 2016). An example item would be:

"Professor Kellan, the director of a teacher preparation program, was designing a new course in human development and needed to select a textbook for the new course. She had narrowed her decision down to one of two textbooks: one published by Pearson and the other published by McGraw. Professor Kellan belonged to several professional organizations that provided Web-based forums for its members to share information about curricular issues. Each of the forums had a textbook evaluation section, and the websites unanimously rated the McGraw textbook as the better choice in every category rated. Categories evaluated included quality of the writing, among others. Just before Professor Kellan was about to place the order for the McGraw book, however, she asked an experienced colleague for her opinion about the textbooks. Her colleague reported that she preferred the Pearson book. What do you think Professor Kellan should do?

a. She should definitely use the Pearson textbook (1 point)

b. She should probably use the Pearson textbook (2 points)

c. She should probably use the McGraw textbook (3 points)

d. She should definitely use the McGraw textbook" (4 points)

Here preference for the McGraw textbook indicates a tendency to rely on the large-sample information in spite of a salient personal testimony. A preference for the Pearson textbook indicates reliance on the personal testimony over the large-sample information. Therefore, a larger preference for McGraw book was assigned with more points. A total score was calculated as an average of responses to all three items.

Gambler's fallacy. Gambler's fallacy refers to the tendency for people to see links between events in the past and events in the future when the two are really independent (Tversky & Kahneman, 1974). We used four items that were either taken from Stanovich et al. (2016) or obtained through personal communication with Predrag Teovanović. Consider the following problem which is one of the four we used:

"Imagine you are throwing a fair coin (there is a 50% chance of it falling on either side) and it happened to fall on the tail side 5 times in a row. What do you think is more likely to happen in the sixth throw?

a) A head is more likely in the sixth throw

b) A tail is more likely in the sixth throw

c) Both head and tail are equally likely in the sixth throw"

Here the correct answer is "c". However, people prone to gambler's fallacy would reason that, since there were five tails in a row, head would be more probably in the sixth throw. This does not make sense as a coin does not "remember" previous outcomes and always has a 50% probability of falling on each side. We measured gambler's fallacy with four items. Correct responses were scored as 1 and the total score was the average of responses.

Availability bias. Availability heuristic refers to assessing the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind (Tversky & Kahneman, 1973). Availability or the ease of retrieval certain instances of events is often influenced by the vividness or media exposure and does not necessarily correspond to the true frequency of such instances. For example, people might think that homicide is a much more common cause of death than the diabetes (it is the other way round; this was one of the questions we asked our participants) because homicides are often covered in media while diabetes complications and deaths are rarely discussed publicly. In this study, we followed a paradigm introduced by Lichtenstein, Slovic, Fischhoff, Layman and Combs (1978) and asked the participants which of the four pairs of lethal events is more common. Choosing causes of death that are less common but more vivid and more covered in media is a sign of over-reliance on easily available and retrievable information, (Pachur, Hertwig, & Steinmann, 2012;

Stanovich et al., 2016). Thus, we refer to responses that follow from the availability heuristic even in situations when this does not correspond to reality as the availability bias. We scored the correct responses as 1 and incorrect (based on the availability heuristic) as 0. Again, we calculated the total score as the average of responses.

Sunk cost. Sunk cost effect refers to the tendency to "continue an endeavor once an investment of money, effort, or time has been made" (Arkes & Blumer, 1985, p. 124). We measured the susceptibility to sunk cost using four different scenarios describing situations where a person had to choose between an option reflecting unrecoverable past expenditure (sunk-cost) and a more beneficial option in a given situation (normative option). The items were taken from Aczel et al. (2015), Bruine de Bruin et al. (2007) and Teovanović et al. (2015) studies. For example, one of the tasks were "You paid a fair amount of money to rent a club for a birthday party. Three days before the party, you learn that you can move the party completely free of charge to a much better place, a villa with a pool. All additional costs (food, drinks, cleaning) are the same in both locations. It is too late to cancel the rented space, and the money paid cannot be refunded. What would you do?" Participants were instructed to make a choice between two options by using a rating scale that ranged from 1 (most likely to choose sunk-cost option, in this case to have a party in a club) to 6 (most likely to choose the normatively correct option, in this case to throw a party in a villa with a pool). We averaged the score on four individual items to calculate the total score of susceptibility to sunk cost.

Attribute framing. Framing problems assess the degree to which a judgment is affected by irrelevant variations in problem descriptions. Therefore, similar to other studies that measure susceptibility to framing (some of our items were taken from Bruine de Bruin et al., 2007, and some were obtained through personal communication with Predrag Teovanović), for each framing problem a participant reads two scenarios. Scenarios are constructed so that they are normatively identical – the only difference is the way the information is presented (i.e., framed). For example, one scenario our participants saw was this one: "Imagine going to work by public transport every day and in 80% of cases waiting longer than three minutes for the bus to arrive. How satisfied would you be with public transportation services?" Participants were then instructed to evaluate their satisfaction with public transport on a 6-point scale (1 – completely unsatisfied, 6 – completely satisfied). The other, normatively identical scenario to the previous one, was: "Imagine going to work by public transport every day and in 20% of cases waiting less than three minutes for the bus to arrive. How satisfied would you be with public transportation

services?" Again, participants rated their satisfaction on a 6-point scale. In order to minimize memory effects, two scenarios were presented at two different time points, one at the beginning of the first part of survey and the other at the end of the second part. Generally, this means that at least an hour passed between the two scenarios during which the participants solved a number of different tasks that further interfered with their memory. We scored susceptibility to framing by subtracting the evaluation in the "worse" scenario (i.e. waiting longer than 3 minutes) from the "better" scenario (i.e. waiting less than 3 minutes). The majority of framing scores when assessed in this way was positive, although there were also negative scores (people who would judge "worse" scenario as being better than a "good" one). As the higher result meant higher framing bias in the expected direction, before calculating the total score on attribute framing, we transformed individual item scores by subtracting them from 5, theoretically highest score indicating maximal susceptibility to bias. In this way, similarly to our other CB tasks, higher scores indicated lower susceptibility to bias, i.e. higher rationality. We then averaged the score on the four attribute framing items to get the total score.

Risk framing. Risk framing (Tversky & Kahneman, 1981), similar to attribute framing, measures the degree to which a judgment is affected by irrelevant variations in problem descriptions. However, this time both scenarios face participants with a choice between a "sure thing" and a risky option. The difference is again in the wording of scenarios. In a so-called loss-frame, the outcome is framed in terms of losses while in a gain-frame it is framed in terms of gains. We measured risk framing with four pairs of tasks. For example, we faced our participants with the following options (the items were taken from Bruine de Bruin et al., 2007 and Predrag Teovanović [personal communication]): "Imagine that a recent research showed that a certain pesticide could kill 1,200 endangered animals. Two response options to this pesticide threat have been proposed: a) If you go with option A, 600 animals would be saved for sure. b) If you go with option B, there is a 75% chance that 800 animals would be saved and a 25% chance that no animals would be saved. Which option would you recommend?" The response scale again had six point (1 – certainly option A, 6 - certainly option B). This was a gain-frame as the animals would be saved, as opposed to the following loss-frame where the animals would die: "Imagine that a recent research showed that a certain pesticide could kill 1,200 endangered animals. Two response options to this pesticide threat have been proposed: a) If you go with option A, 600 animals would die for sure. b) If you go with option B, there is a 75% chance that 400 animals would die and a 25% chance that all 1200 animals would die. Which option would you recommend?" We subtracted a gain-frame response from a loss-frame response. Therefore, a higher score indicated a greater susceptibility to framing effects,

so before calculating the total score, we did the same transformation as with the attribute framing (subtraction from five) in order for higher scores to indicate higher rationality.

Outcome bias. Our final CB task was the outcome bias task (Baron & Hershey, 1988) that was, similarly to framing problems, composed of two parallel and relatively equivalent scenarios, one with a positive and the other with a negative outcome (we obtained our items from Aczel et al., 2015, Baron & Hershey, 1988 and Teovanović, 2013). An example of a positive outcome scenario we showed our participants would be the following: "A 54-year-old had heart problems. He had to stop working because of chest pain but he loved his job and wanted to continue working. Pain also affected other things, such as travel and recreation. Successful heart bypass surgery would ease his pain and increase his life expectancy from 65 to 70 years. However, 8% of people who decide to have this operation die because of the operation itself. His doctor decided to go on with a surgery and the surgery was successful. Please rate the doctor's decision to have surgery performed." Participants again rated the quality of the decision on a six-point scale (1 – Very bad decision, 6 – Very good decision). In another scenario, the outcome of the decision was the opposite, in this case a negative one: "A 58-year-old had degenerative hip disease. He was confined to a wheelchair and forced to retire early last year. Due to immobility, he gained weight and was depressed because he could not work or engage in any recreational activities. He loved his job and recreation and did not want to stop with it. He consulted a doctor who told him that successful degenerative hip surgery would ease his pain and prolong his life expectancy by 10 years or more because he could exercise. However, because the surgery is complicated and the man had a milder heart disease, there is a 2% chance of dying from the surgery itself. Unfortunately, there were complications on the operating table and the man died of heart failure. Please rate the doctor's decision to have surgery performed." We used four pairs of tasks to measure outcome bias. We scored the tasks by subtracting the negative outcome rating from the positive outcome rating with the greater score indicating greater susceptibility to outcome bias. Therefore, we again subtracted the scores from five before calculating the total score as the average score on four outcome bias item pairs.

b) Fluid intelligence. We measured fluid intelligence with a 16 items version of the International Cognitive Ability Resource (ICAR; for details see icar-project.com and Condon and Revelle, 2014). ICAR is a broad cognitive ability assessment tool consisting of four different types of tasks: letters and numbers series, matrix reasoning items, verbal reasoning items and three-dimensional rotation items.

The validation of this measure is reported in Condon and Revelle (2014). We formed the total score as an average of responses to these 16 items.

c) Cognitive reflection test. Cognitive reflection test is test that originally (Frederick, 2005) consisted of three items, each with a distinctive characteristic of pitting an intuitive but incorrect responses against a correct one. Probably the best-known item is a bat-and-ball item: "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" The immediate response that comes to mind is 10 cents which is, on a further reflection, incorrect and a correct response I 5 cents. Following the publication of an original three-item test, several studies were published that extended this short form test with additional items (e.g. Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; Thomson & Oppenheimer, 2016; Toplak, West & Stanovich, 2014b). In this study, we used six items, each cuing strong but incorrect intuitive response. The proportion of correct responses represented a total score.

d) Numeracy. We used the Berlin numeracy test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) as a measure of numeracy. The BNT is a four-question test for assessing numeracy and risk literacy. An example of a question is "Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?". The questions are designed in a way that they gradually become harder and a total score is calculated by averaging the responses on the four questions.

e) Actively open-minded thinking. We used a 15-item AOT scale. The items were taken from Stanovich and West (2007) 41-item scale and these specific 15 items were first used in Campitelli and Gerrans (2014). AOT scale is a scale that assesses individual beliefs about proper standards of thinking by asking participants to indicate their level of agreement (1 – strongly disagree to 6 – strongly agree) with items such as "It is OK to ignore evidence against your established beliefs". This particular item would be reverse coded as it goes against the actively open-minded principles of thinking. The total score on this scale was calculated as a mean level of agreement with the items (after recoding the responses so that the higher score indicates higher AOT) and can be anything between 1 and 6.

f) Superstitious thinking. Superstitious thinking was assessed using items from Toplak et al. (2011). The scale consisted of 13 items in total, and the participants rated their level of agreement with each of the items on a six-point scale. The total score was calculated as a mean score of responses on all of the items.

The items tapped into four different concepts: superstitious beliefs (e.g. “When something good happens to me, I believe it is likely to be balanced by something bad.”), luck (e.g. “I have personal possessions that bring me luck at times.”), paranormal beliefs (“Astrology can be useful in making personality judgments.”), and extrasensory perception (“Dreams can provide information about the future.”)

g) Conspiracy beliefs. We measured conspiracy beliefs with 12 items taken from the Generic conspiracist beliefs scale developed by Brotherton, French, and Pickering (2013). For example, one of the items in the scale was “A small, secret group of people is responsible for making all major world decisions, such as going to war.” Participants rated their level of agreement with the claims such as this one on a six-point scale and the total score was the average of these 12 ratings.

h) Decision-outcome inventory. The Decision-Outcome Inventory was developed by Bruine de Bruin et al. (2007) as a measure of decision-making success in terms of avoiding negative decision outcomes. It presents a list of negative outcomes, ranging from mild (e.g. throwing out food or groceries you had bought or destroying clothes because you did not follow the washing instructions on the label) to more serious ones (e.g. participation in a fight or experience of breaking a bone because you fell, slipped, or misstepped). We used 33 different outcomes and our outcomes somewhat differed from the original inventory. Namely, as our Study 1 participants were students, some of the original outcomes were extremely unlikely or totally impossible to have happened to them and were thus removed. For example, we removed some of the most serious negative outcomes for this reason (e.g. been in a jail cell overnight for any reason or got divorced). We also added some outcomes that seems appropriate for college students such as “Overslept classes multiple times” or “Had to do an additional assignment because you missed too many lectures.” The total score was calculated in a following way: first, for some outcomes the participants were asked whether they had the opportunity to experience them (for example, someone who does not have a driving license could not have had it seized from him/her – therefore, for these kinds of outcomes, we first asked the participants whether they could or could not experience it). Next, to account for the severity of outcomes, possible outcomes were weighted by the proportion of participants who reported not experiencing them (thus, more severe outcomes or the ones that were experienced by less people were weighted more). Finally, a total score was calculated by averaging these weighted scores.

Results

In order to answer our research questions, we did two things. First, we factor-analyzed our 10 CB measures to investigate the dimensionality of our tasks. Second, we correlated our measures in order to

validate our rationality factor(s). Prior to reporting the results of our main analyses, in Table 12 we report the descriptive statistics of raw scores and the reliabilities of all the measures we used in the study. In addition, we also report the effect sizes of biased responding for each of the CB tasks, i.e. the size of the difference between the mean of the responses and the normative value for each of the tasks.

Table 12. Descriptive statistics and reliabilities of the Study 1 measures with the effect sizes of biased responding in comparison with normative responding on cognitive bias tasks

Measure	**M**	**SD**	**Min**	**Max**	**Cronbach α**	**ω_h**	**Normative value**	**Cohen's d**
Belief-bias	0.74	0.21	0	1	.79	.77	1	1.20
Base-rate neglect	0.58	0.39	0	1	.93	.91	1	1.08
Causal base-rate	2.96	0.51	1.33	4	.55	.56	4	2.05
Four-card selection	0.34	0.33	0	1	.86	.80	1	2.00
Attribute framing	4.73	0.65	2.25	7.50	.35	.39	5	0.42
Outcome bias	4.34	0.94	0.25	6.75	.65	.59	5	0.70
Sunk cost	4.41	0.95	2	6	.56	.51	6	1.67
Availability bias	0.68	0.30	0	1	.77	.64	1	1.08
Gambler's fallacy	0.84	0.21	0	1	.76	.69	1	0.75
Risk framing	4.86	0.77	2.25	7.50	.24	.14	5	0.19
ICAR	0.67	0.18	0.1	1	.83	.61	-	-
Numeracy	0.42	0.27	0	1	.64	.61	-	-
Cognitive reflection	0.65	0.30	0	1	.83	.72	-	-
AOT	4.62	0.65	1.87	6	.84	.61	-	-
Conspiracy thinking	3.22	0.79	1	5.42	.87	.73	-	-
Superstitious thinking	2.01	0.63	1	4.46	.83	.56	-	-
DOI	16.72	7.67	1.29	42.33	.66	.53	-	-

Note. ICAR = International cognitive ability resource; AOT = Actively open-minded thinking; DOI = Decision-Outcome Inventory.

Cohen's ds for cognitive bias tasks were calculated by the formula $d = (M - \mu) / SD$, where M is the mean CB score and μ is normative value.

Table 12 shows that the participants found CB tasks to be of varying difficulties. Specifically, the highest scores, i.e. the highest number of correct/normative responses, were obtained on the gambler's fallacy

tasks and belief-bias tasks, indicating that our participants were the least susceptible to these biases. Conversely, the four-cards selection task was by far the most difficult revealing strong susceptibility of our participants to confirmation bias as assessed by this task. Overall, and in line with college students being on average relatively smart and rational, our participants on average correctly solved two thirds of fluid intelligence tasks, more than half of the cognitive reflection tasks and just under half numeracy tasks (this numeracy test is generally considered to be very hard; for example, college students scored 40% correct in Cokely et al., 2012, study), at the same time mostly disagreeing with the items measuring conspiracy and superstitious beliefs.

The correlations among our focal variables are presented in the Table 13. The raw correlations are presented above the diagonal, while the correlations corrected for attenuation are presented below the diagonal. It can be seen that the correlations among our CB tasks are generally positive, albeit low, replicating previous findings (Stanovich & West, 1998, 2000; Parker & Fischoff, 2005; Teovanović et al., 2015). Most of our CB tasks are significantly correlated with three to six other tasks. However, there are three tasks that are more problematic in this way. Specifically, sunk cost is correlated only with two other tasks, risk framing with one, while gambler's fallacy failed to correlate with any of the other CB tasks. Perhaps the main problem with gambler's fallacy was ceiling effect in its scores, as majority of students solved these tasks correctly, while the risk framing tasks had serious issues with reliability that could have substantially diminished its correlations with other tasks. Interestingly, only four of our CB tasks (belief-bias syllogisms, base-rate neglect, four-card selection task and sunk cost) correlated significantly with fluid intelligence, but these correlations were all relatively small. This suggests that the performance on CB tasks was mostly independent of fluid intelligence. However, it was not as independent from numeracy, as both numeracy and cognitive reflection test that mostly captures numeracy construct (Attali & Bar Hillel, 2020), each correlated with six of the CB tasks. Moreover, the thinking disposition of actively open-minded thinking was significantly positively correlated with all but one of the CB tasks, indicating that this thinking disposition plays an important role in success on rationality tasks. It is interesting to note that AOT was more important for CB tasks than for success on the fluid intelligence measure, judging from a very low correlation between the two. Finally, it seems that our CB tasks were largely irrelevant for conspiracy thinking and real-life decision outcomes (DOI), but somewhat relevant for superstitious thinking, with four of them being negatively related with this measure.

Table 13. Correlations among the Study 1 variables. Raw correlations are above the diagonal while the disattenuated[7] correlations are below the diagonal.

	RAT	BBS	BRN	CBR	FCS	ATF	OB	SC	AVB	GF	RIF	ICAR	NU	CRT	AOT	CON	SUP	DOI
RAT		.62*	.66*	.62*	.47*	.31**	.44**	.25**	.47**	-.04	.15*	.30**	.38**	.44**	.40**	-.10	-.23**	-.09
BBS	/		.27**	.19**	.23**	.12	.20**	.14*	.11	-.06	.01	.26**	.37**	.41**	.24**	-.05	-.15*	-.03
BRN	/	.32		.27**	.18**	.14*	.17**	.01	.19**	.08	-.01	.19**	.24**	.28**	.24**	-.11	-.15*	-.10
CBR	/	.29	.38		.12	.13*	.15*	.14*	.23**	-.09	.10	.12	.19**	.24**	.19**	-.09	-.19**	-.04
FCS	/	.28	.20	.17		.04	.14*	.05	.13*	-.11	.13*	.19**	.12	.20**	.21**	-.07	.05	-.10
ATF	/	.23	.25	.30	.07		.05	.06	.08	.02	.12	.09	.13*	.10	.13*	.03	-.07	.03
OB	/	.28	.22	.25	.19	.11		.06	.12	.09	.00	.09	.13*	.15*	.22**	.01	-.04	-.01
SC	/	.21	.01	.25	.07	.14	.10		.10	.06	.05	.15*	.11	.21**	.15*	-.01	-.07	.08
AVB	/	.14	.22	.35	.16	.15	.17	.15		-.04	.04	.09	.14*	.12	.22*	-.03	-.23**	-.06
GF	/	-.08	.10	-.14	-.14	.04	.13	.09	-.05		.02	.00	-.07	-.04	.11	-.02	-.09	.05
RIF	/	.02	-.02	.27	.28	.41	.00	.14	.09	.05		.01	-.03	-.05	.13*	-.04	-.01	-.01
ICAR	.46	.32	.22	.18	.22	.17	.12	.22	.11	.00	.02		.40**	.44**	.16*	.01	-.04	.03
NU	.67	.52	.31	.32	.16	.28	.20	.17	.20	-.10	-.05	.55		.46**	.18**	.01	-.12	.04
CRT	.67	.51	.32	.36	.24	.19	.20	.31	.15	-.05	-.11	.53	.63		.17**	-.03	-.09	.01
AOT	.61	.29	.27	.28	.25	.22	.30	.22	.27	.14	.29	.19	.25	.20		-.26**	-.22**	-.10
CON	-.15	-.06	-.12	-.13	-.08	.05	.01	-.01	-.04	-.02	-.09	.01	.01	-.04	-.30		.36**	.23**
SUP	-.35	-.19	-.17	-.28	.05	-.13	-.05	-.10	-.29	-.11	-.02	-.05	-.17	-.11	-.26	.43		.11
DOI	-.15	-.04	-.13	-.07	-.13	.06	-.02	.13	-.08	.07	-.02	.04	.06	.01	-.13	.30	.13	

Note. * $p < .05$; ** $p < .01$

RAT = Rationality factor; BBS = Belief bias syllogisms; BRN = Base-rate neglect; CBR = Causal base-rate; FCS = Four-card selection task; AVB = Availability bias; ATF = Attribute framing; OB = Outcome bias; SC = Sunk cost; GF = Gambler's fallacy; RIF = Risk framing; ICAR = Fluid intelligence; NU = Numeracy; CRT = Cognitive reflection test; AOT = Actively open-minded thinking; CON = Conspiracy thinking; SUP = Superstitious thinking; DOI = Decision outcome inventory.

[7] We disattenuated correlations based on the coefficient alpha calculated using the omega () function from psych R package (Revelle, 2021). This function allows for estimation of reliability based on polychoric correlations among the items, leading to less overcorrection.

Before conducting a factor analysis on our CB tasks, we checked whether our data was adequate for performing such analysis. Therefore, we calculated the Kaiser-Meyer-Olkin factor (KMO) and Bartlett's test of sphericity. KMO reached an acceptable level of KMO = .66 and Bartlett's test indicated that the correlation matrix is significantly different from an identity matrix ($\chi2(45) = 143.01$, $p < .001$). To identify the most appropriate structure for our data, we conducted a parallel analysis using a fa.parallel function from the R - package "psych" (Revelle, 2021) and inspected the scree plot. Both the parallel analysis and the scree plot indicated a one-factor solution as the most appropriate one (the outputs from this analysis are presented in Appendix B). We then conducted a factor analysis using a maximum likelihood extraction method, extracting one factor with good fit indices ($\chi2(35) = 35.68$, $p = .44$) that was able to explain 12% of the variance among CB tasks. This is substantially lower amount of explained variance than in the case of the decision-making competence factors obtained in earlier studies (e.g. one factor explained 25% of the variance in Parker & Fischoff [2005] and 30% of the variance in Bruine de Bruin et al. [2007][8]). We called this factor a rationality factor. This factor was most saturated with base-rate neglect, causal base-rate neglect and belief bias scores and least saturated with sunk cost, risk framing and gambler's fallacy scores.

Table 14. One factor model of CB tasks

Cognitive bias task	Loadings
Base-rate neglect	.51
Causal base-rate	.48
Belief bias	.48
Four-card selection	.36
Availability bias	.36
Outcome bias	.34
Attribute framing	.24
Sunk cost	.19
Risk framing	.12
Gambler's fallacy	-.03

[8] However, it seems that these authors conducted a principal components analysis that inflates the communalities and the proportion of variance explained in comparison to factor analysis.

Although the general factor of CB tasks was quite weak, we still wanted to check if it has a meaningful pattern of relationships with the other variables from our study. In Table 13 we show all the correlations among our variables. We calculated rationality factor score for each of the participants (i.e., the score reflecting an individual's standing on the factor) using the regression method. Looking at the Table 13, several things are notable. First, the fact that Rationality factor exhibits relatively strong correlations with numeracy and cognitive reflection and somewhat lower with fluid intelligence confirms that rationality is closely tied to these well-known abilities. As these correlations, even the disattenuated ones, are not perfect, this means that rationality factor captured somewhat different constructs than these other abilities. Second insight from the correlation table is thus related to one of these potential additional constructs. Namely, the rationality factor was strongly related to actively open-minded thinking, and the magnitude of this correlation is greater than any of the correlations between actively open-minded thinking and other cognitive abilities.

In an additional analysis, we wanted to see how much of the variance in rationality factor is explained by numeracy (assessed with CRT), fluid intelligence (ICAR) and actively open-minded thinking, the strongest correlates of rationality factor. Therefore, we regressed the rationality factor on CRT, ICAR and actively open-minded thinking factors using SEM. Specifically, we defined our four latent variables by their corresponding manifest variables (rationality latent variable was defined as a second order factor of ten cognitive bias factors that were each defined by their corresponding manifest variables, while CRT, ICAR and AOT latent variables were defined as first order factors by their respective manifest variables, i.e. six CRT items, 16 ICAR items and 15 AOT items) and did calculations using a "sem" function with maximum likelihood estimation method from the "lavaan" R-package (Rosseel, 2012). Only numeracy and actively open-minded thinking were significant predictors of rationality, explaining 61% of its variance, with unique effect of ICAR being virtually non-existent (description of model fits and beta ponders are shown in Appendix B). This points to the conclusion that the constructs assessed by rationality measures are largely cognitive capacities and dispositions captured by the cognitive reflection test and actively open-minded thinking. To quantify the relative importance of numeracy and actively open-minded thinking for rationality obtained in the regression analysis, we conducted a dominance analysis (Azen & Budescu 2003; Budescu, 1993) by calculating the general dominance coefficient for each of the predictors. In short, this is done by averaging the added variance explained by the predictor in all possible subsets of regression models. In our case, there are only two possible models for our two predictors – one where the predictor is alone in the model and the other where both of the predictors are

included. The dominance coefficient showed that, of the 61% variance explained in the rationality factor, numeracy was responsible for 47% and actively open-minded thinking for 14% of the variance. Finally, neither rationality factor nor individual CB factors were correlated with the decision-outcome inventory score as a potential real-life indicator of bad decisions. This means that either rationality, as measured with these CB tasks, is not important for the chosen real-life decisions or that this version of the decision-outcome inventory is not a particularly good measure of real-life decision quality. However, the rationality factor negatively correlated with superstitious thinking. Given that the actively open-minded thinking correlates with superstitious thinking but neither of the cognitive abilities does, rationality might be related to these epistemically suspect beliefs because of its correlation with actively open-minded thinking. In sum, Study 1 findings show existence of a weak rationality factor that has meaningful relationship with other individual differences in cognitive variables and that mostly reflects numeracy and AOT.

Study 2

The goal of Study 2 was to replicate and expand on findings from the Study 1. In comparison with Study 1, Study 2 had several benefits. Probably the most important one relates to the sample – participants in our Study 2 were not college students but a community sample consisting of participants of different age and education. Second, we expanded the variety of variables for testing convergent and predictive validity of rationality factor(s). Specifically, we measured personality traits, as well as job and career satisfaction as potential additional real-life outcomes of good decisions. Third, we wanted to see whether greater rationality would be reflected in person's quality of decision-making as evaluated by his/her peers. Fourth, to additionally test the role of actively open-minded thinking in rationality, we developed and tested a new, more direct measure of actively open-minded thinking. Specifically, we developed a short construct-driven situational judgment test (SJT; Guenole, Chernyshenko, & Weekly, 2017; Lievens, 2017) of actively open-minded thinking. Finally, to further minimize the potential effects of memory on results in cases where the bias is measured by two parallel items (such as attribute framing and outcome bias), participants solved two parts of the questionnaire one week apart. However, Study 2 also had several weaknesses. One weakness of Study 2 is that, due to time constraints, we captured seven instead of ten CBs and we did not measure conspiracy and superstitious thinking. Furthermore, our research set did not include measures fluid intelligence or numeracy, but only the cognitive reflection test as the only cognitive ability measure. The reason for this was that these three measures were very highly correlated in our first study, as well as the findings from some recent studies (e.g. Attali & Bar Hillel, 2020; Erceg,

Galić & Ružojčić, 2020) showing that cognitive reflection test actually represents a fairly good numeracy test (and little beyond that).

In sum, in Study 2 we measured the following variables: seven CBs (belief-bias, attribute framing, outcome bias, causal base-rate, base-rate neglect, sunk cost and four-card selection task), cognitive reflection, actively open-minded thinking scale and SJT , the "Big Five" personality traits (extraversion, emotional stability, agreeableness, openness and conscientiousness) and four potential real-life decision-making outcomes (decision-outcome inventory, job satisfaction, career satisfaction and peer-rated decision-making quality).

Methods

Participants

In total, 210 participants participated in our study. As our study was divided in two parts separated by a week time, there was some loss of participants. Specifically, 183 participants also participated in and completed second part of the study. Therefore, our sample size is somewhat different depending on the analysis. There were 79 male and 103 female participants (28 unknown). The average participants' age was M = 34.31 (SD = 10.63, Min = 19, Max = 63). Regarding the education, we had one participant with primary school only, 38 with high school, 41 with college education, 87 with higher education and 16 with PhDs.

Procedure

Similarly as Study 1, Study 2 was conducted in two parts, only this time with a longer time distance between them. In the first part of the study, the participants solved the cognitive reflection test, seven CB tasks and actively open-minded thinking self-report questionnaire and several other measures not reported in this study. In the second part, they completed actively open-minded thinking SJT, personality questionnaire, decision-outcome inventory, job and career satisfaction scale and the second part of the two-part CB tasks (attribute framing and outcome bias). Apart from solving these two parts of our study, we asked our participants to forward an additional link containing several other-report measures to their peers who were asked to rate them on these measures. Of relevance for this study is a peer-rated overall decision-making quality. In total, 192 of our participants received peer ratings. They were mostly rated by two peers (N = 144), some by only one peer (N = 39) and several were rated by three peers (N = 9).

Instruments

In this part, we will only describe those instruments that were either not used or were somewhat changed from Study 1. If we used an instrument that is not listed here, it is the same as was used in Study 1. All of the items from described instruments are presented in the Appendix A.

Belief-bias syllogisms. Unlike the first study, in the second study we had only four BBS items and they were all of the same type, namely with unbelievable, but logically correct conclusions. We decided to cut on some of the biases in this study, including some of the BBS task, as we believed that too much of the tasks would have discouraged participants from participating in our study or from finishing it completely, leaving us with much smaller sample size in the end.

Actively open-minded thinking SJT. We developed a three-item construct-driven SJT (Lievens, 2017) for measuring actively open-minded thinking in realistic work situations. Specifically, each task consisted of a description of a realistic situation that could be encountered in real life or work and four potential responses to this scenario. Participant's task was to choose, out of these four options, what he/she would do in that situation. The response options were developed in a way that they increasingly reflect actively open-minded thinking, starting from a first option that reflected almost complete lack of this type of thinking and was scored as 1 (i.e. deciding now without looking for additional information or competing arguments) and ending with a final option that reflected this type of thinking in a high degree and was scored as 4 (i.e. not only taking more time or looking for additional information before deciding, but specifically looking for those information and arguments that counteracts person's current opinion). Here is a sample item:

"You were recently promoted to the position of HR Manager of a large company. Management expects you to make some changes to motivate employees. An older colleague, a long-term employee of the human resources department, believes based on his practice and experience that rewarding employees according to performance is the best way to motivate them. This sounds like a good idea to you too - it makes sense to you that people will work harder if they are paid according to the work they did and you don't see any objective disadvantage of this method. What will you do?

a) You can't see any major drawbacks to this approach, so you'll be introducing a performance-based reward system as soon as possible. This will increase employee motivation and at the same time meet the requirements of management.

b) You will talk to an older colleague who advocates this system and knows more about it than you do. If his reasons for introducing this approach are reasonable and good, you will implement it immediately.
c) You will engage and try to find on the internet what other experts think about why such a system should be introduced.
d) Although it seems that this approach is generally supported, you will do your best to find key arguments against it or identify possible problems with the introduction of this human resource management practice."
By looking at previous example, it is possible to identify the key features of response options. In response a), an actor relies exclusively on his/her current arguments and knowledge for making a decision and makes immediate decision. In response b), an actor seeks out for additional information, but only for the information that is confirming his/her current views. In option c), an actor seeks out a wider range of available information, although still not specifically looking for one that will disconfirm his/her position. Finally, in response d), an actor actively tries to find key counterarguments to his/her own views in order to balance the direction of arguments before deciding and potentially coming up with more objective view of the situation. The total score on this measure was the average of the scores on individual items and could thus be between one and four.

Actively open-minded thinking scale. Actively open-minded thinking scale was somewhat different that the one in Study 1. Specifically, we removed some of the items and replaced them with the items from the currently recommended measure of actively open-minded thinking (http://www.sjdm.org/dmidi/Actively_Open-Minded_Thinking_Beliefs.html). For example, items "No one can discourage me from something I know is right" and "In general, I know everything I need to know about the important things in life" were omitted and items such as "People should take into consideration evidence that goes against conclusions they favor" and "It is important to be loyal to your beliefs even when evidence is brought to bear against them" were included in this version of questionnaire. In total, questionnaire had 13 items that were scored in the same way as the ones in the Study 1.

Mini International Personality Item Pool questionnaire (Mini IPIP). Mini IPIP (Donnellan, Oswald, Baird, & Lucas, 2006) is a 20-item personality measure, measuring the Big 5 traits each with four items. Participants were instructed to rate the accuracy of the description (e.g. "Am the life of the party.", "Sympathize with others' feelings.", "Get chores done right away.", "Have frequent mood swings.",

"Have a vivid imagination.") on a five-point scale (1 = Completely incorrect 5 = Completely correct) and the total score is the average of the ratings on the four items per trait, after the items were recoded so that a higher score indicates a higher degree of a trait.

Decision-outcome inventory. In comparison with Study 1, Decision-outcome inventory was significantly shorter, consisting of 18 items. We dropped number of outcomes that were appropriate for student population, but inappropriate for adults. We also added several finance-related items such as "Took out an unfavorable short-term loan." and "Spent more money in a month than you could afford." The items were again weighted by the percentage of participants who did not experience given outcome, thus giving more weight to more serious negative outcomes. The total score is average of item scores.

Job satisfaction. We measured job satisfaction with one item: "Think about your current job. Weigh all its advantages and disadvantages and then assess how satisfied you are, on the whole, with your job." Participants were instructed to rate their satisfaction on a five-point scale (1 = Very dissatisfied, 5 = Very satisfied).

Career satisfaction. We measured career satisfaction with a five item Career satisfaction scale (Greenhouse, 1990). Participants were instructed to indicate to what extent they agreed or disagreed (1 = Completely disagree, 5 = Completely agree) with statements such as "I am satisfied with the success I have achieved in my career." or "I am satisfied with the progress I have made towards meeting my overall career goals." (the rest of the statements are in the Appendix). The total score is calculated as a mean of the scores on these five statements and higher score indicates higher career satisfaction.

Peer-rated general decision-making quality. This scale was developed by Wood (2012) and consists of four items (e.g. "The decisions my friend makes are quality ones", "The decisions my friend makes end up working out well"). Participants' peers were instructed to rate their agreement with the statements describing their colleague/friend participants on a five-point scale (1 = Completely disagree, 5 = Completely agree). Each participant was rated by up to three different peers. Therefore, before calculating the final scores we averaged peers' ratings on these four items. The final score is calculated as an average of ratings and its higher values indicate peers' more positive perception of decision-making abilities of their colleague.

Results

In Study 2, we used the same analyses as in the Study 1. This means that we factor-analyzed our CB tasks in order to investigate their dimensionality and then correlated the factor score(s) with different measures in order to validate them. Before showing results from these analyses, we present descriptive statistics and reliabilities of our measures in Table 15.

Table 15. Descriptive statistics and reliabilities of the Study 2 measures with the effect sizes of biased responding in comparison with normative responding on cognitive bias tasks

Measure	M	SD	Min	Max	Cronbach α	ω_h	Normative value	Cohen's d
Belief-bias	0.43	0.35	0	1	.82	.64	1	1.64
Base-rate neglect	0.37	0.40	0	1	.95	.93	1	1.57
Causal base-rate	2.75	0.62	1	4	.48	.49	4	2.00
Four-card selection	0.16	0.24	0	1	.80	.77	1	3.56
Attribute framing	4.67	0.71	2.25	6.75	.17	.23	5	0.47
Outcome bias	3.78	1.11	1.25	7.25	.68	.66	5	1.09
Sunk cost	4.31	0.97	1.25	6	.39	.27	6	1.74
Cognitive reflection	0.51	0.34	0	1	.88	.82	-	-
AOT	4.58	0.88	1	6	.85	.75	-	-
AOT SJT	2.87	0.65	1	4	.23	.48	-	-
Extraversion	3.50	0.76	1.25	5	.76	.73	-	-
Agreeableness	3.92	0.64	2	5	.71	.56	-	-
Conscientiousness	3.73	0.73	1.75	5	.74	.70	-	-
Emotional stability	3.48	0.71	1.5	5	.74	.73	-	-
Openness	3.77	0.74	1.25	5	.75	.64	-	-
DOI	19.02	9.49	0.76	47.50	.62	.24	-	-
Job satisfaction	3.94	0.90	1	5	/	/	-	-
Career satisfaction	3.57	0.80	1.40	5	.82	.79	-	-
PRDMQ	4.34	0.49	3	5	.79	.79	-	-

Note. AOT = Actively open-minded thinking; AOT SJT = Actively open-minded thinking situational judgment test; DOI = Decision-Outcome Inventory; CWB = Counterproductive work behavior; PRDMQ – Peer-rated decision-making quality. Cohen's ds for cognitive bias tasks were calculated by the formula $d = (M - \mu) / SD$, where M is the mean CB score and μ is normative value.

Not unexpectedly, our community sample scored somewhat lower on basically every comparable task than college students. Looking at the correlations among the CB tasks in Table 16, it can be seen that they are again mostly low or moderate and positive. The exception is attribute framing that was uncorrelated with any of the other tasks. Possible reason for this is that this measure exhibited very low reliability, lower than in the Study 1. Again, numerical abilities and actively open-minded thinking proved to be relevant for the performance on our CB tasks. Specifically, cognitive reflection test score, as a measure of numeracy, was significantly correlated with each of the seven CB tasks, while actively open-minded thinking was related to five out of seven CB tasks, or six out of seven when measured with a situational judgment test. Conversely, personality traits seemed to matter less, with emotional stability (significantly correlated with five of the CB tasks) and openness (significantly correlated with four of the tasks) being the most important of the personality traits. Finally, looking at the correlations between CB tasks and potential outcomes of decisions (DOI, job and career satisfaction and peer-rated decision-making quality), there is not much to see. Only few of the correlations were significant, and even those that were significant were relatively small (all of the correlations were lower than $r = .20$, disattenuated lower than $r = .30$).

Table 16. Correlations among the Study 2 variables. Raw correlations are above the diagonal while the disattenuated correlations are below the diagonal.

	RAT	BBS	BRN	CBR	FCS	ATF	OB	SC	CRT	AOT	SJT	EXT	AGR	CON	STA	OPE	DOI	JS	CS	DMQ
RAT	/	.63**	.75**	.73**	.52**	.16*	.53**	.45**	.60**	.43**	.31**	.11	-.04	-.17*	.29**	.30**	-.05	.18*	.19**	.11
BBS	/		.35**	.30**	.25**	.13	.28**	.10	.41**	.13	.16*	.04	-.11	-.11	.13	.15*	.06	.02	.18*	.12
BRN	/	.40		.37**	.29**	.05	.27**	.21**	.39**	.40**	.28**	.07	.01	-.05	.22**	.28**	-.03	.16*	.15	.15*
CBR	/	.48	.55		.26**	.05	.25**	.32**	.38**	.35**	.24**	.06	-.06	-.12	.16*	.13	-.07	.12	.16*	.09
FCS	/	.31	.33	.42		-.03	.16*	.16*	.26**	.26**	.21**	.09	-.05	-.16*	.13	.14	-.02	.05	.03	.06
ATF	/	.34	.12	.17	-.08		.13	.11	.21**	-.02	.16*	.12	.09	-.11	.15*	.05	-.03	-.01	.10	.13
OB	/	.37	.34	.44	.22	.38		.15*	.34*	.16*	.06	-.03	.02	-.23**	.25**	.28**	-.06	.19*	.09	.04
SC	/	.18	.34	.74	.29	.42	.29		.34**	.31**	.15*	.15*	.04	-.08	.24**	.17*	-.11	.15	.09	.01
CRT	.81	.48	.43	.59	.31	.54	.44	.56		.42**	.32**	.06	-.06	-.25**	.31**	.26**	-.13	.14	.15*	.06
AOT	.59	.16	.44	.55	.32	-.05	.22	.54	.49		.25**	-.01	.09	-.16*	.19**	.28**	-.07	.02	.04	.19**
SJT	.81	.37	.60	.72	.49	.80	.15	.50	.71	.56		.01	.20**	-.16*	.08	.19**	.06	-.05	.07	.17*
EXT	.16	.05	.08	.10	.12	.33	-.04	.28	.06	-.01	.02		.17*	-.09	.16*	.15*	.10	.09	.22**	-.05
AGR	-.06	-.14	.01	-.10	-.07	.26	.03	.08	-.08	.12	.49	.23		-.06	.05	.16*	.02	.15*	.21**	.16*
CON	-.25	-.14	-.05	-.20	-.21	-.30	-.32	-.15	-.31	-.20	-.39	-.12	-.08		.13	-.17*	-.30**	.01	.12	-.01
STA	.43	.17	.26	.27	.17	.42	.36	.45	.37	.24	.19	.21	.07	.18		.06	-.23**	.27**	.22**	.09
OPE	.44	.19	.33	.22	.18	.14	.39	.32	.32	.35.	.46	.20	.22	-.23	.08		.11	.16*	.06	.11
DOI	-.08	.08	-.04	-.13	-.03	-.09	-.09	-.22	-.16	-.10	.16	.15	.03	-.44	-.34	.16		-.10	-.21**	-.05
JS	.27	.03	.20	.19	.07	-.03	.28	.29	.18	.03	-.12	.12	.21	.01	.38	.22	-.15		.50**	.14
CS	.26	.22	.16	.26	.04	.26	.12	.16	.19	.05	.16	.28	.28	.15	.28	.08	-.29	.66		.15
DMQ	.17	.15	.17	.16	.03	.19	.04	.02	.05	.20	.23	-.04	.23	.04	.12	.14	-.06	.20	.25	

Note. * $p < .05$; ** $p < .01$

RAT = Rationality factor; BBS = Belief bias syllogisms; BRN = Base-rate neglect; CBR = Causal base-rate; FCS = Four-card selection task; ATF = Attribute framing; OB = Outcome bias; SC = Sunk cost; CRT = Cognitive reflection test; AOT = Actively open-minded thinking; SJT = Actively open-minded thinking situational judgment test; EXT = Extraversion; AGR = Agreeableness; CON = Conscientiousness; STA = Emotional stability; OPE = Openness; DOI = Decision outcome inventory; JS = Job satisfaction; CS = Career satisfaction; DMQ = Peer rated decision-making quality.

Before proceeding with factor-analyzing our CB measures, we checked if the data was adequate for performing such analysis. KMO (KMO = .75) was again acceptable and the Bartlett's test showed that the correlation matrix was not an identity matrix ($\chi2(21) = 131.27$, $p < .001$) meaning that our data was appropriate for conducting factor analysis. We again conducted parallel analysis and scree plot inspection in order to decide on the most appropriate number of factors. Again, both the parallel analysis and scree plot indicated that a one-factor solution was the most appropriate one (output of this analysis is shown in the Appendix B). A one-factor solution obtained using a maximum likelihood extraction method is presented in Table 17. This solution showed an excellent fit to the data ($\chi2(14) = 11.54$, $p = .64$) and one factor was able to explain 22% of the variance in our CB tasks. Although a single factor managed to account for substantially larger part of the CB tasks variance compared to the Study 1, this is still relatively modest amount of common variance shared by CB tasks. However, it seems that the co-variation in our CB tasks scores was not random but at least to some degree under the influence of factor that could be labeled as the rationality factor. As in Study 1, this factor was mainly saturated with base-rate neglect, causal base-rate and belief bias scores, with outcome bias, four-card selection and sunk cost scores showing greater loadings than in the Study 1. Conversely, attribute framing exhibited somewhat lower loading compared to Study 1, being the only score whose loading did not exceed the value of .30.

Table 17. One-factor model of Study 2 CB tasks

Cognitive bias task	Loadings
Base-rate neglect	.62
Causal base-rate	.61
Belief bias	.52
Outcome bias	.44
Four-card selection	.43
Sunk cost	.38
Attribute framing	.13

Again, to investigate the nature of our rationality factor and test its convergent and predictive validity, we correlated it with number of different variables and outcomes. We again calculated the rationality score using a regression method based on factor loadings. We report these correlations in Table 16. Raw correlations are presented above the diagonal while the disattenuated ones are below the diagonal.

Table 16 offers several interesting insights. First, the rationality factor again correlated highly with the numeracy measure (cognitive reflection), as well as with two actively open-minded thinking measures. Indeed, these three were the highest correlations between the rationality factor and any of the variables, confirming that cognitive abilities such as numeracy, as well as actively open-minded thinking, lie at the core of the rationality factor that we extracted.

We again conducted a SEM regression analysis to investigate whether both numeracy and actively open-minded thinking independently predict rationality factor and to identify the portion of the variance that they account for. Similarly to the Study 1, rationality latent variable was defined as a second-order factor of seven CBs that were each defined by their corresponding manifest variables. Numeracy (again captured with the Cognitive Reflection Test) and actively open-minded thinking were first order factors defined by their six and 13 manifest variables respectively (measurement model fits and outcomes of the regression analysis are described in the Appendix B). Both numeracy and actively open-minded thinking were significant predictors of the rationality factor and together they explained 75% of the variance in the rationality factor. Dominance analysis showed that numeracy contributed by explaining 56% and actively open-minded thinking with additional 19% of the variance in Rationality factor. Other notable correlations of the rationality factor are with two of the personality traits, namely emotional stability and openness. Both of the traits are positively and moderately correlated with the rationality factor. We investigated whether any of these traits were able to account for additional variance in rationality beyond cognitive reflection and actively open-minded thinking by conducting two additional regression analysis, each with cognitive reflection, actively open-minded thinking and one of the personality traits as predictors. However, neither of the personality factors was significant predictor in the regression equations and the proportion of explained variance practically did not change after including these two traits as predictors. One plausible explanation is that emotional stability and openness are related with rationality because people higher on these two traits are better at actively open-minded thinking and perhaps more reflective, therefore more rational. This replicates the findings of Study 1 that cognitive capacities, especially numerical ability, as well as dispositions (actively open-minded thinking, reflection) make up rational thinking.

Finally, correlations between the rationality factor and real-life outcomes (DOI, job and career satisfaction and peer's perception of decision quality) were generally low. However, it is interesting that

those that scored higher on rationality measures were somewhat more satisfied with their jobs ($r = .18$; $p < .05$; disattenuated $r = .27$) and careers ($r = .19$, $p < .01$, disattenuated $r = .26$).

Discussion

Across the two studies presented in this manuscript, we aimed to investigate a) the dimensionality of relatively large set of CB tasks, b) the correlations between the uncovered rationality factor(s) and other cognitive abilities as well as other measures from its/their nomological network (i.e. actively open-minded thinking, epistemically suspect beliefs and personality traits), and c) the correlations between rationality factor(s) and some real-life outcomes that could depend on good decision making.

Regarding the first research questions, our results were comparable to some of the similar previous studies. The correlations were mostly positive, but small in size in both of our studies (the highest correlations in the Study 1 were between base-rate neglect and belief bias/causal base-rate, [both correlations were $r = .27$], and between base-rate neglect and causal base-rate in Study 2, $r = .37$). Factor-analyzing these tasks, we found that a single factor solution was the best one in both studies. In Study 1, one factor was able to explain 12% of the variance among our variables, while in Study 2 it was able to account for and 22% percent of the variance. Although a single-factor solution turned out to be the most appropriate for our data in both studies, the extracted factors in both studies were quite weak. This leads to the conclusion that is in line with majority of previous studies that investigated the dimensionality of CB tasks: these tasks are quite heterogenous and idiosyncratic which reflects in low levels of shared variance among them (e.g. Aczel et al., 2015; Blacksmith et al., 2019; Ceschi et al., 2019; Berthet, 2021; Teovanović et al., 2015). Therefore, although our results indicate that there exists some common core that accounts for a success on many different CB tasks, it seems that this core is small and not robust enough to be replicated across the studies. However, although small, this common core, showed meaningful relationship with other variables in our studies.

There are several things apparent from the Table 13 and Table 16 that are relevant for our discussion about the validity of the rationality factor that we extracted. First, our rationality score was positively and moderately correlated with fluid intelligence in Study 1. The fact that superior decision making on normative tasks is positively, but modestly, related with fluid intelligence confirms some of the previous findings (e.g. Teovanović et al., 2015; Sobkow, Olszewska & Traczyk, 2020), although some other studies obtained somewhat higher correlations between these types of tasks (e.g. Blacksmith et al., 2019). Therefore, it seems to make sense to separate the construct of rationality from fluid intelligence, as some

researchers advocate (e.g. Bruine de Bruin et al., 2020; Stanovich, 2012; Stanovich et al., 2016). While fluid intelligence certainly underpins quality decision making, it is not a synonym for it, resulting in, among other, many cases of "dysrationalia" (Stanovich, 2002, 2009b; Erceg et al., 2019).

Second, the highest correlations that the rationality score showed with any of the variables was with measures of numeracy and cognitive reflection (that seems to be indistinguishable from numeracy; Attali & Bar-Hillel, 2020; Erceg et al., 2020). This is also a common finding as many previous studies have shown that numeracy is the strongest predictor of good decision-making both in real life and on CB tasks (e.g. Allan, 2018; Cokely, Feltz, Ghazal, Allan, Petrova & Garcia-Retamero, 2018; Garcia-Retamero, Sobkow, Petrova, Garrido & Traczyk, 2019). In fact, strong disattenuated correlations between the rationality score and numeracy/cognitive reflection show that these two constructs overlap to a large extent. Cokely et al. (2018), as part of their skilled decision theory, propose that numeracy and decision-making skills share number of common processes, including metacognitive, heuristic, intuitive, affective, subjective, gist-based, and number-sense processes. They conclude that statistical numeracy, a type of numeracy measured by the Berlin Numeracy Test, predicts decisions because statistical numeracy tests are relatively representative judgment and decision-making tasks whose solving requires the same kinds of reasoning and metacognitive skills essential for good decision making. These metacognitive skills that are shared among numeracy and rationality task could be related with the disposition to be more careful, thorough and elaborate in solving problems. For example, numeracy and cognitive reflection predicted a higher number of verbalized considerations on risk decision-making tasks which was positively related both to the number of normative correct responses and to the response times (Cokely & Kelley, 2009). Similarly, it has been shown that participants that scored higher on statistical numeracy performed better on various tasks (lotteries, intertemporal choice, denominator neglect, and confidence judgments) because they deliberated more during decision making and, in that way, more accurately evaluated their judgments (Ghazal, Cokely & Garcia-Retamero, 2014). Therefore, apart from fluid intelligence and quantitative reasoning, thinking dispositions that predispose a person towards more careful and elaborate cognition seem to also be important for different cognitive biases and, thus, lie at the core of our rationality factor.

This brings us to the third notable conclusion following from our correlation tables which suggests that the most important disposition that underpins rationality and makes a person more careful, thorough and elaborate in solving problems and making decisions is actively open-minded thinking. In fact, in both of our samples the correlations between the rationality score and actively open-minded thinking were

comparable to those of rationality and numeracy. In this regard, our results replicate previous findings on the importance of actively open-minded thinking for success on heuristics and biases tasks (e.g. Stanovich et al., 2016; West, Toplak & Stanovich, 2008). Given that rationality tasks are performance-based measures while actively open-minded thinking was measured with self-report scales, correlations this high are quite remarkable. Even a newly developed and short situational judgment measure of actively open-minded thinking exhibited a quite high correlation with the rationality score, especially when looking at disattenuated correlations. In fact, constructs of numeracy and actively open-minded thinking explained a majority of variance in the rationality factor (61% in Study 1 and 75% in Study 2), with numeracy apparently being somewhat more important predictor, as indicated by the dominance analyses. In addition to fluid intelligence, numeracy and actively open-minded thinking, personality traits of emotional stability and openness were also relatively highly correlated with rationality. However, neither of these traits managed to explain additional portion of variance in the rationality score above what was explained by numeracy and actively open-minded thinking.

Therefore, our findings paint a picture of rationality as a complex construct reflecting cognitive abilities, such as fluid intelligence and quantitative reasoning, as well as thinking dispositions, such as dispositions to be reflective (as opposed to being impulsive; Baron [2018] calls this disposition reflection/impulsivity or R/I) and actively open-minded. Putting these findings into a broader perspective, we can conclude that they align nicely with the so-called tripartite theories that extend the popular dual-process theories (e.g. Stanovich, 2009a, 2012; Evans, 2019). Stanovich's tripartite theory distinguishes between "three different minds", an autonomous mind (that is called Type 1 processing in dual-processes theory), an algorithmic and a reflective mind (these two are the extensions of Type 2 processes). According to this theory, unlike for the classical fluid intelligence tasks, when solving tasks that assess rationality, a person needs to first recognize the need to suppress and override responses generated by the autonomous mind (reflective mind) and only then to have sufficient computational power to replace this initial response by calculating a new, correct one (algorithmic mind). Therefore, in order to be successful on rationality tasks, a person needs to possess adequate dispositions by which he/she will be able to recognize the need to suppress and correct an initial, autonomously generated response, as well as adequate intelligence that will allow him/her to come up with a correct response. In his recent work, Evans (2019) introduced the so-called Type 3 processes that predispose a person to check one's intuition, therefore being conceptually similar to Stanovich's reflective mind, with both of these concepts clearly relating to the previously discussed dispositions of reflection and actively open-minded thinking.

Our results are remarkably in line with the ideas put forward by Baron (1985) who claimed that biased responding, i.e. departures from normative responding, is most often in one direction: searching too little for possibilities, evidence and goals and ignoring and discounting evidence against our favorite position. In order to combat these biases, he proposed the prescriptive model of actively open-minded thinking, defining the way people should form beliefs and make judgments and decisions so that their responses approximate normative ones as much as possible given cognitive and environmental limitations. Practically, in this framework, actively open-minded thinking is rational thinking. We believe that our results corroborate this idea that has at least two important implications. First, measuring rationality using CB tasks is not the only way and probably not even the best way of capturing rationality in thinking and decision making as the tasks are highly idiosyncratic and heterogeneous. Developing good indicators of real world actively open-minded thinking would perhaps be a more fruitful direction, one which we tried pursuing by developing a pilot version of situational judgment test of this disposition. This is something that definitely deserves more work in the future. Second, unlike fluid intelligence and many other types of cognitive abilities that predominately depend on individual's capacities that are relatively stable and hard to change, actively open-minded thinking is a thinking disposition that could be teachable and changeable. In other words, it is probably possible to teach a person to be more rational through teaching him/her how to apply the principles of actively open-minded thinking in real-life settings, while it would be quite hard to substantially raise someone's score on fluid intelligence tests through instruction.

In discussing the broader perspectives, another question that our findings could inform is the one related to the position of rationality and/or decision-making skills among other cognitive abilities, especially within probably the broadest and most known model of human intelligence, the Cattell-Horn-Carroll (CHC) model. This model represents human cognitive abilities along the three stratums differing in the specificity/generality of cognitive abilities. The most specific abilities constitute the first stratum. These specific abilities then group together to form broad abilities that constitute the second stratum and are perhaps most commonly discussed in the literature (e.g. fluid intelligence, crystalized intelligence, short and long-term memory, speed of processing). Finally, the most general factor, the g-factor constitutes the third stratum and represents the shared variance of second stratum abilities. As we and many other researchers have shown that rationality and/or decision-making competence represents a somewhat distinct construct from the fluid intelligence, it appears that it is not adequately represented in the most complete and broadest taxonomy of cognitive abilities. Others have already took notice of this and suggested the ways in which rationality could be accommodated in this taxonomy (e.g. Alan, 2018; Cokely et al, 2018). These suggestions also align nicely with our results, ending up with the proposition

that rationality/decision-making skills should be represented as a separate broad ability in the second stratum of CHC taxonomy, one that is more underpinned by thinking dispositions than cognitive capacities and one that could be defined and measured with different CB tasks, statistical numeracy that seems to be highly dependent on thinking dispositions, and different measures of thinking dispositions, especially actively open-minded thinking.

The final point of our discussion related to our main problems represents the relationship between the rationality score and different outcomes in which two things become apparent. First, rationality was related to holding more correct beliefs about the world, as seen from its negative correlation to superstitious thinking in Study 1 (although, it did not correlate with conspiracy thinking). As holding more correct beliefs about the world is the definition of epistemic rationality (Stanovich et al., 2016), it seems that rationality as measured with ability to suppress cognitive biases is related to epistemic rationality. Second, rationality exhibited quite low and non-significant correlations with real life decision-making outcomes (DOI). The question is why, and we believe that the part of the answer can be deduced from the relationships between different personality traits and these outcomes. As is evident from the Table 16, traits of conscientiousness and emotional stability were the most predictive of DOI. Therefore, it seems that these long-term outcomes are more affected by stable personality characteristics than the quality of reasoning. This means that, no matter how good and rational someone reasoning is, in affecting real life outcomes this will probably be overcome by whether one is careful or self-disciplined (i.e. conscientious) or anxious and prone to excessive emotional reactions (i.e. emotional stable). A glimpse of the relevance of the rationality perhaps comes from its low, but positive correlation with job and career satisfaction. However, these low correlations between rationality and real-life outcomes need to be viewed from a broader perspective, namely the fact that practically neither of the cognitive ability variables in our two studies (i.e. fluid intelligence, numeracy, cognitive reflection) were meaningfully related with these real life outcomes. This probably reflects the fact that multiple determinants, including luck, work together and sometimes probably in opposite direction to produce these outcomes, diminishing the effects of single individual determinants.

Finally, we will briefly comment on what we see as the probably the biggest downside of our study, namely the low reliability of some of our measures. This in particular refers to two framing measures whose score is calculated as the difference of the scores on two different question versions. These types of scores, the difference scores, are long known to suffer from the reliability problems (e.g. Peter, Churchill Jr & Brown, 1993). Although the discussion about the reasons for and remedies of this problem

is beyond the scope of this article, we can offer a brief speculation about what caused low reliability in our difference scores measures and what could perhaps be done about it. In his article, Trafimow (2015) writes that the reliability of difference score depends on the reliabilities of each of the two forms and the correlation between them. If we take our attribute framing task from the Study 1, the reliabilities of each of the forms were very low ($\alpha = .15$ for the first form, $\alpha = .29$ for the second form). This reflects our items assessing preferences in completely different domains, diminishing the relationships between these preferences. For example, if a person A prefers a consulting firm (attribute framing item 1) more than a person B, this certainly does not mean that he/she will be more satisfied with the quality of public transport (attribute framing item 4) than person B. Therefore, in order to increase the difference score reliability, it will be necessary to raise the reliability of individual forms. Unlike our approach here, where we tried to sample over a large range of domains, it would probably be better to devise domain-specific questions where the reliability would not suffer and the framing effects could still be observed. The other CB tasks had a satisfactory reliability given that they were measured with only few items, with the exception of very low reliability of sunk cost in Study 2, although the items were the same as in the Study 1. Granted, these "satisfactory" reliabilities were lower than it is generally deemed acceptable. This was expected given that each of the biases was assessed with relatively few items. However, this could have lowered the relationships between or CBs. Given this, it is possible that, had the CBs been measured more reliably, perhaps with more items, the positive manifold of CB tasks would perhaps be greater than the one we found. However, we do not think that this would fundamentally change our conclusions in terms of the nature of rationality factor – it would probably still explain quite variance in CB tasks and exhibit similar correlations with other variables that we measured.

Conclusion

Across two studies, we investigated a validity of the rationality factor(s) as assessed by different CB tasks. Our findings can be summed up in following way: a) one factor, a rationality factor, could be extracted from responses on different CB tasks although it accounted for relatively small amount of variance among the tasks; b) this factor of rationality is separate from fluid intelligence, but closely related to numeracy and dispositions of reflection and actively open-minded thinking; and c) consequently, our results add credence to the view that rationality should find its place in the taxonomy of cognitive abilities, at the same level as some other broad abilities such as fluid and crystallized intelligence.

5. STUDY 4: INCREMENTAL VALIDITY OF DECISION-MAKING STYLES IN PREDICTING REAL-LIFE AND WORK-RELATED OUTCOMES

This chapter was previously published as: Erceg, N., & Galić, Z. (2023). Incremental Validity of Decision-Making Styles in Predicting Real-Life and Work-Related Outcomes. *Journal of Individual Differences.*

Introduction

Decision-making styles are defined as learned, habitual response patterns exhibited by an individual when confronted with decision situations (Scott and Bruce, 1995). More recently, Thunholm (2004) broadened the definition to include some aspects of well-known psychological characteristics of decision-makers. According to him, in addition to task and decision-making situations, habitual patterns of responses are influenced by individual differences in basic cognitive abilities, self-evaluation and self-regulation. A number of studies supported this more encompassing definition by showing that decision-making styles are substantially correlated with cognitive abilities and personality traits (e.g., Dewberry et al., 2013a; Gambetti & Giusberti, 2019; Juanchich et al., 2016; Ülgen et al., 2016; Ward, 2016; Wood, 2012; Wood & Highhouse, 2014). However, this has raised the questions about incremental validity of decision-making styles for important real-life and work outcomes, beyond well-established effects of personality traits and cognitive ability (Schmitt, 2014). In response to the recent calls (e.g., Dalal & Brooks, 2014; Dalal et al., 2010; Moore & Flynn, 2008), this paper reports results of a research program that investigated whether decision-making styles exhibit incremental validity in predicting various outcomes over and above cognitive abilities and personality traits.

Decision-making styles: definition and measurement

One of open questions in the decision-making styles realm is the question of number of styles that can adequately capture various ways in which people approach decision-making process. The propositions range from two (e.g., rational decision style and intuitive decision style; Hamilton et al., 2016) to seven styles (e.g., vigilant, intuitive, spontaneous, dependent, anxious, brooding, and avoidant decision-making style; Leykin and DeRubeis, 2010). Thunholm (2004) noticed that rational (or vigilant in different models) and intuitive decision-making style resemble analytic and intuitive dimensions of general cognitive styles (Kozhenikov, 2007) and reflect differences in approach to information gathering and evaluation. However, the other decision-making styles such as dependent or avoidant style cannot be positioned on the same information gathering/evaluation dimension but reflect different psychological

processes. Thus, individuals' approaches to decision-making cannot be evaluated only in terms of whether they are looking for as much information as possible prior to making decisions in the case of rational, or whether they are relying on hunches and feeling when making decisions as in the case of intuitive decision-making style, but the conceptual framework needs to be expanded in order to accommodate for other identified styles.

In one such attempt, Dewberry et al. (2013b) distinguished between differences in cognitive processes that people use to make decisions (captured by rational/vigilant and intuitive styles) and regulatory processes concerned with choice regulation (captured by avoidant, dependent, or anxious style). More specifically, they related the three styles concerned with cognitive processes in decision making (rational/vigilant, intuitive and spontaneous) with well-known dual-process theory (Kahneman, 2011) that differentiates between System 1 (intuitive, automatic, associative, fast) and System 2 (analytic, explicit, rule-based, relatively slow). Consistent with the distinction between the two systems, intuitive and spontaneous styles seemed to indicate the extent to which individuals rely on heuristic/System 1 processing, while rational/vigilant style pertains to deliberate/System 2 processing. In their model, the remaining styles are not related with the two systems of information processing but more with the regulation of choice - the extent to which decisions tend to be delayed or avoided (avoidant style), referred to others (dependent style) or followed by negative affect (anxious style). The main insight here is that the feeling of anxiety over making decisions underpins the three regulatory styles.

The most used and widely cited model of decision-making styles seems to be the one proposed by Scott and Bruce (1995). Their model is broad enough to encompass both styles related with cognitive processes and those related with regulation processes. Specifically, Scott and Bruce (1995) proposed five different decision-making styles and defined them in behavioral terms: rational (a tendency towards thorough search for and logical evaluations of alternatives), intuitive (an inclination to rely on hunches and feelings), dependent (a propensity to search for advice and direction from others), avoidant (a proclivity to avoid decision making) and spontaneous (a sense of urgency to finish decision-making process as soon as possible). They also showed that the decision-making styles were not highly intercorrelated indicating that individuals do not rely on a single style but can use several of them to various extent. The factorial structure of Scott and Bruce's model of decision making operationalized with General Decision Making Styles (GDMS) questionnaire was further supported in several additional studies (Loo, 2000; Spicer and Sadler-Smith, 2005; Thunholm 2004).

Predictive validity of decision-making styles

Most of the studies testing validity of decision-making styles focused exclusively on rational and intuitive styles. For example, recent meta-analysis by Phillips et al. (2016) showed that rational/analytical style positively predicted both decision performance (normatively correct responding) and decision experience (speed and enjoyment) whereas intuitive style was negatively related to performance but positively with decision experience. However, absolute sizes of the meta-analytically estimated correlations were small (between .06 and .14). When it comes to work-related outcomes such as job performance or job attitudes, empirical evidence seems to be sparse and also limited to the two styles. For example, a recent meta-analysis (Alaybek et al., 2021a) found a significant positive correlation between rational style and task performance, but non-existent relationship between intuitive style and the same outcome. Similarly, a study conducted in vocational behavior domain by Singh and Greenhaus (2004) showed that the use of the rational style was related to higher levels of person-job fit, whereas there was no relationship between intuitive style with the same criterion. Contrary to this, Crossley and Highhouse (2005) showed that more frequent reliance on both rational and intuitive decision-making style while making job choice reflected in higher job satisfaction and satisfaction with the job search process.

Several studies explored relevance of decision-making styles in organizational leadership context, but again only looking into rational and intuitive styles. For example, Agor (1986) showed that frequency of intuitive decision-making increases on higher levels of organizational hierarchy and that managers experienced excitement and harmony when making intuitive judgments. Sadler-Smith (2004) reported a significant correlation between entrepreneurs' inclination towards using intuitive decision-making style with financial and non-financial performance of small and medium sized companies. In the same study, the rational style showed non-significant relation with the criteria.

A handful of studies did examine a broader range of decision-making styles, consistently finding that the choice-regulation styles are relevant for various outcomes, sometimes even more than rational and intuitive style. For example, Bruine de Bruin et al. (2007), using a community sample, showed that participants who relied more on rational and intuitive decision-making style experienced lower number of negative outcomes that followed their decisions, whereas those that relied more on spontaneous style or avoided making decisions experienced more negative decision outcomes, as measured by the Decision Outcome Inventory (DOI). Dependent decision-making style did not show significant relationship to the criterion Wood (2012) also correlated individual differences in Scott and Bruce's decision-making styles with peer reports of decision-making quality on a sample of undergraduates. In her study, rational style

was positively correlated with peer ratings of general decision-making quality but also with a reputation of reasonable decision maker, while the spontaneous style was negatively correlated with those outcomes. The avoidant style was also negatively correlated with reputation of a reasonable decision-maker.

There seems to be only few studies that explored the relationship of all the five decision-making styles concurrently with work-related outcomes. For example, Russ et al. (1996) correlated the decision-making styles of sales managers with ratings of their overall performance and behavioral performance ratings (operationalized as supervisor's ratings of nine behaviors affecting the sale force). Rational, avoidant, and spontaneous styles predicted overall performance ratings whereas behavioral performance ratings was predicted only by avoidant style. Consistent with expectations, rational style correlated positively and avoidant/ spontaneous styles showed a negative correlation with the criteria.

Thunholm (2008, 2009) explored validity of the five decision styles in military context. Thunholm, (2009) showed that military team leaders rated themselves as more spontaneous but less rational, dependent or avoidant decision-makers than their team members. In another study (Thunholm, 2008), the same author related the decision-making styles with negative stress experienced by leaders when making demanding decisions. Out of the five styles, only individual differences in avoidant style were (positively) correlated with stress as indicated with saliva cortisol levels.

Incremental validity of decision-making styles above the effects of cognitive ability and personality traits

To the best of our knowledge there is a lack of evidence about incremental validity decision-making styles can add over and above more often studied predictors, such as cognitive abilities and personality traits. The research on the relationship between decision-making styles and cognitive abilities is relatively sparse. Recent meta-analysis (Alaybek et al., 2021b) covered only rational and intuitive decision-making styles and showed that only rational style was significantly related to intelligence but the size of the relationship was small. Similarly, it has been shown that there is very little overlap between decision-making styles and cognitive reflection as measured with the well-known Cognitive Reflection Test (Juanchich et al., 2016). Among the five decision-making styles, the intuitive style was the only one that was significantly related to cognitive reflection and even that correlation was low in size ($r = -.15$). Therefore, it seems that, to the extent that decision-making styles manage to explain some part of the variance in different outcomes, that variance should be largely independent from cognitive abilities. This was recently partly confirmed in a meta-analysis where (Alaybek et al.(2021a) showed that rational

style incrementally predicted task performance beyond both intelligence and conscientiousness. However, but the picture was much less clear for other styles indicating a need for further research.

Although some research found that the styles were predictive of certain outcomes beyond the effects of personality (e.g., Alaybek et al., [2021a] meta-analysis, also Wood [2012] who found that rational style uniquely predicted peer-rated decision-making quality after accounting for the effects of personality traits), it is nevertheless still a question how much of a predictive validity the five Scott and Bruce's decision-making styles can add over personality traits. For example, in Juanchich et al. (2016) study, the five decision-making styles failed to predict DOI beyond the effects of personality traits. Furthermore, it has consistently been shown that the decision-making styles are substantially related with most of the Big 5 traits. Most notable of those correlations are generally medium to high correlations between conscientiousness and rational style (positive), conscientiousness and avoidant style (negative), and neuroticism and avoidant style (positive; Dewberry et al., 2013a; Juanchich et al., 2016; Wood, 2012; in Ülgen et al., 2016).

Current study

Our short literature review reveals several gaps in the literature. First, although the two styles related to cognitive processes of information gathering and evaluation (rational and intuitive) were relatively extensively studied, the styles related to choice regulation (e.g. avoidant and dependent) have been less frequently researched, especially in the workplace context. Second, given that the decision-making styles are correlated with cognitive ability and, especially, personality traits, it is still not clear whether they can add anything above these more common predictors in predicting important real-life outcomes.

Therefore, the goal of our research program reported in this paper was to extend the findings of relevance of decision-making styles for a broad range of outcomes across three different samples (undergraduate students in Study 1, employed adults in Study 2 and entrepreneurs/leaders in Study 3) and to see whether styles can uniquely contribute to predictiveness of outcomes beyond effects of cognitive abilities (Study 1), personality (Study 2), and motivational variables (Study 3).

Study 1

In Study 1, we wanted to see whether the five Scott and Bruce's decision-making styles will be able to predict decision outcomes in a sample of undergraduates. In this study, we wanted to see if we can replicate earlier findings about the relationship of decision-making styles with negative decision

outcomes (Bruine de Bruin et al., 2007) and academic achievement (Baiocco et al., 2009) but also expand the criterion domain with a variable that parallels counterproductive work behavior (Spector et al., 2010) in the academic domain. Moreover, we wanted to investigate whether the decision styles will remain important for decision outcomes after the effects of cognitive abilities, such as intelligence and numeracy, are taken into account.

Methods

Sample

A total of 253 undergraduate University of Zagreb students participated in this study (214 from the Faculty of humanities and social studies – mostly psychology students, 34 from other faculties and five undeclared). There were 187 females, 62 males and four participants refused to give their gender. The mean participants' age was 21.47 (SD = 1.89; Min = 18, Max = 29). We describe the detailed procedure in the Appendix B.

Instruments

General decision-making style. To measure decision-making styles, we used Scott and Bruce (1995) General Decision Making Style (GDMS) scale that assesses five different decision-making styles: rational, intuitive, avoidant, spontaneous and dependent. Each of the styles is assessed by five self-report items and the task of participants was to indicate their level of agreement with each of the statements on a five-point scale (1 = completely disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, 5 = completely agree). Total scores were calculated by averaging responses on items capturing each of the styles. We report all of the items for all our instruments in the Appendix A.

Intelligence. We measured intelligence with a 16 items version of the International Cognitive Ability Resource (ICAR; for details see icar-project.com and Condon and Revelle, 2014). ICAR is a broad cognitive ability assessment tool consisting of four different types of tasks: letters and numbers series, matrix reasoning items, verbal reasoning items and three-dimensional rotation items. We formed total score as an average of responses to these 16 items.

Numeracy. We used The Berlin Numeracy Test (BNT; Cokely et al. 2012), a four-question test assessing numeracy and risk literacy. The questions are designed in a way that they gradually become harder and we calculated total score as a ratio of number of correct responses and total number of responses.

Negative decision outcomes. We measured negative decision outcomes (NDO) with a 33-item version of Decision-Outcome Inventory (DOI). DOI was developed by Bruine de Bruin et al. (2007) to capture decision-making success by assessing avoidance of negative decision outcomes. Total score was calculated by averaging the experienced outcomes weighted by their severity[9].

Counterproductive academic behavior. Apart from the total NDO score, we also calculated a score that parallels counterproductive work behavior in the academic domain. Specifically, among the 33 DOI items, there were six that were related with students' behavior related to academic obligations. We averaged the score on these six items to get a counterproductive academic behavior indicator.

Academic achievement. We measured academic achievement by asking our participant to report their average college grade (grades in Croatia range from 1 = unsatisfactory to 5 = excellent).

Results

Prior to calculating correlations among our variables, we have examined their distributions and calculated skewness and kurtosis values. Given that Pearson's product-moment correlation is fairly robust to minor departures from normal distribution (e.g. Bishara & Hittner, 2012; De Winter et al., 2016) and that our data (across all three studies) did not show any extreme departures from normal distribution (the highest skewness absolute value was 1.50 and the highest kurtosis absolute value was 2.62 across all three samples which is far from extreme non-normality; e.g. Kline, 2015), we report Pearson's correlations in every study. When calculating correlation coefficients, we excluded missing values pairwise which is why the sample sizes for correlations vary between N = 244 and N = 253, depending on the variable (e.g. grade point average had the most missing data, thus the sample size for analyses that included this variable was N = 244). Other than missing values, we did not exclude additional data from the analyses. The descriptive statistics and reliabilities of Study 1 focal variables, together with correlations among them, are shown in Table 18.

[9] See our Appendix B for a detailed explanation about items and total score calculations.

Table 18. Descriptive statistics and reliabilities of Study 1 focal variables, together with Pearson's correlations among them

	Mean	SD	Min	Max	α	Int	Avo	Spon	Dep	FI	BNT	NDO	CAB	AA
Rat	4.13	0.59	1	5	.80	-.02	-.11	-.34**	.19**	.09	-.09	-.13*	-.13*	.04
Int	3.37	0.71	1	5	.82		.12	.45**	.13*	-.02	-.22**	.08	.07	-.18**
Avo	2.74	1.07	1	5	.92			.22**	.30**	.08	.02	.30**	.25**	-.15*
Spon	2.58	0.76	1	5	.80				.01	-.03	-.02	.25**	.28**	-.21**
Dep	3.64	0.84	1	5	.83					.06	-.06	.12	-.04	-.06
FI	0.67	0.18	0.06	1	.70						.37**	.02	.00	.05
BNT	0.42	0.27	0	1	.42							.04	.10	-.04
NDO	16.72	7.67	1.29	42.33	.69								.74**	-.20**
CAB	15.52	15.95	0	65.17	.55									-.31**
AA	3.90	0.57	2.30	5.00	/									

Note.

** p < .01; * p < .05 (two-tailed)

α = Cronbach alpha; Rat = Rational decision-making style; Int = Intuitive decision-making style; Avo = Avoidant decision-making style; Spon = Spontaneous decision-making style; Dep = Dependent decision-making style; FI = Fluid intelligence; BNT = Berlin Numeracy Test; NDO = Negative decision outcomes; CAB = Counterproductive academic behavior; AA = Academic achievement.

It is evident from Table 18 that, in our undergraduate sample, cognitive abilities (intelligence and numeracy) were not correlated with either NDO, CAB or academic achievement. At the same time, decision-making styles were related to each of the three outcomes. Specifically, students that tend to avoid making decisions or to make them quickly and spontaneously experienced more negative real-life and college related outcomes and had lower college grade point average. Moreover, these decision-making styles were practically unrelated to the ability measures indicating that they do not tap into same constructs. Still, it has to be noted that the internal consistencies of some of these measures were quite low (especially for the numeracy and CAB measures) which could artificially reduce the size of correlations.

Study 2

In Study 2 we wanted to extend our Study 1 findings in several important ways. First, we recruited a sample of employed adults that is more heterogeneous in its socio-demographic characteristics than was the Study 1 student sample. The Study 2 participants were recruited by psychology undergraduates who

received course credits in exchange. Second, along with the five decision-making styles, we also measured personality traits, enabling us to investigate the styles' incremental validity above the personality traits. Third, we captured a much wider range of potential decision outcomes. In addition to DOI, we measured peer-reported decision-making quality and important work-related outcomes – job attitudes (job and career satisfaction), and job performance (in-role performance and counterproductive work behavior).

Methods

Sample

In total, 210 participants and 354 of their peers (colleagues from work who rated our participants – see Appendix B for explanation and description of peers) participated in our Study 2. As our study was divided in two parts separated by a week time (see Appendix B for a description of the procedure), there was some loss of participants. Specifically, 183 participants also participated in the second part of the study. There were 79 male and 103 female participants (28 unknown). The average participants' age was M = 34.31 (SD = 10.63, Min = 19, Max = 63). Regarding the education, we had one participant with primary school only, 38 with high school, 41 with college education, 87 with higher education and 16 with PhDs.

Instruments

In addition to Scott and Bruce (1995) GDMS scale described earlier, we used following instruments:

Mini International Personality Item Pool questionnaire (Mini IPIP). Mini IPIP (Donnellan et al., 2006) is a 20-item personality measure, measuring the Big 5 traits each with four items. Participants were instructed to rate the accuracy of the description on a five-point scale (1 = completely incorrect, 2 = somewhat incorrect, 3 = neither correct nor incorrect, 4 = somewhat correct, 5 = completely correct) and total score is the average of the ratings on the four items per trait, after the items were recoded so that a higher score indicates a higher degree of a trait.

Cognitive reflection test (CRT). CRT consist of items with a distinctive characteristic of pitting an intuitive, but incorrect responses against a correct one. In this study, instead of the original version of the

test (Fredrick, 2005), we used six items, each cuing strong but incorrect intuitive response[10]. The average of correct responses represented total score.

Negative decision outcomes. We measured NDO again using DOI. However, in comparison with Study 1, DOI was significantly shorter, consisting of 18 items[11].

Job satisfaction. We measured job satisfaction with a following single item: "Think about the job you are doing now. Weigh all its advantages and disadvantages and then assess how satisfied you are, as a whole, with your job." Participants rated their satisfaction on a five-point scale (1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = neither satisfied nor dissatisfied, 4 = somewhat satisfied, 5 = very satisfied).

Career satisfaction. We measured career satisfaction with the five-item Career Satisfaction Scale (Greenhaus et al., 1990). The participants were instructed to indicate to what extent they agreed or disagreed (1 = completely disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, 5 = completely agree) with career-related statements (see Appendix A). Total score is calculated as a mean of the scores on these five statements and higher score indicates higher career satisfaction.

In-role job performance. This construct was measured with the In-role Behaviors Scale (Williams & Anderson, 1991) consisting of seven statements. The participants rated their agreement with these statements on a five-point scale (1 = completely disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, 5 = completely agree). Total score was calculated by averaging the ratings on the seven statements and higher scores indicated better in-role performance.

Counterproductive work behavior (CWB). We measured CWB with a short, 10-item measure developed by Spector et al. (2010). The participants were instructed to rate how frequently they displayed different counterproductive work behaviors at their workplace during the last six months (1 = never, 2 = once, 3 = twice, 4 = several times, 5 = once a month, 6 = once a week, 7 = every day). We averaged these ratings to calculate the total score.

[10] See Appendix A for a more detailed description of CRT items.

[11] See Appendix A for the items.

Peer-rated general decision-making quality. The scale was developed by Wood (2012) and consists of four items. The participants' peers were instructed to rate their agreement with statements describing their colleague/friend participants on a five-point scale (1 = completely disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, 5 = completely agree). Each participant was rated by up to three different peers. Therefore, before calculating final scores for these two scales, we averaged peer's ratings on the four items per scale. The final score is calculated as an average of ratings and the higher it is, the better peers perceive decision-making abilities of their colleague.

Results

We report descriptive statistics and reliabilities of our focal variables in Table 19 and the correlations among those variables in the Table 20. Pairwise exclusion of cases this time resulted in sample sizes varying between N = 166 and N = 187. Majority of the missing values was due to attrition between two time points of data collection, as described previously, not every participant managing to obtain peer-ratings or participants and peers leaving some questions unanswered.

Table 19. Descriptive statistics and reliabilities of focal Study 2 variables

	M	SD	Min	Max	α
Education	3.43	0.93	1	5	/
Rational	4.33	0.44	3	5	.71
Intuitive	3.36	0.78	1	5	.83
Avoidant	2.36	1.00	1	5	.90
Spontaneous	2.46	0.73	1	4.20	.74
Dependent	3.84	0.68	1.60	5	.76
Extraversion	3.50	0.76	1.25	5	.76
Agreeableness	3.92	0.64	2	5	.70
Consc.	3.73	0.73	1.75	5	.74
Emotional st.	3.48	0.71	1.50	5	.74
Openness	3.77	0.74	1.25	5	.75
CRT	0.51	0.34	0	1	.77
NDO	19.02	9.49	0.76	47.59	.52
Job Satisf	3.94	0.90	1	5	/
Career Satisf	3.57	0.80	1.40	5	.82
In role bahavior	4.49	0.47	2.43	5	.84
CWB	1.79	0.67	1	4.40	.72
DQ-G	4.34	0.49	3	5	.75

Note. α = Cronbach alpha; Consc. = Conscientiousness; Emotional st. = Emotional stability; CRT = Cognitive reflection test; NDO = Negative decision outcomes; CWB = Counterproductive work behavior; DQ-G = Peer-rated general decision-making quality.

Table 20. Pearson's correlations among the Study 2 variables

	Sex	Edu	Rat	Int	Avo	Spon	Dep	Extra	Agr	Cons	Stab	Open	CRT	DOI	JSat	CSat	InR	CWB	DQ-G
Age	.09	-.08	.11	.03	-.19*	-.15	-.15*	-.08	.05	.19*	.11	-.21**	-.15*	-.23**	.03	.09	.02	-.13	.02
Sex		.03	-.03	.23**	.02	.02	.07	.01	.27**	.16*	-.16*	-.16*	-.38**	.04	.01	.02	.22**	-.04	-.01
Edu			-.04	-.06	-.09	-.09	.14	.07	.00	.01	.13	.08	.24**	-.21**	.27**	.38**	.25**	-.23**	-14
Rat				.00	-.21**	-.25**	.17*	-.14	.01	.28**	.07	-.06	-.04	-.21**	.01	.06	.13	-.11	.24**
Int					.17*	.39**	.10	.06	.01	.07	-.10	-.19*	-.36**	-.03	.12	.01	-.09	.06	-.20**
Avo						.25**	.26**	-.20**	.03	-.39**	-.30**	.04	-.11	.26**	-.17*	-.22**	-.30**	.18*	-.12
Spon							.04	.20**	-.06	-.18*	-.12	-.03	-.16*	.21**	-.03	-.15*	-.13	.22**	-.20**
Dep								-.15*	.16*	-.04	-.15*	-.18*	.03	.05	-.07	.02	.06	-.08	-.05
Extra									.17*	-.09	.16*	.15*	.06	.10	.09	.22**	.22**	.10	-.05
Agr										-.06	.05	.16*	-.06	.02	.15*	.21**	.28**	-.14	.16*
Cons											.13	-.17*	-.25**	-.30**	.01	.12	.20**	-.18*	-.01
Stab												.06	.31**	-.23**	.27**	.22**	.22**	-.30**	.09
Open													.26**	.11	.16*	.06	.09	-.06	.11
CRT														-.13	.14	.16*	.02	-.13	.06
DOI															-.10	-.21**	-.29**	.41**	-.05
JSat																.50**	.23**	-.43**	.14
CSat																	.38**	-.26**	.15
InR																		-.39**	.14
CWB																			-.10
DQ-G																			

Note. Rat = Rational decision-making style; Int = Intuitive decision-making style; Avo = Avoidant decision-making style; Spon = Spontaneous decision-making style; Dep = Dependent decision-making style; Extra = Extraversion; Agr = Agreeableness; Cons = Conscientiousness; Stab = Emotional stability; Open =Openness; NDO = Negative decision outcomes; JSat = Job satisfaction; CSat = Career satisfaction; InR = In-role performance; CWB = Counterproductive work behavior; DQ-G = Peer-rated general decision-making quality.

** $p < .01$; * $p < .01$ (two-tailed)

Table 20 shows us that most of the decision-making styles and all of the personality traits were correlated to at least one of our outcomes. Conversely, cognitive reflection as a measure of cognitive ability was only significantly (positively) correlated with career satisfaction. Regarding the decision-making styles, avoidant style was related to five outcomes, consistently in a negative way. Specifically, it was related to more negative decision-making outcomes, lower career-satisfaction, higher frequency of counterproductive work behavior, lower job satisfaction and lower self-reported job in-role performance. Spontaneous style was related to more negative decision-making outcomes, lower career satisfaction and greater counterproductive work behavior. Moreover, it was significantly related to more negative peer ratings of participant's general decision-making quality. Rational style was related to two of the seven outcomes, negatively with DOI and positively with peer-rated general decision-making quality, while intuitive style's only significant relationship was a negative one with peer- rated decision-making quality. Dependent style was the only style one that failed to show a significant relationship with any of the outcomes.

Regarding the personality traits, emotional stability was correlated with five out of the six outcomes, failing to correlate only with peer-rated decision-making quality. Agreeableness was positively related to peer-rated decision-making quality, job and career satisfaction and self-reported in-role performance, while conscientiousness showed negative correlations with DOI and counterproductive work behavior and positive with in-role performance. Extraversion was related to two outcomes (positive relationship with career satisfaction and self-reported in-role performance), whereas openness showed only a small positive relationship with job satisfaction.

To investigate whether decision-making styles can add some predictive validity over personality traits in predicting the outcomes, we conducted two types of analysis. First, we conducted classical hierarchical linear regressions where we wanted to test whether decision-making styles show incremental validity over personality traits for our outcomes. Predictors were entered in two subsequent blocks, first the five personality traits followed by five decision-making styles. We have conducted a sensitivity power analysis using G*Power (Faul et al., 2007) that showed that, even when doing analyses on the smallest available sample (N = 166), we were able to detect a small-to-medium effect ($f^2 = 0.11$) with a power of $1 - \beta = 0.8$. This is in line with Tabachnick and Fidell (2013) rules of thumb according to which, given our ten predictors, we should have at least N = 140 cases to detect medium effect size.

However, this type of analysis has serious problems due to imperfect reliabilities of measures (e.g. Baron, Gürçay, & Metz, 2017), resulting in often serious Type I error inflation and unreliable conclusions about significance and importance of predictors (e.g., Westfall & Yarkoni, 2016). To alleviate this problem and to select, among all of the possible models the most appropriate constellation of predictors for our outcomes given the data, we implemented a Bayesian Model Averaging method (BMA; Hinne et al., 2020; van den Bergh et al., 2020). The detailed description of this analysis can be found in the Appendix B.

We report the results of these two analyses in Table 21. For each of the outcomes, we report two columns. In the first column we report beta weights from the second step of the "classical" hierarchical regression analysis (in this step all of the predictors are included), indicating which of the predictors are significant for that specific outcome, as well as the R^2 and ΔR^2. In the second column we report Bayes factors (BFs) indicating the strengths of evidence that each of the predictors is important predictor for that specific outcome and indicate the best model from the BMA analysis together with its model BF (BFm). Bayes factor (BF) can be interpreted in the following way: BFs ranging from 1 to 3 means anecdotal or insufficient evidence for model/predictors, BFs from 3 to 10 means moderate evidence, BFs from 10 to 30 means strong evidence, BFs from 30 to 100 means very strong evidence and BFs greater than 100 means extremely strong evidence.

Table 21. The results of a classical hierarchical regression analysis and Bayesian regression analysis using a Bayesian Model Averaging method

	DOI		JSat		CSat		In-role		CWB		DQ-G	
	β	BF	β	BF	β	BF	β	BF	β	BF	β	BF
Extra	.06	0.40	-.03	0.33	.22[a]	12.29[c]	.16[b]	1.34	.14	1.10	-.08	0.40
Agr	-.03	0.30	.11	1.07	.14	2.49[c]	.22[a]	62.70[c]	-.11	0.99	.18[b]	2.58[c]
Consc	-.16[+]	8.04[c]	-.08	0.44	.08	0.83[c]	.12	0.83	-.10	0.98[c]	-.08	0.46
Stab	-.19[b]	8.37[c]	.25[a]	56.58[c]	.13	1.71[c]	.11	1.02	-.28[a]	232[c]	0.4	0.37
Open	.05	0.39	.16[b]	2.43[c]	.03	0.32	.06	0.39	-.08	0.39	.05	0.48
Rat	-.10	1.02	-.04	0.33	.00	0.36	.04	0.40	.02	0.33	.26[a]	30.42[c]
Int	-.09	0.47	.23[a]	4.93[c]	.07	0.41	-.04	0.36	-.04	0.33	-.15	3.08[c]
Avo	.12	0.96	-.17[b]	1.27[c]	-.12	1.17	-.21[b]	17.69[c]	.11	0.60	-.01	0.34
Spon	.13	1.25[c]	-.05	0.38	-.16	2.00[c]	-.05	0.40	.14	2.08[c]	-.04	0.47
Dep	.03	0.31	.01	0.33	.09	0.52	.14	1.09	-.14	1.07[c]	-.12	0.86
R^2	.19[a]		.17[a]		.17[a]		.23[a]		.19[a]		.15[a]	
ΔR^2	.04		.06		.04		.06[b]		.03		.10[a]	
BFm		34.68		43.18		22.33		26.91		18.49		47.00

Note.

[a] $p < .01$; [b] $p < .05$; [+] $p = .051$;

[c] indicates that the predictor is included in the most appropriate model, i.e. the one with the highest model Bayes factor (BFm)

β = standardized regression coefficient; BF = Bayes factor.

Extra = Extraversion; Agr = Agreeableness; Cons = Conscientiousness; Stab = Emotional stability; Open =Openness; Rat = Rational decision-making style; Int = Intuitive decision-making style; Avo = Avoidant decision-making style; Spon = Spontaneous decision-making style; Dep = Dependent decision-making style; DOI = Decision Outcome Inventory; JSat = Job satisfaction; CSat = Career satisfaction; In-role = In-role performance; CWB = Counterproductive work behavior; DQ-G = Peer-rated general decision-making quality.

Looking at the significance of β coefficients and the magnitude of BFs, it appears that the BMA is somewhat more conservative than the "classical" regression. Specifically, in four instances where the β coefficients for predictors were significant, BMA indicated that there is insufficient evidence for the

importance of predictor (e.g. openness and avoidant style as predictors of job satisfaction, extraversion as predictor of in-role performance and agreeableness as predictor of decision-making quality). Conversely, there was only one instance where the β was not significant, but the BF showed moderate evidence for predictor, although the strength of evidence here barely surpassed the cut-off of BF = 3 for moderate evidence (intuitive style as predictor of general decision-making quality). To reduce the chance of false positives, we will focus on the BMA analysis. Thus, it is apparent that the personality traits, at least one of them, are important predictors for five out of six outcomes (none seem to be important for the prediction of peer-rated general decision quality), whereas decision-making styles are important for only three outcomes (job satisfaction, in-role performance and peer-rated general decision quality).

Study 3

In Study 3 we collected a small, but highly specific sample of entrepreneurs/owners/managers of small businesses. Recent reviews of research showed that individual differences in entrepreneurs' traits predict both entrepreneurs' behavior and business performance (e.g., Collins et al., 2004; Frese & Gielnik, 2014; Pekkala Kerr et al., 2017). Considering that decision making is an essential part of every leadership position (Mintzberg, 2009; Yukl, 2013), we believed that decision-making styles should reflect in subordinate ratings of entrepreneurs' in-role performance, and, specifically, in their leadership behavior thus determining employees' job attitudes and job performance.

In our study, we assessed entrepreneurs' performance with employees' perception of efficacy of their employer in the entrepreneurial role, but also a number of variables reflecting employees' job attitudes and job performance. To be more specific, in our study employees' self-reported about perceived organizational support, job satisfaction and intention to leave the organization (job attitudes), as well as about their own in-role performance and counterproductive work behavior (the two key dimensions of job performance). In this study, we used motivational trait of need for achievement as a benchmark against which we compared the effects of decision-making styles. Several meta-analyses showed that the need for achievement is one of the key psychological characteristics that both differentiate between entrepreneurs and non-entrepreneurs (Collins et al., 2004; Stewart & Roth, 2007; Zhao & Seibert, 2006) as well as between successful and non-successful entrepreneurs (e.g. Collins et al., 2004; Rauch & Frese, 2007).

Methods

Sample

A total of 53 entrepreneurs who were owners and CEOs of small businesses (up to 30 employees) participated in our study (34 males and 19 females; Mage = 47.15, SDage = 10.66, Min = 27, Max = 69). 37% of them had high school, 49% completed college while 14% obtained master's degree or PhD. On average, they started 1.74 companies (SD = 1.04; Median = 1) and have 10.68 employees (SD = 7.95). Our entrepreneurs were also rated by a total of 154 of their employees (average 2.91 per employer). Most of the entrepreneurs were rated by three of their employees, but the number of employees that rated each entrepreneur ranged from one to seven (see Appendix B for detailed description of the procedure).

Instruments

Instruments that were the same as in the previous studies were GDMS (responded by entrepreneurs), in-role performance and counterproductive work behavior (both responded by entrepreneurs' employees). We describe only the new instruments here. In order to avoid any confusion, we put in brackets data source.

Need for achievement (entrepreneurs). We measured need for achievement (NfA) using the Unified motive scale (UMS) developed by Schönbrodt and Gerstenberg (2012). Achievement motivation is assessed with items where the participants rated importance of achievement-related goals on a six-point scale (1 = not important to me, 2 = of little importance to me, 3 = of some importance to me, 4 = important to me, 5 = very important to me, 6 = extremely important to me). Total score is calculated as an average of importance ratings on all the items.

Perceived entrepreneurial efficacy (employees). We revised a self-report scale of entrepreneurial self-efficacy from Slabbinck et al. (2018) by reframing the items to be appropriate for employee ratings. The efficacy of entrepreneur was measured in terms of his/her employees' confidence in the ability of entrepreneur to perform critical entrepreneurial tasks. Employees rated their employers on a seven-point scale (1 = much worse than other entrepreneurs, 2 = worse than other entrepreneurs, 3 = somewhat worse than other entrepreneurs, 4 = about same as other entrepreneurs, 5 = somewhat better than other entrepreneurs, 6 = better than other entrepreneurs, 7 = much better than other entrepreneurs). Total score was calculated as an average level of ratings on the nine items and the higher the score, the better the perception of participant's efficacy in the entrepreneurial role.

Perceived organizational support (employees). Perceived organizational support represents set of global beliefs concerning the extent to which an organization values contribution of its employees and cares about their well-being. We measured it with eight items taken from the Eisenberger et al. (1986) scale, where participants indicated their level of agreement with the statements on a seven-point scale (1 = completely disagree, 2 = disagree, 3 = somewhat disagree, 4 = neither agree nor disagree, 5 = somewhat agree, 6 = agree, 7 = completely agree). We calculated total score as an average level of agreement on these eight items.

Job satisfaction (employees). We measured job satisfaction with five items taken from longer Index of Job Satisfaction measure (Brayfield & Rothe, 1951). Participants rated their level of agreement with five statements on a five-point scale (1 = completely disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, 5 = completely agree), and total score represents an average of their ratings on these five items.

Intention of leaving organization (employees). We measured this with three items assessing turnover intentions, adopted from Konovsky and Cropanzano, (1991). Participants rated their level of agreement with these statements on a five-point scale (1 = completely disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, 5 = completely agree), and total score was calculated as average of their ratings.

Results

We report descriptive statistics and reliabilities of our Study 3 focal variables in Table 23 and correlations among them in Table 24. This time we had no missing values, so each analysis was done on a full sample of N = 53.

Table 23. Descriptive statistics and reliabilities of Study 3 focal variables

	M	SD	Min	Max	α
Rational	4.48	0.53	2.80	5	.72
Intuitive	3.51	0.79	1.40	5	.79
Avoidant	2.00	0.86	1	4.60	.81
Spontaneous	2.46	0.73	1	4.40	.67
Dependent	3.50	0.74	2	5	.69
NfA	4.81	0.71	2.67	6	.81
Efficacy	5.02	1.04	1.75	7	.91
In-role	4.53	0.28	3.67	5	.72
CWB	1.82	0.49	1	2.93	.78
Perc. org. supp.	5.51	1.14	2.44	6.92	.96
Job sat.	3.97	0.62	2.20	5	.86
Leaving int.	2.03	0.86	1	4.33	.93

Note.

α = Cronbach alpha; NfA = Need for achievement; In-role = In-role job performance; CWB = Counterproductive work behavior; Perc. org. supp. = Perceived organizational support; Job sat. = Job satisfaction; Leaving int. = Intention of leaving organization.

A specific sample such as entrepreneurs can serve as a valuable additional source for validation of decision-making styles. For example, because entrepreneurs are characterized by higher risk tolerance, proactiveness in seeking and utilizing new opportunities and higher need for autonomy (Ahmetoglu, 2015; Frese & Gielnik, 2014; Rauch & Frese, 2007) we can expect that they will be less avoidant in their decision-making and less willing to rely on others when making decisions than population of non-entrepreneurs. We therefore, as a preliminary analysis, conducted five independent samples t-tests to examine the difference in the five decision-making styles between the entrepreneurs from our Study 3 and a more general population "ordinary" employees from our Study 2.

T-tests showed that there were significant differences in rational style, with the entrepreneurs scoring somewhat higher than the general population ($t = 2.02$, $p = .04$, Cohen's $d = 0.31$), dependent style, with the entrepreneurs scoring lower than the general population ($t = 3.18$, $p = .002$, Cohen's $d = 0.48$) and avoidant style, with the entrepreneurs again scoring lower than the general population ($t = 2.42$, $p = .02$,

Cohen's d = 0.39). There were no significant differences in intuitive and spontaneous decision-making styles.

Table 24. Pearson's correlations between Study 3 focal variables

	Sex	Rat	Int	Avo	Spon	Dep	NfA	Eff	In-role	CWB	POS	Jsat	LInt
Age	.07	.20	.13	.03	-.05	-.19	-.13	-.27^{+}	.17	.06	-.21	-.14	.09
Sex		.22	.06	-.18	-.29*	.28*	.06	.07	-.13	-.08	.00	.15	-.11
Rat			-.10	-.10	-.17	.01	-.02	-.21	-.10	.12	-.13	.05	.02
Int				.07	.09	.24^{+}	.12	.06	-.08	-.15	.14	.00	.06
Avo					.23^{+}	.25^{+}	-.10	-.19	-.18	.41**	-.24^{+}	-.24^{+}	.26^{+}
Spon						.02	.22	.19	.14	-.14	.10	.08	-.16
Dep							-.17	.09	-.05	.03	.06	.11	-.04
NfA								.09	-.05	-.09	.22	.18	-.10
Eff									.39**	-.32*	.72**	.69**	-.63**
In-role										-.44**	.46**	.45**	-.46**
CWB											-.47**	-.38**	.41**
POS												.82**	-.74**
Jsat													-.82**
Lint													

Note.

Rat = Rational decision-making style; Int = Intuitive decision-making style; Avo = Avoidant decision-making style; Spon = Spontaneous decision-making style; Dep = Dependent decision-making style; NfA = Need for achievement; Eff = perceived entrepreneurial efficacy; In-role = In-role job performance; CWB = Counterproductive work behavior; POS = Perceived organizational support; JSat = Job satisfaction; LInt = Intention of leaving organization.

$^{+}$ $p < .10$; * $p < .05$; ** $p < .01$ (two-tailed)

In Table 24, we flagged significant correlations between our variables. However, due to relatively small sample, we decided to increase our power by increasing a cut-off p-value for a significant effect from .05 to .10. Therefore, we treat all the effects with p-value lower than .10 as significant. We are aware that at low sample sizes correlation effects are still not stabilized and could therefore be far from true effect

(Schönbrodt & Perugini, 2013) and that this way we risk a greater number of potential false positives. However, we believe that in this case it was sensible to make small trade-off between Type 1 and Type 2 errors in order to be able to detect potential relationships between our variables in this largely explorative study.

Of all decision-making styles, avoidant style appears to be the most relevant for employee related outcomes. In our study it was positively correlated with employees' self-reports of counterproductive work behavior and intentions of leaving organization, and negatively with perceived organizational support and job satisfaction. Conversely, need for achievement, as one of the most relevant personality factors for entrepreneurial success did not seem to matter as much for measured outcomes.

General discussion

Across the three studies and the three different types of samples, we aimed to investigate the predictive validity of decision-making styles for a wide range of significant real-life outcomes, with a focus on work-related criteria. In short, our findings confirmed that the way individuals typically make decisions is related to a range of real-life outcomes. Generally, avoidant decision-making was the most relevant of the five styles for the outcomes we measured. Across the three samples, it was consistently related to worse outcomes, including generally experiencing more negative decision outcomes, lower job and career satisfaction, lower quality of academic/in-role job performance, more counterproductive academic/work behavior, and worse performance in entrepreneurial role. It also differentiated between a sample of entrepreneurs and a sample of "ordinary" employees. Spontaneous decision-making style showed a similar pattern of negative relationships with positive outcomes and positive with negative outcomes, although not within a sample of entrepreneurs. The other styles were also correlated with some of the outcomes, but they seem to be somewhat less relevant for our outcomes than the avoidant and spontaneous styles. Our studies again point to the importance of studying a broader range of decision-making styles, not only ones related to cognitive processes (i.e. rational and intuitive), especially in the workplace context.

Regarding the question of incremental validity, our findings revealed that decision-making styles were largely unrelated to cognitive abilities, except for intuitive style that was somewhat negatively correlated with numerical abilities in the first two studies, paralleling Juanchich et al., 2016. Thus, to the extent that the styles predict any of the outcomes, we can be quite sure that they do it independently of cognitive abilities such as numeracy and intelligence. The picture is not so clear regarding the incremental validity

above personality traits. Specifically, our Study 2 showed that the effects of decision-making styles were frequently decreased after accounting for the effects of the Big 5 personality traits on the outcomes. However, there were still instances where some of the styles remained significant and important predictors, even when personality traits were accounted for. This was especially true in the case of in-role performance where avoidant decision-making style was one of the most important predictors and in the case of decision-making quality where rational decision-making style was by far the most important predictor. This latter result practically replicates the findings about importance of rational style for peer perceptions of decision-making quality (Wood, 2012).

However, the question that begs the answer here is why the effects of avoidant style, that seem to be the most important when bivariate relationships are considered, shrink once the effects of personality traits are accounted for. We believe that the main reason for this is that decision-making styles are in part a product of personality. Specifically, as we explained in the introduction, avoidant decision-making style is posited to follow from negative affects and emotions, such as anxiety (Dewberry et al., 2013b). The same traits are captured in part with the emotional stability scale that was the most potent personality predictor of outcomes in all our studies. The substantial correlation between avoidant style and emotional stability mirrors the relationships of indecisiveness and avoidant style with neuroticism found earlier (e.g. Dewberry et al., 2013a; Germeijs & Verschueren, 2011; Juanchich et al., 2016; Ülgen et al., 2016). Therefore, avoidant decision-making and emotional stability are probably correlated with our outcomes partly because they capture similar traits, making the avoidant scale less able to contribute by explaining additional variance in our analyses. Looking from this angle, it could perhaps be said that decision-making styles could be seen as proximal determinants of outcomes, partly arising from personality traits that could be seen as the distal determinants of the same outcome.

In addition to students and employees, avoidant decision-making style seems also to be undesirable for entrepreneurs. In our Study 3, entrepreneurs' avoidant decision-making style was related with employees' counterproductive work behavior and intentions to leave organizations, as well as employees' job satisfaction and perceptions of organizational support. This finding complements findings from the leadership literature which showed that "laissez-faire" style leadership that is characterized by avoidance or absence of leadership is correlated negatively with leadership outcomes such as follower job satisfaction, follower motivation and group/organization performance (cf., Judge & Piccolo, 2004). Thus, it seems that, from the employee perspective, one of the worst things an entrepreneur can do is being indecisive.

Our study is not without limitations. While the diversity of samples is the strong point of our manuscript, we cannot disregard the small sample size of our third study. Therefore, we were careful not to draw any strong and definitive conclusion from its results. However, we felt that it was more worthwhile to publish those results while being mindful about the limitations of conclusions and generalizability of the results, thus allowing others to build on our results by e.g. conducting replication/extension studies or meta-analyses. Despite these limitations, we still believe that our research program fills the gap about individual differences in decision-making, offering evidence that decision-making styles matter for various important real-life outcomes.

6. STUDY 5: TESTING THE THEORY OF GOOD THINKING AND DECIDING IN ORGANIZATIONAL SETTING - MANY BENEFITS OF LEADER'S ACTIVELY OPEN-MINDED THINKING

This chapter was previously published as: Erceg, N., Galić, Z, & Buljan Šiber, A. (2023). Testing the Theory of Good Thinking and Deciding in Organizational Setting: Many Benefits of Leader's Actively Open-minded Thinking. *Studia Psychologica*.

Introduction

Decision-making is one of the core things a manager does and one of the core skills he or she must possess. Most of the competency-based models of managerial work puts decision-making at the forefront of the managerial duties (e.g., Bartram, 2005; Dierdorff & Rubin, 2006; Tett et al., 2000). For example, Bartram (2005) lists deciding and making judgments among the great eight managerial competencies and Tett et al. (2000) conclude that decision-making is the common core of all twelve models of leadership competencies that they reviewed.

Yet, research indicate that managers are bad at decision-making. For example, Nutt (2002), who studied 400 decisions made by top managers over twenty years came to a startling conclusion that decisions made by top managers fail half of the time, although this by itself does not necessarily mean that the decision-making process was bad. Lovallo and Sibony (2010) describe the McKinsey survey of 2,207 executives of which only 28 percent said that the quality of strategic decisions in their companies was generally good, 60 percent thought that bad decisions were about as frequent as the good ones, and the remaining 12 percent thought good decisions were altogether infrequent.

In addition to establishing that managers are not particularly good decision-makers, studies have tried to uncover the reasons behind the good vs. bad decisions. First, it seems that, when making decisions, managers often rush to conclusions without searching for and considering wide enough array of possibilities or evidence, leading to mistakes such as premature commitment to an idea (Nutt, 2002), relying on the limited set of assumptions (Ketchen & Craighead, 2022), or anchoring to the first piece of information and failing to adjust one's position subsequently (Ketchen & Craighead, 2022; Sibony, 2020). Second problem seems to be selective search and interpretation of evidence. Managers, as other people, have tendency to search for and overweight evidence that is in line with their current, favorite position or idea, while simultaneously avoiding and downplaying evidence that counters it. This can lead to several serious mistakes in decision-making such as: a) the escalation of commitment to the

current idea even when evidence against it appears (Ketchen & Craighead, 2022; Sibony, 2020); b) constructing a coherent story from a selection of facts fueled by tendency to attend only to information that confirm the current idea/position and ignore or discount information that contradict it (i.e. confirmation bias; Kahneman et al., 2011; Ketchen & Craighead, 2022; Sibony, 2020); c) the groupthink trap, as this tendency to shield oneself from counterevidence can also lead to surrounded themselves with likeminded people or those who are afraid to speak against bosses' idea (Sibony, 2020). Finally, the third problem is the overconfidence in one's own conclusions and decisions (Ketchen & Craighead, 2022, Sibony, 2020). If a person only attends to information that confirms his/her initial position, without ever questioning it, this will lead to accumulation of one-sided arguments and to bolstered confidence in one's conclusions.

Actively open-minded thinking as an antidote to the managerial mistakes

Given that it can be argued that these three problems underpin majority of managerial decision failures, it is remarkable that the concept of actively open-minded thinking (AOT; Baron, 2000; 2019; Baron et al., 2015) is practically still non-existent in the managerial literature. According to Baron (2000) who developed this theory, AOT describes what a good thinking should look like, and it consists of three things:

a) a search of information that is sufficient and thorough in proportion to the importance of the question
b) active search for and fair treatment of possibilities other than the one decision-maker initially favors
c) confidence that is appropriate for the amount and quality (direction) of thinking done.

From this definition of AOT, it is immediately clear that this kind of thinking is the direct antidote to the three mistakes in managers' thinking that underpin majority of bad strategic decisions. This is not surprising, as the AOT was developed precisely to be a "prescriptive" theory of rationality, i.e. to prescribe how people should think and make judgments in order to counteract the most prevalent and serious cognitive biases that trump quality decision-making. Our goal within this study is, thus, to test this theory in organizational setting, i.e., to test the benefits of managers' AOT for employee level outcomes.

Empirically documented benefits of actively open-minded thinking

Outside the organizational context, there is plenty of evidence for the beneficial effects of AOT on beliefs, judgments, and decision-making. For starters, evidence suggest that AOT correlates negatively with a wide range of the usual cognitive biases identified in human decision-making, such as

confirmation bias, sunk cost effect, outcome bias, belief bias and others (Erceg et al., 2022; Stanovich & West, 1997; Stanovich et al., 2016; Toplak et al., 2014). Additionally, people higher on AOT are less prone to holding epistemically suspect beliefs such as conspiracy, superstitious or paranormal beliefs (Erceg et al., 2022; Pennycook et al., 2020; Svedholm & Lindeman, 2013; Svedholm-Häkkinen & Lindeman, 2018), are more accurate at a variety of judgments, such as distinguishing between good and bad arguments (Stanovich & West, 1997), forecasting world events (Mellers et al., 2015) or distinguishing between real and fake news (Bronstein, et al., 2019).

Possible beneficial effects of actively open-minded thinking within organizational environment

We also believe that managers that are high on AOT bring about additional beneficial outcomes relevant for the workplaces. It is reasonable to expect that managers that are open to and actively search for opposing information and evidence want to hear what others think and have to say and include them in the decision-making process. In organizational literature, such characteristic is labeled as manager's humility and refers to manifested willingness to view oneself accurately, a displayed appreciation of others' strengths and contributions, and teachability (Owens et al., 2013). This managerial characteristic has been shown to be highly beneficial at the individual (e.g. enhanced trust in the leader, work engagement and job satisfaction, enhanced follower creativity), team (e.g. increased team performance, enhanced information sharing) and organizational level (e.g. lower turnover, higher firm performance; Davis et al., 2016; Ou et al., 2018; Owens et al., 2013; Swain & Murray, 2020). Additionally, it seems that humble leaders tend to create a climate of psychological safety (Swain, 2018; Wang et al., 2018) which refers to the shared belief held by members of a team that the team is safe for interpersonal risk taking, i.e. that no one will be reprimanded or ridiculed for stating their opinion, questioning and disagreeing with others or noticing mistakes (Edmondson, 1999; 2018). Psychological safety has significant benefits for organizations and employees that again include increase in employee job performance, engagement and creativity, enhanced team learning behavior and performance, reduction in errors, increase in organizational commitment and perceived organizational support (see Newman et al., [2017] and Edmondson & Lei [2014] for review, and Frazier et al., [2017] for meta-analysis).

Current study

Therefore, the aim of this study is to examine the role of manager's AOT in positive employee level outcomes. Specifically, across two studies we were interested to see how managers' AOT was related to the subordinates' perceptions of the decision-making quality and intellectual humility of their

managers, as well as to the subordinates' work attitudes such as job satisfaction and perceived organizational support. We planned for two studies as we were also interested to see how stable our eventual effects are, i.e. whether or not they will replicate on two independent samples and using different measures of the target constructs. In Study 2, in addition to the aforementioned variables, we also captured psychological safety as an important and positive team outcome that could be positively influenced by manager's AOT. We advanced the hypotheses that managers' AOT will be positively correlated with all the employee level outcome variables. Specifically, managers' AOT will be positively correlated with:

H1: subordinate perceptions of their superiors' decision-making quality,
H2: subordinate perceptions of their superiors' intellectual humility,
H3: subordinate ratings of their own job satisfaction,
H4: subordinate ratings of perceived organizational support, and
H5: subordinate ratings of psychological safety in their teams.

In addition to this, we wanted to see whether AOT helps explain these important outcomes above the effects of the Big Five personality factors as one of the most important individual difference characteristics explaining both leadership and work outcomes (Judge et al., 2009; Schmitt, 2014). To investigate the incremental validity of AOT above the Big Five factors, we joined the two samples to increase the statistical power.

The main reason we wanted to conduct this incremental validity analysis relates to the practical validity of AOT for the purpose of selection for managerial positions. Big Five personality traits are routinely used in such selection processes so it makes sense to check whether practitioners can gain additional predictive power by also measuring AOT in such situations (although from the conceptual perspective this might mean parsing out the valid variance in AOT). However, incremental validity analyses are problematic in several ways, mainly due to the imperfect reliabilities of the measures. One of the possible solutions is to conduct regression analysis using structural equation modelling (SEM) on latent variables that are free from measurement error (Westfall & Yarkoni, 2016), which is what we did.

Study 1

Methods

Procedure

We instructed psychology students, in exchange for extra course credits, to recruit for the study participants who were employed as managers and had at least three subordinates. The participants were informed that they will participate in a study on managerial competencies/leadership skills and were motivated to take part in the study with feedback about their leadership potential and a gift card valuable about 7$. In addition to the variables that we describe in this study, we collected additional data so that the completion of the survey took about one hour. Upon completing their own survey, managers were asked to forward the link to another survey for subordinates, along with their code that they generated and that we used to match the responses, to their subordinates. The subordinates' survey was substantially shorter than managers' and lasted about 10 minutes.

Participants

Both managers and their subordinates participated in our study. Overall, 124 managers participated in our study (49% males and 51% females) with the mean of age M = 45.16 (SD = 10.47) and mean years of experience in managerial role of M = 12.68 (SD = 9.04). On average, managers had M = 47.84 subordinates (Min = 3, Max = 1400). Education wise, our managers mostly had college degree (65%), but there were also some with only high school (27%) as well as those with PhD (8%). They were mostly employed in private sector (83%), but some also worked in state-owned company (13%) or public institution (4%). Finally, our managers mostly work in small companies with less than 50 employees (40%), followed by big companies with more than 500 employees (36%) and then middle companies with 50 to 500 employees (24%).

Not all of the managers were rated by their subordinates – we managed to obtain subordinate ratings for 95 of the managers, meaning that for 95 of the managers we were able to connect managers responses with subordinate responses. Majority of those 95 managers were rated by two subordinates (81%), some were rated by three subordinates (6%), some by four (2%), and some only had one subordinate rating (11%). In total, 190 subordinates participated in this study. For majority of managers who received more than one peer-rating, we averaged those ratings prior to conducting the analyses. All the subsequent analyses were always done on the largest possible sample, meaning that the descriptive statistics for the managers' self-ratings were done on a larger sample (N = 124) than the descriptive statistics for the subordinate-ratings and the correlations between the self- and subordinate-ratings (N = 95).

Instruments

Managers

Actively open-minded thinking. AOT was measured with a 10-item questionnaire (recommended at the time by the Society for Judgment and Decision-making; https://sjdm.org/) where participants rated their level of agreement with the statements (e.g. "People should take into consideration evidence that goes against conclusions they favor" or "Changing your mind is a sign of weakness" [reverse-coded]) on a 5-point scale (1 = completely disagree, 5 = completely agree), and the final score was calculated by averaging these ratings[12].

Cognitive ability. We measured cognitive ability with a 12 items version of the International Cognitive Ability Resource (ICAR; for details see icar-project.com and Condon and Revelle, 2014). ICAR is a cognitive ability assessment tool consisting of four different types of tasks: letters and numbers series, matrix reasoning items, verbal reasoning items and three-dimensional rotation items (not used in our study). The validation of this measure is reported in Condon and Revelle (2014).

Mini International Personality Item Pool questionnaire (Mini IPIP). Mini IPIP (Donnellan et al., 2006) is a 20-item personality measure, measuring the Big 5 traits each with four items. Participants were instructed to rate the accuracy of the description (e.g. "Am the life of the party.", "Sympathize with others' feelings.", "Get chores done right away.", "Have frequent mood swings.", "Have a vivid imagination.") on a 7-point scale (1 = Completely incorrect 7 = Completely correct).

Subordinates

Manager's Decision-making quality scale (Wood, 2012). This scale consisted of eight items assessing subordinates' perceptions of their managers' decision-making quality and success. Subordinates were instructed to rate their agreement with the statements describing their manager (e.g. "The decisions my superior makes follow reason and logic" or "The decisions my superior makes end up working out well") on a five-point scale (1 = completely disagree, 5 = completely agree).

Manager's Intellectual humility. Intellectual humility was measured with Expressed humility scale (Owens et al., 2013) consisting of nine items (e.g. „This person actively seeks feedback even if it is critical", or "This person acknowledges when others have more knowledge and skills than

[12] If not written otherwise, for each of the variables in the study, we calculated the final score by averaging ratings or correct responses.

him/herself"). The participants rated the degree to which each of the item describes their manager on a five-point scale (1 = completely disagree, 5 = completely agree).

Job satisfaction. We measured subordinates' job satisfaction with a single item (a 5-point scale, ranging from 1 = very unsatisfied to 5 = very satisfied) assessing their satisfaction with their job, overall.

Perceived organizational support (Eisenberger et al., 1986) was measured with eight statements (e.g. „This organization really cares about my well-being", or "The organization does not value my extra effort") for which participants indicated their level of agreement on a seven-point scale (1 = completely disagree, 7 = completely agree).

Results

Prior to conducting main analyses, we calculated the descriptive statistics of our focal variables. This is presented in the Table 25.

Table 25. Descriptive statistics of Study 1 focal variables

Variable	Mean	SD	Min	Max	Theoretical range	ω_t
Managers						
AOT	3.75	0.44	2.10	4.80	1 – 5	.76
ICAR	7.70	3.24	0	12	0 – 12	.87
Openness	4.77	1.24	1.75	7.00	1 – 7	.81
Conscientiousness	5.24	1.20	1.75	7.00	1 – 7	.86
Extraversion	4.95	1.22	1.50	7.00	1 – 7	.84
Agreeableness	5.52	0.92	1.75	7.00	1 – 7	.75
Neuroticism	3.30	1.07	1.25	6.00	1 – 7	.87
Subordinates						
DMQ	4.21	0.58	1.63	5.00	1 – 5	.95
IH	4.26	0.73	1.22	5.00	1 – 5	.98
JS	4.07	0.59	2.50	5.00	1 – 5	/
POS	4.71	1.00	2.63	7.00	1 – 7	.87

Note. ωt = Omega total reliability; AOT = Actively Open-minded thinking; ICAR = International Cognitive Ability Resource; DMQ = Manager's Decision-making quality; IH = Manager's Intellectual humility; JS = Job satisfaction; POS = Perceived Organizational Support.

On average, it seems that managers mostly agreed with statements describing AOT as a good standard of thinking and reasoning. Another noticeable result are high ratings on decision-making quality and intellectual humility – it seems that majority of subordinates see their superiors as intellectually humble, good decision makers. On general, employees/subordinates also seem to be very satisfied with their jobs and feeling somewhat supported by their organization. To investigate the relationships between subordinate-rated outcomes and AOT, we calculated and present bivariate correlations among our variables in Table 26. We have conducted one-tailed tests to obtain the degrees of significance as our main hypothesis was directional (i.e., we predicted that AOT would be positively correlated with the outcomes).

Table 26. Bivariate correlations among Study 1 variables (raw correlations are shown above the diagonal and disattenuated correlations are shown below the diagonal)

	AOT	ICAR	Open.	Cons.	Extra.	Agree.	Neuro.	DMQ	IH	JS	POS
Managers											
AOT	1	.24**	.19*	-.01	-.07	.25**	-.09	.25**	.28**	.15	.12
ICAR	.29	1	-.04	-.12	-.13	.02	-.14	.21*	.13	.10	.08
Open.	.24	-.05	1	.03	.21*	.31**	-.18*	.05	.11	-.20*	-.17
Cons.	-.01	-.14	.04	1	.05	.10	-.21**	.09	.05	-.09	.02
Extra.	-.09	-.15	.25	.06	1	.38**	-.28**	-.01	-.04	-.03	.09
Agree.	.33	.02	.40	.12	.48	1	-.09	.22*	.18*	.20*	.27**
Neuro.	-.11	-.16	-.21	-.24	-.33	-.11	1	.00	.01	-.04	-.12
Subordinates											
DMQ	.29	.22	.06	.10	-.01	.26	.00	1	.72**	.17*	.42**
IH	.32	.14	.11	.05	-.04	.21	.01	.75	1	.18*	.44**
JS	.18	.11	-.22	-.10	-.03	.23	-.04	.17	.18	1	.56**
POS	.15	.08	-.20	.02	.11	.33	-.14	.46	.48	.60	1

Note. ** $p < .01$; * $p < .05$ one-sided
AOT = Actively open-minded thinking; ICAR = International cognitive ability resource; Open. = Openness; Cons. = Conscientiousness; Extra. = Extraversion; Agree. = Agreeableness; Neuro. = Neuroticism; DMQ = Manager's decision-making quality; IH = Manager's intellectual humility; JS = Job satisfaction; POS = Perceived organizational support.

Table 26 shows that manager's AOT is positively related to the subordinate's ratings of their decision-making quality and intellectual humility. This means that managers who agree with AOT as a standard of good thinking are perceived by their subordinates to be better decision makers and more intellectually humble, suggesting that adherence to beliefs about good standards of thinking reflects in managers' observable behaviors. Perceived manager's decision-making quality and intellectual

humility were, in response, positively related with subordinates' perceived organizational support. However, manager's AOT was not bivariately related with either of the subordinates' work attitudes, job satisfaction or perceived organizational support. In addition to this, manager's AOT was positively related to his/her cognitive ability, openness and agreeableness.

Besides being related with manager's AOT, perceived manager's decision-making quality was also positively related with his/her cognitive ability and agreeableness. Except AOT, only cognitive ability and two of the five personality factors, openness and agreeableness, showed some meaningful correlations with the measured outcomes. Other personality factors were basically unrelated with any of the outcomes. In order to replicate and extend our results, we conducted Study 2 where we used some measures that were similar to ones in this study, but also some additional measures (as explained below).

Study 2

Methods

Procedure

In Study 2, we collected data in two ways. We approached several companies and asked them to collaborate with us on a study about leadership competencies. Companies that agreed to participate invited their managers to participate in a 3-hour workshop that within the first part included testing and in the second part a lecture on "state-of-the-art" knowledge on leadership skills. Second way was identical to Study 1 – we instructed several psychology students to recruit managers that had at least three subordinates to participate in the study.

All managers were motivated to participate in the study with the promise of personalized feedback and a gift card voucher (around 13$). The managers completed a battery of tests and questionnaires not all of which are reported in this study. It took them on average around hour and half to complete this battery. Managers were asked to provide us with email addresses of five of their subordinates who were then sent the link to a much shorter questionnaire (10 to 15 minutes long) in which they were asked to rate their manager, as well as to rate their own work attitudes (e.g. job satisfaction and perceived organizational support) and their team psychological safety. Subordinates completed several more questionnaires that we do not report here.

Participants

A total of 126 managers and 335 of their subordinates participated in this study. There were 78 (62%) male and 48 (38%) female managers with mean age M = 42.08 (SD = 7.53), mean years of work experience M = 18.15 (SD = 7.36) and mean number of subordinates of M = 25.44 (Min = 3, Max = 386). Again, they mostly had college degree (66%), but there were also some with only high school (22%) as well as those with PhD (12%). Almost all of our managers worked in a private sector (94%), with minority coming from public sector or NGOs. Regarding the company size, majority of managers worked in companies with 50 to 500 employees (42%), followed by large companies with more than 500 employees (38%) and smaller companies with less than 50 employees (20%).

Out of 126 managers participating in the study, we managed to obtain at least one rating for 108 of them. 19% of managers were rated by five, 21% by four, 22% by three, 23% by two and 15% by only one subordinate. This time, a total of 335 subordinates participated in the study. As in the Study 1, we averaged all of the subordinates' ratings for each of the managers before conducting further analyses.

Instruments

In Study 2, we tried to constructively replicate Study 1 findings (Lykken, 1968). Thus, some of the measures we used were the same as in the Study 1, but we also captured some additional constructs and for some constructs we deliberately used different measures. There is no single criterion for successful replication. At minimum, one can look at three things: whether the effects are in the same direction, whether both effects are statistically significant and whether the original effects fall within 95% CI of the replication effects (e.g. Open Science collaboration, 2015). We will use these three criteria to evaluate whether our Study 2 managed to replicate Study 1 results. Here, we describe instruments that were new or different in comparison to Study 1.

Managers

AOT. This time we measured AOT with currently recommended 11-item scale available at https://sjdm.org/dmidi/Actively_Open-Minded_Thinking_Beliefs.html and scoring was similar as before – calculating average of the participants' ratings on a 5-point scale (1 = completely disagree, 5 = completely agree).

Cognitive reflection test (CRT). As a measure of cognitive ability, this time we used the cognitive reflection test (CRT; Frederick, 2005). For this study, we developed our three items that were

completely similar to the original three in structure, but differed in content as we wanted to make them more work related and face-valid (e.g. “You are a manager in an auto equipment factory. If 5 machines make 5 car parts in 5 minutes, how many minutes would it take for 100 machines to make 100 car parts?”; intuitive response = 100, correct response = 5).

IPIP 50. This time we also decided to measure personality traits with more items than before to get a more reliable measure. To do that we used a Croatian translation of IPIP50 (Mlačić & Šakić, 2008) which captures each of the five factors with ten items. Participants rated their levels of agreement with the statements on a 5-point scale (1 = completely incorrect, 5 = completely correct).
Subordinates

Manager’s AOT perceptions. This time, in addition to measuring managers' AOT, we also decided to measure their subordinates' perceptions about whether their superiors tend to think in an AOT way. We did this specifically to see if managers' AOT thinking and behavior is observable by their subordinates. The subordinates rated two statements about their superiors that we developed and that reflected the core of AOT ("My superior looks for arguments and information that could be contrary to his/her existing views and initial decisions." and "My superior changes his/her opinion if the circumstances change, that is, if there are good arguments for the change.") using a seven-point scale (1 = completely disagree, 7 = completely agree).

Manager’s Decision-making quality. This time we opted for a similar, but a bit different scale used by Wood and Highhouse (2014). Instead of eight, this scale consists of five statements on which, again, subordinates rated their managers on a 5-point scale (1 = completely disagree, 5 = completely agree).
Psychological safety. We measured psychological safety with a 7-items scale developed by Edmondson (1999). Participants rated the statements (“If you make a mistake on this team, it is often held against you” and “Members of this team are able to bring up problems and tough issues”) on a 7-point scale ranging from 1 = “completely incorrect” to 7 = “completely correct.”

The variables of manager’s intellectual humility, job satisfaction and perceived organizational support were measured with identical items as those in Study 1.

Results

Again, before moving to main analyses, we have calculated the descriptive statistics of our variables. This is shown in Table 27.

Table 27. Descriptive statistics of Study 2 focal variables

Variable	Mean	SD	Min	Max	Theoretical range	ωt
Managers						
AOT	3.92	0.42	2.83	4.73	1 - 5	.71
CRT	0.60	0.36	0	1	0 - 1	.70
Openness	3.87	0.46	2.70	4.90	1 - 5	.81
Conscientiousness	4.08	0.58	2.10	5	1 - 5	.89
Extraversion	3.75	0.63	2.10	5	1 - 5	.89
Agreeableness	4.10	0.45	2.70	5	1 - 5	.83
Neuroticism	2.17	0.62	1	4	1 - 5	.90
Subordinates						
AOT p.	5.52	0.83	2.50	7	1 - 7	/
DMQ	4.26	0.46	2.95	5	1 - 5	.91
IH	4.35	0.52	2.44	5	1 - 5	.96
JS	4.09	0.46	2.80	5	1 – 5	/
POS	5.31	0.88	2.92	7	1 - 7	.96
PS	5.83	0.65	3.90	6.86	1 - 7	.86

Note. ωt = Omega total reliability; AOT = Actively open-minded thinking; CRT = Cognitive reflection test; AOT p. = Manager's AOT perceptions; DMQ = Manager's decision-making quality; IH = Manager's intellectual humility; JS = Job satisfaction; POS = Perceived organizational support; PS = Psychological safety.

In general, Study 2 managers mostly agreed with the AOT principles, and their subordinates rated them as on average intellectually humble, good decision-makers. Subordinates were also relatively satisfied with their jobs, perceived their team climate to be quite safe and did not plan to leave the organization in foreseeable time. To examine the relationships between AOT and other relevant variables and outcomes, we have computed bivariate correlations between our variables. Again, to obtain the degrees of significance we have conducted one-tailed tests. We are reporting these correlations in Table 28.

Table 28. Correlations among Study 2 variables (raw correlations are shown above the diagonal and disattenuated correlations are shown below the diagonal)

	AOT	CRT	Open	Consc	Extra	Agree	Neuro	AOT p.	DMQ	IH	JS	POS	PS
Managers													
AOT	1	.18*	.21**	-.19*	.08	.13	-.05	.36**	.18*	.22*	.18*	.09	.21*
CRT	.24	1	.21*	.01	.02	.05	.02	.09	.11	.12	.09	-.03	.31**
Open	.28	.27	1	.19*	.38**	.19*	-.23**	.04	.04	-.01	.17*	.08	.02
Consc	-.24	.01	.22	1	.06	.20*	-.35**	.02	-.03	-.10	-.08	-.08	-.09
Extra	.10	.03	.45	.07	1	.38**	-.33**	.02	.00	-.03	-.02	.05	.09
Agree	.17	.07	.23	.23	.44	1	-.21*	.17*	.13	.21*	.04	.17*	.23**
Neuro	-.06	.03	-.27	-.39	-.37	-.24	1	-.05	-.05	.06	-.02	.08	-.07
Subordinates													
AOT p.	.43	.11	.04	.02	.02	.19	-.05	1	.42**	.61**	.34**	.41**	.41**
DMQ	.22	.14	.05	-.03	.00	.15	-.06	.44	1	.73**	.42**	.53**	.51**
IH	.27	.15	-.01	-.11	-.03	.23	.06	.62	.78	1	.43**	.51**	.47**
JS	.21	.11	.19	-.08	-.02	.04	-.02	.34	.44	.44	1	.61**	.33**
POS	.10	-.04	.09	-.09	.05	.19	.09	.41	.57	.53	.62	1	.43**
PS	.27	.39	.02	-.10	.10	.27	-.08	.44	.58	.52	.34	.47	1

Note. ** $p < .01$, * $p < .05$ one-sided

AOT = Actively open-minded thinking; CRT = Cognitive reflection test; Open = Openness; Consc = Conscientiousness; Extra = Extraversion; Agree = Agreeableness; Neuro = Neuroticism; AOT p. = Perceptions of manager's AOT; DMQ = Manager's decision-making quality; IH = Manager's intellectual humility; JS = Job satisfaction; POS = Perceived organizational support; PS = Psychological safety.

Table 28 shows that managers with higher AOT were perceived as being better decision makers, more intellectually humble and their team climate was perceived as being more psychologically safe compared to managers with lower AOT. Additionally, managers' agreement with AOT standards obviously reflected in their observable behavior judging from the relatively high positive correlation between managers' AOT and subordinates' AOT perceptions. Subordinates perceived organizational support again failed to significantly correlate with the managers' AOT, while the correlation between subordinates' job satisfaction and AOT was positive and significant this time.

Unlike AOT that was consistently related with most of the outcomes in the expected directions and to expected degree, cognitive abilities (CRT) and the Big Five factors exhibited only sporadic correlations with these outcomes. Specifically, CRT was significantly correlated only with psychological safety, indicating that managers with higher cognitive abilities tend to have teams that are more psychological safe, which is generally consistent with the relationship between intelligence and leadership outcomes (Judge et al., 2004). Of the five personality traits, only agreeableness and openness managed to correlate significantly with any of the outcomes – agreeableness with AOT perceptions, intellectual

humility, perceived organizational support and psychological safety, and openness with job satisfaction.

Analysis of AOT's incremental validity on joint sample

After joining two samples for the variables that were present in both studies (details of this procedure are in the Appendix B), we were left with four outcomes: subordinates' ratings of managers decision-making quality and intellectual humility, perceptions of subordinates' job satisfaction and perceived organizational support. The joint sample had between N = 214 and N = 250 cases, depending on the variable.

As we mentioned earlier, we conducted a SEM regression, regressing the outcomes on AOT and Big Five factors. We were interested in whether the beta ponder of AOT will remain statistically significant (one-sided tests due to directionality of hypotheses) after accounting for the effects of the Big Five factors. The results showed that the coefficient for AOT remained significant for each outcome variable except the subordinate rated decision-making quality, although the incremental variance explained by AOT was rather modest (0.018 for decision-making quality, 0.042 for intellectual humility, 0.022 for job satisfaction and 0.027 for perceived organizational support; detailed results are presented in the Appendix B).

Discussion

The guiding rationale behind our two studies was that one of the most important tasks that managers need to do is making decisions that have implications for their teams and companies. Often, these are high stake decisions that would benefit from specific decision-making skills that are, given the literature on the management decision-making failures (e.g., Nutt, 2002), often absent. In the decision-making field of study and literature, the concept of AOT has long theoretical and empirical history. Yet, to our surprise, concept of AOT as a central construct and one of the most important decision-making individual difference variables, is completely missing from the management research. With our studies, we are hoping to fill this gap and respond to several calls to bridge the gap between decision making and industrial/organizational (I/O) research traditions that were so far largely distant and disparate (cf., Dalal et al., 2010; Highouse et al., 2014).

With a fair degree of consistency across our two studies, managers' AOT was positively related with subordinates' ratings of their managers' quality of decision-making and intellectual humility, as well as

with perceived team psychological safety, support a subordinate feels he/she is getting from the organization and overall job satisfaction. Formally, looking at the three criteria for successful replication that we specified, Study 2 effects were always in the same direction as Study 1, and their 95% CI always included Study 1 effects. Regarding the statistical significance criterion, both effects of interest that were significant in Study 1 were also significant in Study 2 (AOT relationship with decision-making quality and intellectual humility ratings). Thus, we conclude that, overall, Study 2 successfully replicated and extended the results of Study 1, confirming that AOT is also relevant in the organizational context and that manager's AOT is related with a range of important outcomes. Importantly, for most of the outcomes, managers' AOT was still important even after controlling for the effects of personality traits. Granted, these effects were not large, but this is not surprising given the different sources of ratings and far from perfect indicators of the target constructs, but also the fact that the outcome variables are affected by various and different aspects that are often even outside managers' control (e.g., employees' personality).

Still, it is also worth noting that the AOT is not a classical self-report measure as participants do not rate how often they personally think or behave in a specific way, but only to what degree they agree that specific way of thinking and behaving represents a standard of good thinking and behaving. The rationale here is that people who believe a certain type of behavior is generally good and desired will more often behave in this way, but this gap between believing that something is good behavior and actually behaving in that way could also diminish the correlations between the AOT and other variables. To obtain effects of AOT that we and the previous studies obtained testifies, in our view, of the validity of the AOT measure and of the importance of the construct for various outcomes in different domains. In short, we believe that our studies provide evidence for the importance of AOT in the organizational settings and good arguments for paying more attention to this concept in future studies.

One surprising and a bit disappointing finding from our point of view was relatively low correlation between self-rated managers' AOT and their decision-making quality as rated by their subordinates. Indeed, this was the only outcome for which AOT did not exhibit statistically significant incremental validity above the effects of personality traits on out joint sample. However, in Study 2 we also measured whether subordinates were able to perceive their managers' AOT and obtained moderate correlations between the managers' AOT and subordinate perceptions of their AOT. This means that managers' AOT was something that was observable by other people in their surroundings. It also

implies that the process, rather than the outcome, of making a decision might be better criteria for assessing the quality of decision-making, as the process is under direct individual control while the outcome can be affected by many other things, outside one's control, including pure luck.

Besides positioning AOT as one of the crucial individual differences underpinning good decision-making, testifying for its benefits at a company level, our studies have shown that managers' AOT could have additional benefits at the individual and team levels. First, high AOT managers were consistently perceived as more intellectually humble meaning that their subordinates noticed that these managers value their opinions and advice, show appreciation for their contributions, and notice and praise their strengths. As noted in the introduction, studies have shown that this has many benefits for individuals, teams and organizations - higher work engagement and job satisfaction, enhanced creativity, increased team performance, lower turnover, higher firm performance, etc. (Davis et al., 2016; Ou et al., 2018; Owens et al., 2013; Swain & Murray, 2020). In addition, AOT managers tended to have teams that are more psychologically safe which also has its own benefits such as heightened job performance, engagement and creativity, enhanced team learning and reduction in errors (Newman et al., 2017; Edmondson & Lei, 2014; Frazier et al., 2017). In sum, it seems that acknowledging one's own limitations, paying attention to others' thoughts and arguments, soliciting and valuing their inputs and advice and changing your mind accordingly have many benefits at every level of the organizational structure, and AOT is thinking disposition that predisposes individuals to think and behave in this way more. Thus, the benefits of managers' AOT permeate organization through different channels and mechanisms not limited solely to enhanced quality of decision making.

Practical implications

One practical implication of our results is that it could be possible to teach managers to make better and more beneficial decisions for their organizations and its employees by teaching them what is AOT and how to think in actively open-minded way. Although there are individual differences in propensity towards AOT, as AOT represents the standards of good thinking, it is in principle teachable. Some of the previous studies showed some promise in this regard. For example, Perkins (2019) showed that it is possible to teach students to develop their arguments better, specifically by including other-side perspective, which is something that does not come naturally to people. Gurcay-Morris (2016) showed that a short, one hour long, AOT online module managed to increase other-side thinking and somewhat decrease overconfidence in one's own judgments. Thus, increasing managers' AOT holds promise as an avenue not only to enhanced quality of decision-making at the high levels of organization, but also

to increased psychological safety of the teams and engagement and satisfaction of employees, with all of the benefits these outcomes bring. This is definitely something worth pursuing in future scientific endeavors.

Limitations

The first limitation relates to AOT and outcome measures. The current AOT measure that assesses one's beliefs about proper standards of thinking is, in a way, a proxy for one's "true" tendency towards AOT. Although such measure has its advantages, it is also possible that there exists a gap between what someone believes and what someone does, possibly lowering the correlations between AOT and outcomes. Other approaches to measuring AOT in organizational contexts could be investigated in the future (for example, the AOT situational judgment test that is currently being developed by our research group; Vrhovnik, 2022).

The other limitation relates to the incremental validity of AOT over and above cognitive ability. In our analyses we have not included a measure of cognitive ability in our incremental validity analysis. The reason for this is simply that we used different ability measures across the studies and therefore could not merge them in a joint sample, and doing incremental analysis on separate samples would not make much sense due to low sample size. In other words, statistical strength of such analyses would be too low to warrant reliable conclusions, and SEM regression especially requires high sample size (Westfall & Yarkoni, 2016). However, as both ability measures (ICAR and CRT) exhibited smaller correlations with the outcomes than AOT, we suspect that the AOT would still be able to explain incremental variance had the ability measures been included in the analysis.

Conclusion

In conclusion, we have conducted two studies in which we demonstrated benefits of managers' AOT at many organizational levels. Specifically, we have showed that the higher the manager's AOT, the better their decision-making capabilities and the higher their intellectual humility as judged by their subordinates. Furthermore, managers with higher AOT tended to have teams that were more psychologically safe and employees that were more satisfied with their jobs and that felt more supported by their organization. We argue that the AOT is the disposition that predisposes some managers to patterns of thinking and behavior that are observable and highly valued by their subordinates, resulting in a range of beneficial outcomes. Thus, one promising way forward is to develop and test educational interventions aimed to teach AOT to highly positioned organizational

leaders with a bid to enhance the quality of decision making in the organization, but also to improve their employees' work-related attitudes, thus affecting different organizational outcomes in positive way.

7. GENERAL DISCUSSION

Although many implications that follow from these five studies were already discussed withing particular studies, I will focus here on three main insights related to assessment and measurement of individual differences in traits relevant for sound reasoning and decision-making. In short, the three main insights that follow from this line of work are:

a) Cognitive reflection test is not a good measure of cognitive reflection,

b) Measuring rational thinking and decision-making with cognitive bias tasks is messy,

c) Quality of the process of thinking and decision making is at least as important for the outcomes as are cognitive abilities.

Cognitive reflection test is not a good measure of cognitive reflection

CRT used to be a paradigmatic example of the dual-process theory in action, especially the default-interventionist paradigm. Because of the lures, our System 1 produces strong intuitions in response to CRT questions, and if we are to respond correctly, we need to intervene, overturn those faulty intuitions and calculate the right response. In this way, CRT was an ideal measure of our motivation to think, i.e. of thinking dispositions that make us careful and willing to question our intuitions, as well as willing to engage in prolonged and deeper processing.

However, results of our first two studies pose serious challenge to basically every aspect of this usual CRT story. First, it seems that lures that cue strong intuitive response actually do not make CRT any different from some other mathematical tasks that do not use lures to cue strong intuitive response (Study 1). Thus, it seems unlikely that overriding incorrect intuitive response is the key mechanism of responding on CRT tasks. Consequently, it is a question to what degree CRT captures dispositions related to careful, reflective responding, or is it simply a measure of numerical (and other cognitive) abilities and in that regard like other numerical tasks. Second, our Study 2 showed that, when solved correctly, CRT is often solved through relying on correct intuitions, and not through overturning faulty intuition. Thus, the main conclusion of our first two studies is that CRT is not particularly good measure of dispositions towards reflective and analytical thinking.

However, maybe the best word for describing CRT would be that it is complex. First, it is complex in the sense that it is not a "pure" measure of any specific ability or disposition, but that it probably captures a range of cognitive abilities and thinking dispositions. Second, it is complex in the sense that it probably

captures these abilities and dispositions to a different degree in different people. For example, what can we conclude if five different people, one exceptionally mathematically gifted, one that had a lot of math practice in life, one with exceptionally high IQ, one that was fooled many times on different stumpers and one that is genuinely careful all correctly solve a CRT task? Correct CRT task is probably simultaneously indicative of person's cognitive ability dispositions, practice and broader life experiences, but to what degree it taps into each of these (and probably much more) characteristics in unknown and different from person to person. Finally, CRT is complex in a sense that it might, in fact, be a good measure of disposition towards deliberative and reflective thinking for majority of people, even if most people solve it by relying on intuition. In other words, although it might not be a good "direct" measure of these dispositions, it could be an indirect measure. As we discussed in our Study 2, success on CRT in part depends on the "mindware", i.e. experience and knowledge relevant for solving the tasks. To the extent that dispositions to engage in analytical and deliberative thinking were conductive of person's repetitive engagement with these kinds of tasks to the point of developing strong, correct intuitions, correctly solved CRT task can still be an indicator of person's dispositions to be reflective and analytical, even if solved intuitively. These remarks call for future studies that should shed additional light to the nature of CRT tasks. However, what seems clear is that referring to the CRT as a measure of primarily reflective or analytical thinking, as many researchers do, is either wrong or at least a grave simplification.

Measuring rational thinking and decision—making with cognitive bias tasks is messy

Our Study 3 confirmed results of some previous studies that indicated that tasks designed to capture susceptibility to different cognitive biases are quite heterogeneous and share little in common. However, unlike previous research, in both of our studies one factor solution was the most appropriate, indicating that there might exist something resembling the rationality factor. This factors, although related to intelligence was nevertheless different from it because it was to a greater degree related to actively open-minded thinking. This points to the conclusion that using a set of cognitive bias tasks to capture individual differences in rational tendencies might not be the best approach. Instead, it seems much easier and more appropriate to put more emphasis on the assessment of actively open-minded thinking, as it was one of the main determinants of rational responding across the tasks.

There are several problems with the approach of using many different cognitive bias tasks to capture individual's rationality. First, these tasks are different one from another to much larger degree than they are similar. For one, each of the tasks requires specific knowledge and/or experience to be solved (e.g.

knowledge of logic, probability, expected utility calculation, statistics etc.). Moreover, the content of the tasks that measure the same bias significantly differ from task to task. For example, the domains of our attribute framing measures ranged from judging about acceptability of cheating at universities to taking an imagined role of research & development manager and deciding whether to pursue an expensive project. Given that the transfer of knowledge and learning between substantially different contexts is improbable (e.g. Perkins & Solomon, 1992), each task will tap into different personal experiences. Therefore, and this leads to second problem, it is questionable whether the cognitive bias tasks capture anything substantial and important outside this common core that all the tasks share (especially given that the normative knowledge is already partially accounted for by cognitive abilities). One possibility is that each task captures some specific ability, knowledge or skill that might be important at least in some aspects of work and life. However, given previous discussion, I believe that non-shared variance between tasks is situation and person specific, and, essentially, impossible to define. An additional case in point for this conclusion would be Greenberg and Harris (2022) investigation in people's perception of sunk-cost fallacy tasks similar to the ones we used. They interviewed people to get them to explain why they fell for sunk fallacy cost and concluded that many of them in fact did not fall prey to the fallacy, even though they responded in a way that would suggest that they did. For example, even though people indicated that they would eat the food they paid for even if they were full and did not like the taste, many explained that they assumed that they would not be eating alone and the did not want to seem weird, or that they felt obliged because chef put a lot of work preparing the meal. Whatever these items capture, it does not seem to be related to the sunk-cost fallacy, at least for some of the people. Therefore, it is possible that many of the tasks are not particularly valid measures of their purported constructs. Finally, giving people many challenging tasks is not practical, requiring far too much effort and time. For all these reasons, as well as the ones I will describe next, it might wise to focus on devising better measures of actively open-minded thinking as a way of measuring rationality.

Quality of the process of thinking and decision making is at least as important for the outcomes as are cognitive abilities

Final two studies of this thesis focused on the importance of decision-making styles and thinking dispositions for real-life and, especially, work related outcomes. Two main conclusions came out of these studies. First, it is crucial not shy away from making decisions. People who admitted in our studies that they are inclined towards avoiding decision making suffered all kinds of negative effects, both in personal and in work lives. This detrimental effect of avoidant decision-making style was a consistent finding across our three different samples, undergraduates, community sample and entrepreneurs. For example,

people who avoid making decisions experience substantially greater amount of unpleasant real-life outcomes as indicated by the Decision-Outcome Inventory, as well as more negative career and work-related outcomes such as lower career and job satisfaction, as well as worse work performance. Similarly, avoidant entrepreneurs were not only perceived to be worse entrepreneurs by their employees, but also tend to have employees that are less satisfied with their jobs and have higher intentions of leaving it. It is fairly easy to construct a plausible story describing why and how avoidance of decision-making produces bad outcomes. For example, someone who avoids making important decisions until the very end might miss out on opportunities to improve his/her career or work conditions, resulting in lower career and job satisfaction. Or an entrepreneur who put off decisions might not respond in time and adapt to changes in market or economic conditions, jeopardizing his/her company. However, we must be careful here not to jump to conclusions as our studies were cross sectional, thus not permitting causality claims, and because there are certainly other plausible explanations of these relationships.

Second, not only is it important to make decisions, it is also important how one makes them. Or more precisely, it is important how a person reasons and search for evidence before making a conclusion and committing to a decision. Our studies showed that the right way to do it is in actively open-minded way. Person's tendency to be actively open-minded (i.e. open to more perspectives and arguments and properly confident) was related to committing far less cognitive and logical fallacies in Study 3, as well as to holding less beliefs that are not backed up by evidence (i.e. conspiracy and superstitious beliefs) and to being judged a better decision-maker by others. For all these reasons, as well as ones described in previous chapter, I believe that operationalizing rationality as a thinking disposition, especially towards actively open-minded thinking, and measuring it with existing or (ideally) new and improved instruments, instead of relying on idiosyncratic cognitive biases tasks, is a promising way forward in the field of individual differences in rational thinking.

Actively open-minded thinking was also related to measurable and important organizational outcomes in Study 5. Specifically, we showed that the benefits of managerial AOT reach beyond decision-making – yes, managers who were more prone to this way of thinking were perceived to be better decision-makers by their subordinates, but the also tended to have subordinates that we more satisfied with their jobs and teams that were psychologically safer compared to managers lower on AOT. Thus, notwithstanding the inability of drawing strong conclusions about the causality, our results indicate that it would be prudent for companies to choose and develop their leaders based on the way they think and make decisions. For example, companies could either select new, talented managers based on their AOT

level, or teach the existing ones how to think and behave in actively open-minded way. Our results, although optimistic, only scratched the surface of understanding about possible benefits of AOT and its teaching in organizational context.

8. CONCLUSION

In short, results of the five studies described in this thesis point to the conclusion that how we think is as important, if not more, for consequential real-life outcomes, than how smart we are. For example, in none of our studies and samples was cognitive ability related to unsubstantiated beliefs or negative decision-outcomes, while thinking styles and dispositions were. Avoiding and postponing decision-making was consistently negatively related to a range of positive outcomes, while thinking in an actively open-minded way, conversely, positively predicted similar outcomes. In concordance with our additional findings that the cognitive reflection test and the battery of different cognitive biases tasks, common approaches to capturing individual differences in rationality, suffer from serious drawbacks, this thesis proposes that rationality is best conceptualized as a thinking disposition towards open-minded and other-side thinking and captured with tests or questionnaires of typical performance focused on this specific way of thinking, rather than relying on construct-problematic, maximal performance tasks that are still used widely. Finally, results suggest that these thinking dispositions should find its place in organizational context, especially when selecting or preparing for leadership positions.

LITERATURE

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015). Measuring individual differences in decision biases: Methodological considerations. *Frontiers in psychology*, *6*, 1770.

Agor, W. H. (1986). The logic of intuition: How top executives make important decisions. *Organizational Dynamics*, *14*(3), 5-18.

Ahmetoglu, G. (2015). *The entrepreneurial personality: A new framework and construct for entrepreneurship research and practice* (Doctoral dissertation, Goldsmiths, University of London).

Alaybek, B., Wang, Y., Dalal, R. S., Dubrow, S., & Boemerman, L. S. (2021a). The relations of reflective and intuitive thinking styles with task performance: A meta-analysis. *Personnel Psychology*.

Alaybek, B., Wang, Y., Dalal, R. S., Dubrow, S., & Boemerman, L. S. (2021b). Meta-analytic relations between thinking styles and intelligence. *Personality and Individual Differences, 168*, 110322.

Allan, J. N. (2018). Numeracy vs. intelligence: A model of the relationship between cognitive abilities and decision making. (Master's thesis, University of Oklahoma, Norman, USA). Retrieved from https://hdl.handle.net/11244/299906.

Arkes, H.R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1), 124–140.

Arnott, D. (2006). Cognitive biases and decision support systems development: a design science approach. *Information Systems Journal*, *16*(1), 55-78.

Attali, Y., & Bar-Hillel, M. (2020). The false allure of fast lures. *Judgment & Decision Making*, *15*(1), 93-111.

Azen R. & Budescu D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8(2), 129–148.

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

Bago, B., & De Neys, W. (2019). The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257-299.

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning, 26*(1), 1-30.

Baiocco, R., Laghi, F., & D'Alessio, M. (2009). Decision-making style among adolescents: Relationship with sensation seeking and locus of control. *Journal of adolescence*, *32*(4), 963-976.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*(3), 241-254.

Baron, J. (1985). *Rationality and intelligence*. Cambridge University Press.

Baron, J. (1985). *Rationality and intelligence*. New York: Cambridge University Press

Baron, J. (2000). *Thinking and Deciding*. Cambridge University Press.

Baron, J. (2018). Individual Mental Abilities vs. the World's Problems. *Journal of Intelligence*, 6(2), 23.

Baron, J. (2019). Actively open-minded thinking in politics. *Cognition, 188*, 8-18.

Baron, J. (2019). Actively open-minded thinking in politics. *Cognition*, *188*, 8–18. doi:10.1016/j.cognition.2018.10.004

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of personality and social psychology*, 54(4), 569-579.

Baron, J., Gürçay, B., & Metz, S. E. (2017). Reflection, intuition, and actively open-minded thinking. In M. Toplak & J. Weller (Eds.), *Individual differences in judgment and decision making: A developmental perspective*. Psychology Press.

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284. doi:10.1016/j.jarmac.2014.09.003

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

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015b). The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior, 48*, 473-480.

Bartram, D. (2005). The Great Eight competencies: a criterion-centric approach to validation. *The Journal of Applied Psychology*, *90*(6), 1185–1203. doi:10.1037/0021-9010.90.6.1185

Berthet, V. (2021). The Measurement of Individual Differences in Cognitive Biases: A Review and Improvement. *Frontiers in psychology, 12*, 419.

Berthet, V., & de Gardelle, V. (2021, April 20). Measuring individual differences in cognitive biases: The Cognitive Bias Inventory. https://doi.org/10.31219/osf.io/7wfvb

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological methods, 17*(3), 399-417.

Blacksmith, N., Behrend, T. S., Dalal, R. S., & Hayes, T. L. (2019). General mental ability and decision-making competence: Theoretically distinct but empirically redundant. *Personality and Individual Differences, 138*, 305-311.

Blacksmith, N., Yang, Y., Behrend, T. S., & Ruark, G. A. (2019). Assessing the validity of inferences from scores on the cognitive reflection test. *Journal of Behavioral Decision Making*. 1-14.

Böckenholt, U. (2012). The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika*, *77*(2), 388-399.

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645.

Brayfield, A. H., & Rothe, H. F. (1951). An index of job satisfaction. *Journal of Applied Psychology*, *35*(5), 307-311.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117. doi:10.1037/h0101832

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in psychology*, 4, 279.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, *92*(5), 938-956.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2020). Decision-making competence: More than intelligence?. *Current Directions in Psychological Science*, *29*(2), 186-192.

Budescu D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3), 542–551.

Byrd, N., Gongora, G., Joseph, B., & Sirota, M. (2021). Tell us what you really think: A think aloud protocol analysis of the verbal cognitive reflection test.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & cognition*, *42*(3), 434-447.

Čavojová, V., Šrol, J., & Jurkovič, M. (2020). Why should we try to think like scientists? Scientific reasoning and susceptibility to epistemically suspect beliefs and cognitive biases. *Applied Cognitive Psychology*, *34*(1), 85-95.

Ceschi, A., Costantini, A., Sartori, R., Weller, J., & Di Fabio, A. (2019). Dimensions of decision-making: an evidence-based classification of heuristics and biases. *Personality and Individual Differences*, *146*, 188-200.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, *4*(1), 20-33.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). Decision making skill: From intelligence to numeracy and expertise. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbook of expertise and expert performance (2nd ed.)*. New York, NY: Cambridge University Press.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). *Skilled decision theory: From intelligence to numeracy and expertise.* In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbooks in psychology. The Cambridge handbook of expertise and expert performance* (p. 476–505). Cambridge University

Cokely, E.T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*, 25-47.

Collins, C. J., Hanges, P. J., & Locke, E. A. (2004). The relationship of achievement motivation to entrepreneurial behavior: A meta-analysis. *Human performance, 17*(1), 95-117.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43,* 52-64.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281-302.

Crossley, C. D., & Highhouse, S. (2005). Relation of job search and choice process with subsequent satisfaction. *Journal of Economic Psychology*, *26*(2), 255-268.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, e28.

Dalal, R. S., & Brooks, M. E. (2013). Individual differences in decision making skill and style. In S. Highhouse, R. Dalal, & E. Salas (Eds.), *Judgment and decision making at work* (pp. 80–101). New York, NY: Taylor & Francis.

Dalal, R. S., Bonaccio, S., Highhouse, S., Ilgen, D. R., Mohammed, S., & Slaughter, J. E. (2010). What if industrial–organizational psychology decided to take workplace decisions seriously?. *Industrial and Organizational Psychology*, *3*(4), 386-405.

Damnjanović, K., Novković, V., Pavlović, I., Ilić, S., & Pantelić, S. (2019). A cue for rational reasoning: Introducing a reference point in cognitive reflection tasks. *Europe's journal of psychology*, *15*(1), 25.

Davis, D. E., Hook, J. N., DeBlaere, C., & Placeres, V. (2016). Humility at Work. In *The Wiley Blackwell Handbook of the Psychology of Positivity and Strengths Based Approaches at Work* (pp. 191–209). doi:10.1002/9781118977620.ch12

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. P*erspectives on Psychological Science, 7*(1), 28-38.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*(2), 169-187.

De Neys, W. (2015). Heuristic bias and conflict detection during thinking. In *Psychology of learning and motivation* (Vol. 62, pp. 1-32). Academic Press.

De Neys, W. (2017). Bias, conflict, and fast logic: Towards a hybrid dual process future? In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 47–65). Oxon, UK: Routledge.

De Neys, W. (2020). Rational rationalization and System 2. *Behavioral and Brain Sciences, 43*.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*(3), 1248-1299.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS one, 6*(1), e15954.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20*(2), 269-273.

De Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods, 21*(3), 273-290.

Del Missier, F., Mäntylä, T., & De Bruin, W. B. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making, 25*(4), 331-351.

Deppe, K. D., Gonzalez, F. J., Neiman, J. L., Jacobs, C., Pahlke, J., Smith, K. B., & Hibbing, J. R. (2015). Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology. *Judgment & Decision Making, 10*(4), 314-331.

Dewberry, C., Juanchich, M., & Narendran, S. (2013a). Decision-making competence in everyday life: The roles of general cognitive styles, decision-making styles and personality. *Personality and Individual Differences, 55*(7), 783-788.

Dewberry, C., Juanchich, M., & Narendran, S. (2013b). The latent structure of decision styles. *Personality and Individual Differences, 54*(5), 566-571.

Dierdorff, E. C., & Rubin, R. S. (2006). *Toward a comprehensive empirical model of managerial competencies*. Technical report presented to the MERInstitute of the Graduate Management Admission Council, McLean, VA.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological assessment, 18*(2), 192-203.

Edmondson, A. C., & Lei, Z. (2014). Psychological safety: The history, renaissance, and future of an interpersonal construct. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(1), 23–43. doi:10.1146/annurev-orgpsych-031413-091305

Edmondson, A.C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. doi:10.2307/2666999

Edmondson, A.C. (2018). *The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth*. John Wiley & Sons.

Eisenberger, R., Huntington, R., Hutchison, S., & Sowa, D. (1986). Perceived organizational support. *The Journal of Applied Psychology*, *71*(3), 500–507. doi:10.1037/0021-9010.71.3.500

Erceg, N., Galić, Z., & Bubić, A. (2019). "Dysrationalia" Among University Students: The Role of Cognitive Abilities, Different Aspects of Rational Thought and Self-Control in Explaining Epistemically Suspect Beliefs. *Europe's journal of psychology*, *15*(1), 159.

Erceg, N., Galić, Z., & Bubić, A. (2022). Normative responding on cognitive bias tasks: Some evidence for a weak rationality factor that is mostly explained by numeracy and actively open-minded thinking. *Intelligence*, *90*(101619), 101619. doi:10.1016/j.intell.2021.101619

Erceg, N., Galić, Z., & Ružojčić, M. (2020). A reflection on cognitive reflection-testing convergent/divergent validity of two measures of cognitive reflection. *Judgment & Decision Making, 15*(5), 741-755.

Evans J. S. B., Stanovich K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Evans, J. S. B. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378-395.

Evans, J. S. B. (2019). Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, 25(4), 383-415.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods , 39* , 175-191.

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging*, *25*(2), 271-288.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive psychology*, *18*(3), 253-292.

Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pezeshkan, A., & Vracheva, V. (2017). Psychological safety: A meta-analytic review and extension. *Personnel Psychology*, *70*(1), 113–165. doi:10.1111/peps.12183

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

Frese, M., & Gielnik, M. M. (2014). The psychology of entrepreneurship. *Annu. Rev. Organ. Psychol. Organ. Behav.*, *1*(1), 413-438.

Frey, D., Johnson, E. D., & De Neys, W. (2017). Individual differences in conflict detection during reasoning. *The Quarterly Journal of Experimental Psychology*, *71*(5):1188-1208.

Gambetti, E., & Giusberti, F. (2019). Personality, decision-making styles and investments. *Journal of Behavioral and Experimental Economics*, *80*, 14-24.

Garcia-Retamero, R., Sobkow, A., Petrova, D., Garrido, D., & Traczyk, J. (2019). Numeracy and risk literacy: What have we learned so far? *The Spanish Journal of Psychology*, 22. e10. Doi:10.1017/sjp.2019.16

Germeijs, V., & Verschueren, K. (2011). Indecisiveness and Big Five personality factors: Relationship and specificity. *Personality and individual differences*, *50*(7), 1023-1028.

Gervais, W. M. (2015). Override the controversy: Analytic thinking predicts endorsement of evolution. *Cognition*, *142*, 312-321.

Ghasemi, O., Handley, S., Howarth, S., Newman, I. R., & Thompson, V. A. (2021, May 18). Logical Intuition Is Not Really About Logic. https://doi.org/10.31219/osf.io/wtupm

Ghazal, S. (2014). Component numeracy skills and decision making.

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Daking*, *9*(1), 15-34.

Greenberg, S. & Harris, C. (2022, April 05). Our research shows how word choice can have a huge impact on survey results. https://www.clearerthinking.org/post/our-research-shows-how-word-choice-can-have-a-huge-impact-on-survey-results

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal, 33*, 64–86.

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, 17(3), 234-252.

Gurcay-Morris, B. (2016). *The Use of Alternative Reasons in Probabilistic Judgment*. University of Pennsylvania.

Hamilton, K., Shih, S. I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of personality assessment, 98*(5), 523-535.

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In *Psychology of learning and motivation* (Vol. 62, pp. 33-58). Academic Press.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science, 3*(2), 200-215.

Hogan, R., & Kaiser, R. B. (2005). What we know about leadership. *Review of general psychology, 9*(2), 169-180.

Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters*, *110*(2), 97-100.

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta psychologica, 164*, 56-64.

Juanchich, M., Dewberry, C., Sirota, M., & Narendran, S. (2016). Cognitive reflection predicts real-life decision outcomes, but not over and above personality and decision-making styles. *Journal of Behavioral Decision Making*, *29*(1), 52-59.

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: a meta-analytic test of their relative validity. *Journal of applied psychology*, *89*(5), 755-768.

Judge, T. A., Colbert, A. E., & Ilies, R. (2004). Intelligence and leadership: a quantitative review and test of theoretical propositions. *The Journal of Applied Psychology*, *89*(3), 542–552. doi:10.1037/0021-9010.89.3.542

Judge, T. A., Piccolo, R. F., & Kosalka, T. (2009). The bright and dark sides of leader traits: A review and theoretical extension of the leader trait paradigm. *The Leadership Quarterly, 20*(6), 855–875. doi:10.1016/j.leaqua.2009.09.004

Kahneman & Frederick (2005). A model of heuristic judgment. In K.J. Holyoak & R.G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267-293). New York, NY, US: Cambridge University Press.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Strauss, Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Kahneman, D., Lovallo, O. & Sibony, O. (2011). Before You Make That Big Decision. *Harvard Business Review, 89*(6), 50-60.

Ketchen, D. J., Jr, & Craighead, C. W. (2022). Cognitive biases as impediments to enhancing supply chain entrepreneurial embeddedness. *Journal of Business Logistics*. doi:10.1111/jbl.12307

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Konovsky, M. A., & Cropanzano, R. (1991). Perceived fairness of employee drug testing as a predictor of employee attitudes and job performance. *Journal of applied psychology, 76*(5), 698-707.

Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style. *Psychological bulletin, 133*(3), 464.

Leykin, Y., & DeRubeis, R. J. (2010). Decision-making styles and depressive symptomatology: Development of the Decision Styles Questionnaire. *Judgment and Decision making, 5*(7), 506-515.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25*(4), 361-381.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 551-578.

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing, 17*(3), 269-276.

Loo, R. (2000). A psychometric evaluation of the general decision-making style inventory. *Personality and individual differences, 29*(5), 895-905.

Lovallo, D. & Sibony, O. (2010). The Case for Behavioral Strategy. *McKinsey Quarterly*, 30-43.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3p1), 151-159.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition, 17*(1), 11-17.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., … Tetlock, P. (2015). The psychology of intelligence analysis: drivers of prediction accuracy in world politics. *Journal of Experimental Psychology. Applied*, *21*(1), 1–14. doi:10.1037/xap0000040

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. J*ournal of Cognitive Psychology*, *27*(2), 227-237.

Mintzberg, H. (2009). *Managing*. Pearson Education.

Mlačić, B. & Šakić, I. (2008). The development of Croatian markers for the big-five personality model. *Društvena istraživanja: časopis za opća društvena pitanja, 17*(1-2 (93-94)), 223-46.

Moore, D. A., & Flynn, F. J. (2008). The Case for Behavioral Decision Research in Organizational Behavior. *Academy of Management Annals, 2*(1), 399-431.

Newman, A., Donohue, R., & Eva, N. (2017). Psychological safety: A systematic review of the literature. *Human Resource Management Review*, *27*(3), 521–535. doi:10.1016/j.hrmr.2017.01.001

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2*(2), 175-220.

Nutt, P. (2002). *Why Decisions Fail: Avoiding the Blunders and Traps That Lead to Debacles*. San Francisco: Berrett-Koehler Publishers Inc.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Oreg, S., & Bayazit, M. (2009). Prone to bias: Development of a bias taxonomy from an individual differences perspective. *Review of General Psychology*, *13*(3), 175-193.

Ou, A. Y., Waldman, D. A., & Peterson, S. J. (2018). Do humble CEOs matter? An examination of CEO humility and firm outcomes. *Journal of Management*, *44*(3), 1147–1173. doi:10.1177/0149206315604187

Owens, B. P., Johnson, M. D., & Mitchell, T. R. (2013). Expressed humility in organizations: Implications for performance, teams, and leadership. *Organization Science*, *24*(5), 1517–1538. doi:10.1287/orsc.1120.0795

Pachur, T., Hertwig, R., & Steinmann, F. (2012). How do people judge risks: availability heuristic, affect heuristic, or both?. *Journal of Experimental Psychology: Applied*, 18(3), 314.

Pachur, T., Hertwig, R., & Steinmann, F. (2012). How do people judge risks: availability heuristic, affect heuristic, or both?. *Journal of Experimental Psychology: Applied, 18*(3), 314-330.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making, 18*(1), 1-27.

Parker, A. M., Bruine de Bruin, W., Fischhoff, B., & Weller, J. (2018). Robustness of decision-making competence: Evidence from two measures and an 11-year longitudinal study. *Journal of behavioral decision making, 31*(3), 380-391.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163-177.

Pekkala Kerr, S., Kerr, W., & Xu, T. (2017). Personality traits of entrepreneurs: a review of recent literature.

Pennycook, G. (2022, August 31). A framework for understanding reasoning errors: From fake news to climate change and beyond. https://doi.org/10.31234/osf.io/j3w7d

Pennycook, G., & Rand, D. G. (2019). Cognitive reflection and the 2016 US Presidential election. *Personality and Social Psychology Bulletin, 45*(2), 224-239.

Pennycook, G., & Ross, R. M. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology, 7*, 9.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). The role of analytic thinking in moral judgements and values. *Thinking & Reasoning, 20*(2), 188-214.

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., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2015a). Is the cognitive reflection test a measure of both reflection and intuition?. *Behavior Research Methods, 48*(1), 341-348.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition?. *Behavior Research Methods, 48*(1), 341-348.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment & Decision Making, 15*(4).

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition, 123*(3), 335-346.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning?. *Cognition, 124*(1), 101-106.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science, 24*(6), 425-432.

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.

Perkins, D. (2019). Learning to Reason: The Influence of Instruction, Prompts and Scaffolding, Metacognitive Knowledge, and General Intelligence on Informal Reasoning about Everyday Social and Political Issues. *Judgment and Decision Making, 14*(6), 624-43.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. *International encyclopedia of education, 2*, 6452-6457.

Peter, J. P., Churchill Jr, G. A., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of consumer research*, 19(4), 655-662.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological science, 17*(5), 407-413.

Phillips, W. J., Fletcher, J. M., Marks, A. D., & Hine, D. W. (2016). Thinking styles and decision making: *A meta-analysis. Psychological Bulletin, 142*(3), 260-290.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making, 29*(5), 453-469.

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2020). Domain-specific experience and dual-process thinking. *Thinking & Reasoning, 27*(2), 239-267.

Putarek, V., & Vlahović-Štetić, V. (2019). Metacognitive Feelings, Conflict Detection and Illusion of Linearity. *Psihologijske teme, 28*(1), 171-192.

Raoelison, M. T., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition, 204*, 104381.

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: the development of logical intuitions. *Thinking & Reasoning*, 1-25.

Rauch, A., & Frese, M. (2007). Let's put the person back into entrepreneurship research: A meta-analysis on the relationship between business owners' personality traits, business creation, and success. *European Journal of work and organizational psychology, 16*(4), 353-385.

Revelle W (2020). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.0.8, https://CRAN.R-project.org/package=psych.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological bulletin*, *135*(6), 943-973.

Reyna, V. F., Rahimi-Golkhandan, S., Garavito, D. M., & Helm, R. K. (2017). *The fuzzy-trace dual process model*. In Dual process theory 2.0 (pp. 82-99). Routledge.

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, 48(2), 1–36. http://www.jstatsoft.org/v48/i02/.

Royzman, E. B., Landy, J. F., & Goodwin, G. P. (2014). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment and Decision Making*, *9*(3), 176-190.

Russ, F. A., McNeilly, K. M., & Comer, J. M. (1996). Leadership, decision making and performance of sales managers: A multi-level approach. *Journal of Personal Selling & Sales Management*, *16*(3), 1-15.

Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of educational psychology*, *91*(3), 497.

Sadler-Smith, E. (2004). Cognitive style and the management of small and medium-sized enterprises. *Organization Studies*, *25*(2), 155-181.

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(1), 45–65. doi:10.1146/annurev-orgpsych-031413-091255

Schönbrodt, F. D., & Gerstenberg, F. X. (2012). An IRT analysis of motive questionnaires: The unified motive scales. *Journal of Research in Personality*, *46*(6), 725-742.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality*, *47*(5), 609-612.

Schubert, A. L., Ferreira, M. B., Mata, A., & Riemenschneider, B. (2021). A diffusion model analysis of belief bias: Different cognitive mechanisms explain how cognitive abilities and thinking styles contribute to conflict resolution in reasoning. *Cognition*, *211*, 104629.

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement*, *55*(5), 818-831.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423-428.

Shtulman, A., & McCallum, K. (2014). Cognitive reflection predicts science understanding. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Sibony, O. (2020). *You're About to Make a Terrible Mistake!: How Biases Distort Decision-Making and What You Can Do to Fight Them*. Swift Press.

Singh, R., & Greenhaus, J. H. (2004). The relation between career decision-making strategies and person–job fit: A study of job changers. *Journal of vocational behavior*, *64*(1), 198-221.

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. *Journal of Behavioral Decision Making*, *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. *Journal of behavioral and experimental economics*, *74*, 18-25.

Slabbinck, H., Van Witteloostuijn, A., Hermans, J., Vanderstraeten, J., Dejardin, M., Brassey, J., & Ramdani, D. (2018). The added value of implicit motives for management research Development and first validation of a Brief Implicit Association Test (BIAT) for the measurement of implicit motives. *PloS one*, *13*(6), e0198094.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, *119*(1), 3-22.

Sloman, S. A. (2014). Two systems of reasoning: An update. In Sherman, J., Gawronski, B., & Trope, Y. (Eds.). *Dual process theories of the social mind*. Guilford Press.

Slugoski, B. R., Shields, H. A., & Dawson, K. A. (1993). Relation of conditional reasoning to heuristic processing. *Personality and Social Psychology Bulletin*, *19*(2), 158-166.

Soane, E., & Chmiel, N. (2005). Are risk preferences consistent?: The influence of decision domain and personality. *Personality and Individual Differences*, *38*(8), 1781-1791.

Sobkow, A., Olszewska, A., & Traczyk, J. (2020). Multiple numeric competencies predict decision outcomes beyond fluid intelligence and cognitive reflection. *Intelligence*, *80*, 101452.

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology*, *95*(4), 781–790.

Spicer, D. P., & Sadler-Smith, E. (2005). An examination of the general decision making style questionnaire in two UK samples. *Journal of Managerial Psychology*. *20* (2005), 137-149

Šrol, J. (2020, January 9). Individual differences in epistemically suspect beliefs: The role of susceptibility to cognitive biases. https://doi.org/10.31234/osf.io/4jcf7

Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning*, *27*(1), 38-68.

Stankov, L. (2017). Overemphasized "g". *Journal of Intelligence*, *5*(4), 33.

Stanovich K. E. (2009). Rational and irrational thought: The thinking that IQ tests miss. *Scientific American Mind*, *20*(6), 34–39.

Stanovich, K. E. (2002). Rationality, intelligence, and levels of analysis in cognitive science: Is dysrationalia possible? In R. J. Sternberg (Ed.), *Why smart people can be so stupid* (pp. 124–158). New Haven, CT: Yale University Press

Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic and autonomous minds: Is it time for a tri-process theory? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford, England: Oxford University Press

Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought*. Yale University Press.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: implications for understanding individual differences in reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 433–455). New York: Oxford University Press.

Stanovich, K. E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423-444.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342–357. doi:10.1037/0022-0663.89.2.342

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of experimental psychology: general*, *127*(2), 161.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. Behavioral and brain sciences, 23(5), 645-665.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning*, 13, 225–247.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology*, *94*(4), 672.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. In *Advances in child development and behavior* (Vol. 36, pp. 251-285). JAI.

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

Stewart, W.H. & P.L. Roth. (2007). A meta-analysis of achievement motivation differences between entrepreneurs and managers. *Journal of Small Business Management, 45*(4), 401-421

Stupple, E. J., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning, 19*(1), 54-77.

Stupple, E. J., Pitchford, M., Ball, L. J., Hunt, T. E., & Steel, R. (2017). Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. *PloS one, 12*(11), e0186404.

Svedholm, A. M., & Lindeman, M. (2013). The separate roles of the reflective mind and involuntary inhibitory control in gatekeeping paranormal beliefs and the underlying intuitive confusions. *British Journal of Psychology (London, England: 1953), 104*(3), 303–319. doi:10.1111/j.2044-8295.2012.02118.x

Svedholm-Häkkinen, A. M., & Lindeman, M. (2018). Actively open-minded thinking: development of a shortened scale and disentangling attitudes towards knowledge and people. *Thinking & Reasoning*, *24*(1), 21–40. doi:10.1080/13546783.2017.1378723

Swain, J. E. (2018). Effects of leader humility on the performance of virtual groups. *Journal of Leadership Studies*, *12*(1), 21–37. doi:10.1002/jls.21552

Swain, J. E., & Murray, E. D. (2020). Assessing leader humility. *Journal of College and Character*, *21*(3), 204–211. doi:10.1080/2194587x.2020.1781657

Swan, A. B., Calvillo, D. P., & Revlin, R. (2018). To detect or not to detect: A replication and extension of the three-stage model. *Acta psychologica*, *187*, 54-65.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. *Thinking & Reasoning, 23*(3), 207-234.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics: Pearson new international edition.* Pearson Higher Ed.

Teovanović, P. (2019). Dual Processing in Syllogistic Reasoning: An Individual Differences Perspective. *Psihologijske teme, 28*(1), 125-145.

Teovanović, P. R. (2013). *Sklonost kognitivnim pristrasnostima*. Универзитет у Београду.

Teovanović, P., Knežević, G., & Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intelligence, 50*, 75-86.

Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. J. (2000). Development and content validation of a 'hyperdimensional' taxonomy of managerial competence. *Human Performance*, *13*(3), 205–251. doi:10.1207/s15327043hup1303_1

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215-244.

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. (2018). Do smart people have better intuitions?. *Journal of Experimental Psychology: General*, *147*(7), 945-961.

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive psychology*, *63*(3), 107-140.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, *11*(1), 99-113.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, *11*(1), 99.

Thunholm, P. (2004). Decision-making style: habit, style or both?. *Personality and individual differences*, *36*(4), 931-944.

Thunholm, P. (2008). Decision-making styles and physiological correlates of negative stress: Is there a relation?. *Scandinavian Journal of Psychology*, *49*(3), 213-219.

Thunholm, P. (2009). Military leaders and followers–do they have different decision styles?. *Scandinavian Journal of Psychology*, *50*(4), 317-324.

Tishman, S., & Andrade, A. (1996). Thinking dispositions: A review of current theories, practices, and issues. *Cambridge, MA. Project Zero, Harvard University*.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition*, *39*(7), 1275-1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014a). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental psychology*, *50*(4), 1037-1048.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014b). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147- 168.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making*, *30*(2), 541-554.

Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics*, 2(1), 1064626.

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109-118.

Trippas, D., & Handley, S. (2017). The parallel processing model of belief bias: Review and extensions. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 28–46). Oxon, UK: Routledge.

Trippas, D., Pennycook, G., Verde, M. F., & Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. T*hinking & Reasoning*, *21*(4), 431-445.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

Ülgen, B., Sağlam, M., & Tuğsal, T. (2016). Managers' personality traits over management styles and decision-making styles. *International Journal of Commerce and Finance*, *2*(1), 125-136.

van den Bergh, D., Clyde, M. A., Raj, A., de Jong, T., Gronau, Q. F., Ly, A., & Wagenmakers, E. J. (2020). A Tutorial on Bayesian Multi-Model Linear Regression with BAS and JASP.

Vrhovnik, A. (2022). *Prilog validaciji testa situacijske prosudbe za mjerenje aktivnog otvorenog mišljenja* (Doctoral dissertation, University of Zagreb. Faculty of Humanities and Social Sciences. Department of Psychology).

Wang, Y., Liu, J., & Zhu, Y. (2018). Humble leadership, psychological safety, knowledge sharing, and follower creativity: A cross-level investigation. *Frontiers in Psychology*, *9*, 1727. doi:10.3389/fpsyg.2018.01727

Ward, M. K. (2016). The Relationships between Decision-Making Styles of Entrepreneurs and Organizational Performance.

Wason, P. C. (1966). Reasoning. In B. M. Foss, (Ed.), *New Horizons in psychology 1*. Harmondsworth: Penguin.

Wason, P. C. (1968). Reasoning about the rule. *Quarterly Journal of Experimental Psychology, 20*, 3, 273–281.

Weaver, E. A., & Stewart, T. R. (2012). Dimensions of judgment: Factor analysis of individual differences. *Journal of Behavioral Decision Making, 25*(4), 402-413.

Weller, J. A., Moholy, M., Bossard, E., & Levin, I. P. (2015). Preadolescent decision-making competence predicts interpersonal strengths and difficulties: A 2-year prospective study. *Journal of Behavioral Decision Making*, *28*(1), 76-88.

Weller, J., Ceschi, A., Hirsch, L., Sartori, R., & Costantini, A. (2018). Accounting for individual differences in decision-making competence: Personality and gender differences. *Frontiers in psychology*, *9*, 2258.

Welsh, M., Burns, N., & Delfabbro, P. (2013). The cognitive reflection test: How much more than numerical ability?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*(4), 930-941.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. PloS one, 11(3), e0152719.

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of management, 17*(3), 601-617.

Wood, N. L. (2012). *Individual differences in decision-making styles as predictors of good decision making* (Doctoral dissertation, Bowling Green State University).

Wood, N. L., & Highhouse, S. (2014). Do self-reported decision styles relate with others' impressions of decision quality?. *Personality and Individual Differences, 70*, 224-228.

Yukl, G. (2013). *Leadership in organizations* (8th ed.). Englewood Cliffs, NJ: Prentice Hall.

Zhao, H. & S.E. Seibert. (2006). The big five personality dimensions and entrepreneurial status: A metaanalytical review. *Journal of Applied Psychology, 91*, 259-271.

APPENDIX A

Appendix A consists of instruments that we used across the studies.

Study 1 Instruments

CRT

A bat and a ball together cost 110 kunas. The bat costs 100 kunas more than the ball. How much does the ball cost? Correct: 5; Lure: 10.

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? Correct: 5; Lure: 100.

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 the lake? Correct: 47; Lure: 24.

Josip received a grade that is at the same time the fifteenth highest and the fifteenth lowest in the class. How many students are there in his class? Correct: 29; Lure: 30.

Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has:

a. broken even in the stock market,

b. is ahead of where he began, (lure)

c. has lost money (correct)

If 3 elves can wrap 3 toys in 1 hour, how many elves are needed to wrap 6 toys in 2 hours? Correct: 3; Lure: 6.

In an athletic team, tall athletes are three times more likely to win a medal than short athletes. This year the team has won 60 medals so far. How many of those medals were won by short athletes? Correct: 15; Lure: 20.

A square shaped garage roof with 6 meters long edge is covered with 100 tiles. How many tiles of the same size are covering a neighbouring roof, which is also square shaped, but with a 3 meters long edge? Correct: 25; Lure: 50.

There are two swimming pools in a swimming facility and in the summer they need to be filled with water. 100 liters of water are required to fill the cube-shaped pool. How many liters of water does it take to fill a cube-shaped pool but with a 3 times longer edges? Correct: 2700; Lure: 300.

. 25 soldiers are standing in a line 3 meters apart from each other. How many meters is the line long? Correct: 72; Lure: 75.

Belief bias syllogisms (all are believable, but logically incorrect)

1. Premise 1: All unemployed people are poor. Premise 2: Todorić* is not unemployed. Conclusion: Todorić is not poor.
2. Premise 1: All flowers have petals. Premise 2: Roses have petals. Conclusion: Roses are flowers.
3. Premise 1: All Eastern countries are communist. Premise 2: Canada is not an Eastern country. Conclusion: Canada is not communist.
4. Premise 1: All things that have a motor need oil. Premise 2: Automobiles need oil. Conclusion: Automobiles have motors

* Todorić is a well-known Croatian rich businessman

Berlin numeracy test

1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent. Correct response: **25 %**
2. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? Correct response: **30** out of 50 throws.
3. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6? Correct response: **20** out of 70 throws.
4. In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? Correct response: **50**

Verbal reasoning

1. What number is one fifth of one fourth of one ninth of 900?

   2; 3; 4; **5 (correct)**; 6; 7.

2. Zach is taller than Matt and Richard is shorter that Zach. Which of the following statements would be most accurate?
    1. Richard is taller than Matt.
    2. Richard is shorter than Matt.
    3. Richard is as tall as Matt.
    4. **It's impossible to tell. (correct)**
3. Joshua is 12 years old and his sister is three times as old as he. When Joshua is 23 years old, how old will his sister be?

    35; 39; 44; **47 (correct)**; 53; 57.
4. If the day after tomorrow is two days before Thursday then what day is today?

    Friday; Monday; Wednesday; Saturday; Tuesday; **Sunday (correct)**.

AOT

1. There are two kinds of people in this world: those who are for the truth and those who are against the truth.
2. Changing your mind is a sign of weakness.
3. I believe we should look to our religious authorities for decisions on moral issues.
4. No one can talk me out of something I know is right.
5. Basically, I know everything I need to know about the important things in life.
6. Considering too many different opinions often leads to bad decisions.
7. There are basically two kinds of people in this world, good and bad.
8. Most people just don't know what's good for them.
9. It is a noble thing when someone holds the same beliefs as their parents.
10. I believe that loyalty to one's ideals and principles is more important than "open-mindedness."
11. Of all the different philosophies which exist in the world there is probably only one which is correct.
12. One should disregard evidence that conflicts with your established beliefs.
13. I think that if people don't know what they believe in by the time they're 25, there's something wrong with them.
14. I believe letting students hear controversial speakers can only confuse and mislead them.
15. Intuition is the best guide in making decisions.

Base-rate neglect

Among the 1000 people that participated in the study, there were 995 nurses and 5 doctors. John is randomly chosen participant in this research. He is 34 years old. He lives in a nice house in a fancy neighborhood. He expresses himself nicely and is very interested in politics. He invests a lot of time in his career. Which is more likely?

a) John is a nurse. (correct)
b) John is a doctor.

Among the 1000 people that participated in the study, there were 100 engineers and 900 lawyers. George is randomly chosen participant in this research. George is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and developing computer programs. Which is more likely?

a) George is an engineer.
b) George is a lawyer. (correct)

Among the 1000 people that participated in the study, there were 50 16-year-olds and 950 50-year-olds. Helen is randomly chosen participant in this research. Helen listens to hip hop and rap music. She likes to wear tight T-shirts and jeans. She loves to dance and has a small nose piercing. Which is more likely?

a) Helen is 16 years old.
b) Helen is 50 years old. (correct)

Among the 1000 people that participated in the study, there were 70 people whose favorite movie was "Star wars" and 930 people whose favorite movie was "Love actually." Nikola is randomly chosen participants in this research. Nikola is 26 years old and is studying physics. He stays at home most of the time and loves to play video games. Which is more likely?

a) Nikola's favorite movie is "Star wars"
b) Nikola's favorite movie is "Love actually" (correct)

One international student conference was attended by 50% of Germans, 30% of Italians and 20% of Poles. One of the participants, an architecture student, described himself as a temperamental but friendly, fan of football, good weather and pretty girls. In your opinion, the participant is from:

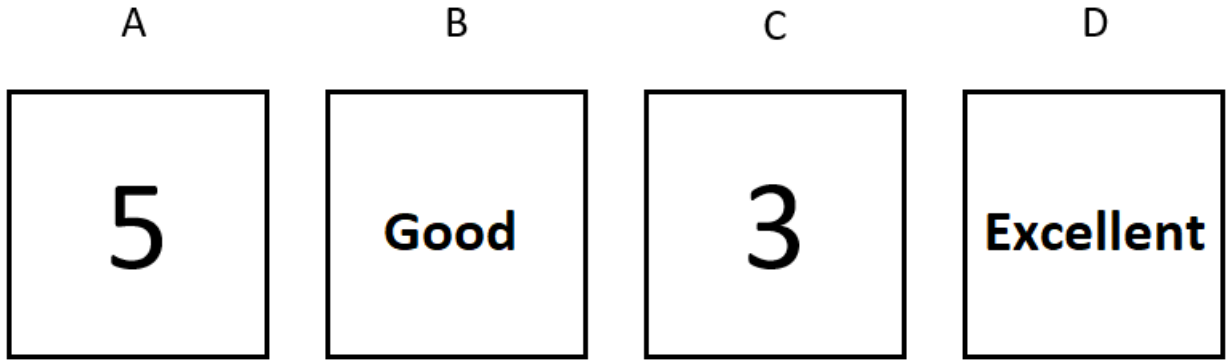
a) Germany (correct)
b) Italy
c) Poland

Four card selection task

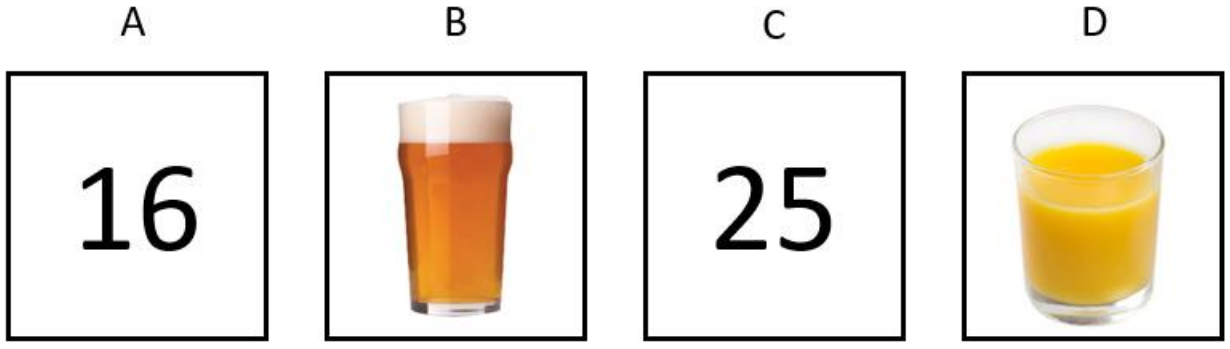
The cards you see in front of you are printed on both sides. The content of the cards is determined by some rule. In this task, a rule is proposed to determine the content of these cards. However, this rule may or may not be correct.

To find out if this rule is correct or not, we give you the opportunity to turn two cards and see what's on the back of those cards. So, your job is to check that the rule described in the task is correct by only turning two cards.

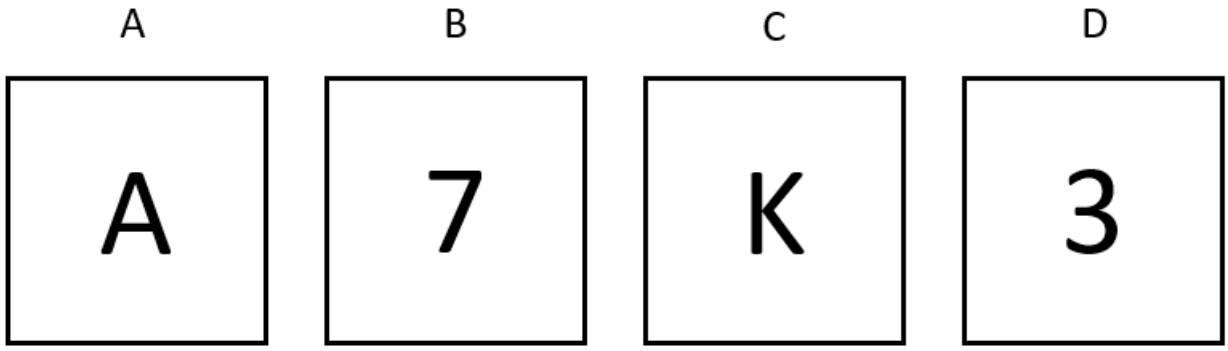
1. Rule: If a card shows “5” on one face, the word "excellent" is on the opposite face. Which two cards would you choose to turn to check the accuracy of this rule? Correct: cards A and B.



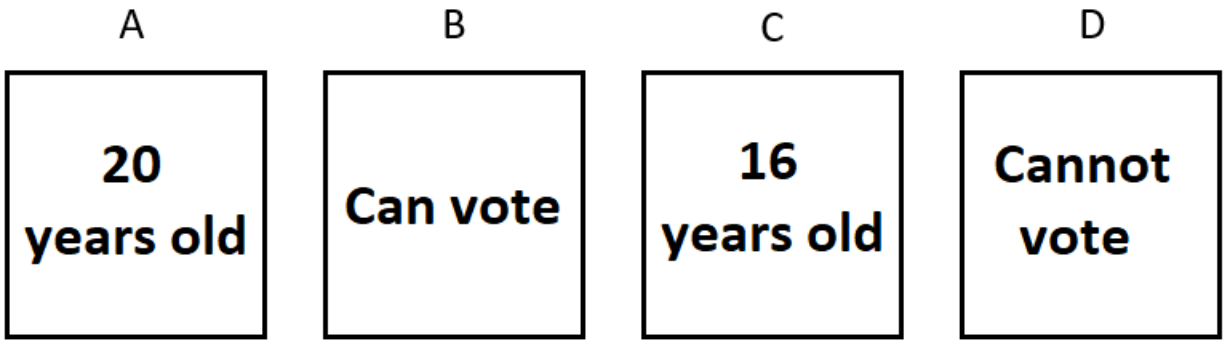
2. Rule: If a person drinks beer, he/she must be over 18 years old. Which two cards would you choose to turn to check the accuracy of this rule? Correct: B and A.



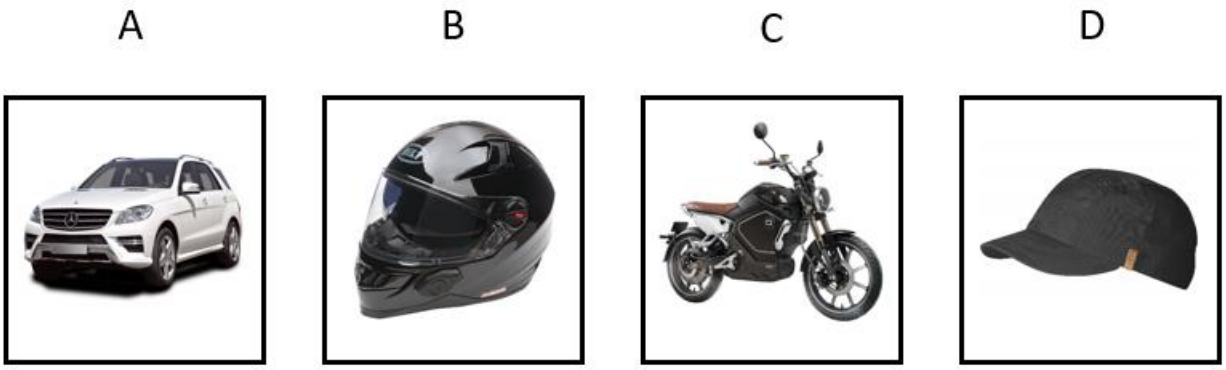
3. Rule: If a card shows letter A on one face, a number 3 is on the opposite face. Which two cards would you choose to turn to check the accuracy of this rule? Correct: A and B.



4. Rule: If a person is over 18 years old, he/she has the right to vote. Which two cards would you choose to turn to check the accuracy of this rule? Correct: A and D.



5. Rule: If a person rides a motorcycle, then he/she wears a helmet. Which two cards would you choose to turn to check the accuracy of this rule? Correct: C and D.



Causal base-rate

1. As the Chief Financial Officer of a corporation, you are planning to buy new laptops for the workers of the company. Today, you have to choose between two types of laptops that are almost identical with regard to price and the most important capabilities. According to statistics from trusted sources, type “A” is much more reliable than type “B”. One of your acquaintances, however, tells

you that the motherboard of the type “A” laptop he bought burnt out within a month and he lost a significant amount of data. As for type “B”, none of your acquaintances have experienced any problems. You do not have time for gathering more information. Which type of laptop will you buy?

a) Definitely type A

b) Probably type A

c) Probably type B

d) Definitely type B

2. Professor Kellan, the director of a teacher preparation program, was designing a new course in human development and needed to select a textbook for the new course. She had narrowed her decision down to one of two textbooks: one published by Pearson and the other published by McGraw. Professor Kellan belonged to several professional organizations that provided Web-based forums for its members to share information about curricular issues. Each of the forums had a textbook evaluation section, and the websites unanimously rated the McGraw textbook as the better choice in every category rated. Categories evaluated included quality of the writing, among others. Just before Professor Kellan was about to place the order for the McGraw book, however, she asked an experienced colleague for her opinion about the textbooks. Her colleague reported that she preferred the Pearson book. What do you think Professor Kellan should do?

a) Should definitely use the Pearson textbook

b) Should probably use the Pearson textbook

c) Should probably use the McGraw textbook

d) Should definitely use the McGraw textbook

3. The Caldwells had long ago decided that when it was time to replace their car they would get what they called "one of those solid, safety-conscious, built-to-last Swedish" cars -- either a Volvo or a Saab. When the time to buy came, the Caldwells found that both Volvos and Saabs were expensive, but they decided to stick with their decision and to do some research on whether to buy a Volvo or a Saab. They got a copy of Consumer Reports and there they found that the consensus of the experts was that both cars were very sound mechanically, although the Volvo was felt to be slightly superior on some dimensions. They also found that the readers of Consumer Reports who owned a Volvo reported having somewhat fewer mechanical problems than owners of Saabs. They were about to go

and strike a bargain with the Volvo dealer when Mr. Caldwell remembered that they had two friends who owned a Saab and one who owned a Volvo. Mr. Caldwell called up the friends. Both Saab owners reported having had a few mechanical problems but nothing major. The Volvo owner exploded when asked how he liked his car. "First that fancy fuel injection computer thing went out: $400 bucks. Next I started having trouble with the rear end. Had to replace it. Then the transmission and the brakes. I finally sold it after 3 years at a big loss." What do you think the Caldwells should do?

a. They should definitely buy the Saab.

b. They should probably buy the Saab.

c. They should probably buy the Volvo.

d. They should definitely buy the Volvo.

Gambler's fallacy

1. When playing slot machines, people win something 1 out of every 10 times. Julie, however, has just won on her first three plays. *What are her chances of winning the next time she plays?*
____ out of _____ (Correct: 1 out of 10).

2. Imagine that we are tossing a fair coin (a coin that has a 50/50 chance of coming up heads or tails) and it has just come up heads 5 times in a row. For the 6th toss do you think that:

a. It is more likely that tails will come up than heads.

b. It is more likely that heads will come up than tails.

c. Heads and tails are equally probable on the sixth toss. (correct)

3. The coin was tossed five times, but you were not present. You asked acquaintances what the order of the heads and tails was. Dinko told you that the order was "head-head-head-head-head", and Vinko that the order was "tail-tail-head-tail-head"? Who do you think is more likely to tell the truth?

a) Dinko

b) Vinko

c) It is equally likely that they are both telling the truth (correct)

4. People typically have a 50% chance of having a male and a 50% chance of having a female child. However, Ilija and Ivana currently have four daughters and are expecting their fifth child. What is the probability that Ivana will give birth to a son?

a) Less than 50%

b) 50% (correct)

c) More than 50%

5. Four babies were born in one hospital today. As usual, two local newspapers reported this news. "Daily Events" newpaper reported that the order of births was "Boy - Boy - Boy - Boy", while "World in Your Hand" newspaper reported that the order was "Girl - Boy - Boy - Girl". Only one of these two sources reported accurate information. What is the probability that the order reported by the "Daily Events" is correct?

a) Less than 50%

b) 50% (correct)

c) More than 50%

Availability bias

Which cause of death is more likely?

1. Suicide (less likely) vs. Diabetes
2. Homicide (less likely) vs. Diabetes
3. Commercial airplane crash (less likely) vs. Bicycle-related
4. Shark attack (less likely) vs. Hornet, wasp or bee bite

Study 2 instruments

Study 1 items

CRT

1. A bat and a ball together cost 110 kunas. The bat costs 100 kunas more than the ball. How much does the ball cost?
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

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 the lake?
4. Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has: a. broken even in the stock market, b. is ahead of where he began, c. has lost money.
5. A farmer had 15 sheep and all but 8 died. How many are left?

CRT control items

1. The magazine and banana cost 29 kunas. The magazine costs 20 kunas. How much does a banana cost?
2. If it takes 1 machine 5 minutes to make 5 widgets, how long would it take 1 machine to make 100 widgets?
3. In a lake, there is a patch of lily pads. Every day, the patch grows for 10 m2. If it takes 30 days for the patch to cover the 400 m2 of the lake, how long would it take for the patch to cover 390 m2 of the lake?
4. On Monday, the air temperature was 22 °C. Two days later, temperatures dropped by 50%. Fortunately, by Saturday the temperature had risen again by 125%. Compared to Monday, on Saturday it was: a. Equally warm; b. Warmer; c. Colder.
5. A man had 15 apples, but he decided to share 8. How many apples are left?

BBS items

1. Premise 1: All unemployed people are poor. Premise 2: Todorić* is not unemployed. Conclusion: Todorić* is not poor.
2. Premise 1: All flowers have petals. Premise 2: Roses have petals. Conclusion: Roses are flowers.
3. Premise 1: All Eastern countries are communist. Premise 2: Canada is not an Eastern country. Conclusion: Canada is not communist.
4. Premise 1: All things that have a motor need oil. Premise 2: Automobiles need oil. Conclusion: Automobiles have motors

Study 2 items

CRT items

1. A bat and a ball together cost 110 kunas. The bat costs 100 kunas more than the ball. How much does the ball cost?
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
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 the lake?
4. A farmer had 15 sheep and all but 8 died. How many are left?

5. A square shaped garage roof with 6 meters long edge is covered with 100 tiles. How many tiles of the same size are covering a neighboring roof, which is also square shaped, but with a 3 meters long edge?
6. If you were running a race, and you passed the person in 2nd place, what place would you be in now?
7. Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?

BBS items

1. Premise 1: All unemployed people are poor. Premise 2: Todorić* is not unemployed. Conclusion: Todorić* is not poor.
2. Premise 1: All things that are smoked are good for the health. Premise 2: Cigarettes are smoked. Conclusion: Cigarettes are good for the health.
3. Premise 1: All flowers have petals. Premise 2: Roses have petals. Conclusion: Roses are flowers.
4. Premise 1: All animals love water. Premise 2: Cats do not like water. Conclusion: Cats are not animals.
5. Premise 1: All Eastern countries are communist. Premise 2: Canada is not an Eastern country. Conclusion: Canada is not communist.
6. Premise 1: All mammals walk. Premise 2: Whales are mammals. Conclusion: Whales walk.
7. Premise 1: All things that have a motor need oil. Premise 2: Automobiles need oil. Conclusion: Automobiles have motors

Study 3 instruments

Cognitive biases tasks (the rest of the tasks are the same as the ones from Study 1, see above)

Belief bias syllogisms

1. Premise 1: All unemployed people are poor. Premise 2: Todorić* is not unemployed. Conclusion: Todorić is not poor. (incorrect)
2. Premise 1: All flowers have petals. Premise 2: Roses have petals. Conclusion: Roses are flowers. (incorrect)
3. Premise 1: All Eastern countries are communist. Premise 2: Canada is not an Eastern country. Conclusion: Canada is not communist.(incorrect)
4. Premise 1: All things that have a motor need oil. Premise 2: Automobiles need oil. Conclusion: Automobiles have motors. (incorrect)
5. Premise 1: All things that are smoked are healthy. Premise 2: Cigarettes are smoked. Conclusion: Cigarettes are healthy. (correct)

6. Premise 1: All four-legged animals are dangerous. Premise 2: Poodles are not dangerous. Conclusion: Poodles do not have four legs. (correct)
7. Premise 1: All mammals can walk. Premise 2: Whales are mammals. Conclusion: Whales can walk. (correct)
8. Premise 1: All animals like water. Premise 2: Cats do not like water. Conclusion: Cats are not animals. (correct)

Attribute framing

1. A. You would like to apply for funding for the development of a new technology. A consulting firm offers to write the application for you. The firm is one of the more expensive consulting firms. According to the information available to you, this firm loses applications in 5 out of 20 cases. Based on this information, would you accept the company's offer to write an application?

Please indicate how willing you would be to hire the consulting firm.

(1 - Definitely not hire; 6 - Definitely hire)

1. B. You are the owner of a rural hotel. To repair and expand the building, you would like to apply for funding. A consulting firm offers to write the application for you. The firm is one of the more expensive consulting firms. According to the information available to you, the firm wins applications in 15 out of 20 cases. Based on this information, would you accept the company's offer to write the application?

Please indicate how willing you would be to hire the consulting firm.

(1 - Definitely not hire; 6 - Definitely hire)

2.A. In a recent confidential survey completed by graduating seniors, 35% of those completing the survey stated that they had never cheated during their college career. Considering the results of the survey, how would you rate the incidence of cheating at your university?
(1 – Very low; 6 – Very high)
2.B. In a recent confidential survey completed by graduating seniors, 65% of those completing the survey stated that they had cheated during their college career. Considering the results of the survey, how would you rate the incidence of cheating at your university?
(1 – Very low; 6 – Very high)

3.A. As R&D manager, one of your project teams has come to you requesting an additional 600,000 HRK in funds for a project you instituted several months ago. The project is already behind schedule

and over budget, but the team still believes it can be successfully completed. Evaluating the situation, you believe there is a fair chance the project will not succeed, in which case the additional funding would be lost; if successful, however, the money would be well spent. You also noticed that of the projects undertaken by this team, 30 of the last 50 have been successful. What is the likelihood you would fund the request?
(1 – Very unlikely; 6 – Very likely)
3.B. As R&D manager, one of your project teams has come to you requesting an additional 600,000 HRK in funds for a project you instituted several months ago. The project is already behind schedule and over budget, but the team still believes it can be successfully completed. Evaluating the situation, you believe there is a fair chance the project will not succeed, in which case the additional funding would be lost; if successful, however, the money would be well spent. You also noticed that of the projects undertaken by this team, 20 of the last 50 have been unsuccessful. What is the likelihood you would fund the request?
(1 – Very unlikely; 6 – Very likely)

4.A. Imagine going to work by public transport every day and in 80% of cases having to wait longer than three minutes for the transport to arrive. How satisfied would you be with public transportation services?
(1 – Very dissatisfied; 6 – Very satisfied)
4.B. Imagine going to work by public transport every day and in 20% of cases having to wait less than three minutes for the transport to arrive. How satisfied would you be with public transportation services?
(1 – Very dissatisfied; 6 – Very satisfied)

Outcome bias
1.A. In two days you have an important presentation of your project in front of potential investors. It's a beautiful day and friends have invited you over for a barbecue. You accepted the invitation. You had a great time there and stayed almost until morning. The next day you spent a good part of the day preparing for the presentation, but the presentation was not very successful and the investors decided not to finance you. How good was your decision to have a barbecue with friends?
(1 – Very bad decision; 6 – Very good decision)

1.B. You have an exam in two days. Yesterday, a friend invited you to a party. You have decided to go to the party. You had a great time there and stayed almost until morning. The next day you studied a good part of the day and passed the exam. How good was your decision to go to a party?
(1 – Very bad decision; 6 – Very good decision)

2. A. You needed shoes. As the model you really liked was not available from the local stores, you have decided to order it online, where it was also slightly cheaper than you expected. Only, you weren't sure if you guessed the right size as it was expressed with a number from the American footwear metric system. The shoes arrived after a week, nicer and more comfortable than you imagined. You were very pleased with them for the next few years. How good was your decision to buy shoes online?
(1 – Very bad decision; 6 – Very good decision)

2.B. Ivan is a writer who is claimed to have considerable creative potential, but has so far made good money writing the lyrics of commercial songs. He recently came up with a "big" idea for his first novel. If he writes it, and the audience accepts it, it will be a qualitative leap in his career. On the other hand, if readers do not accept it, he will spend a great deal of time and energy on a project that will not pay off for him. Ivan, however, decided to devote time to writing the novel. Unfortunately, the novel went unnoticed. How good was Ivan's decision to write the novel?
(1 – Very bad decision; 6 – Very good decision)

3.A. The biotechnology company is considering investing in the development of a completely new technology. If the technology is recognized in the market, the investment will pay off many times over. However, experts believe that the investment is quite risky because the company would have to take out a fairly large loan to finance it. According to them, there is a 10% chance that the project will fail and that the whole company will go bankrupt as a result. In the end, the company's management decided to invest and the investment was very successful. How good, in your opinion, was company's management decision to invest in new technology?
(1 – Very bad decision; 6 – Very good decision)

3.B. AeroWings management is considering launching an ambitious space tourism project. If the project is successful, the investment will pay off many times over. However, experts consider the project to be very risky because it requires very high financial investments. According to them, there is a 10% chance that the project will fail and that the whole company will go bankrupt as a result. In the end, the company's management decided to invest in the project, but, unfortunately, the project

was not successful and the company went bankrupt because of that. How good, in your opinion, was company's management decision to invest in new project?
(1 – Very bad decision; 6 – Very good decision)

4.A. In a recent conversation, an acquaintance presented you with a rather interesting investment opportunity. Based on reliable economic analysis, there is a 90% chance that you would have a very high return on your investment. However, if you want to get into that investment, you have to invest considerable amount of money. You decided to invest, the business succeeded and your investment brought you a very high return. How good was your decision to pursue this investment opportunity?
(1 – Very bad decision; 6 – Very good decision)
4.B. You are the owner and manager of a small business. You have the opportunity to apply for a tender that, if selected, would ensure sales and a very large income in the coming years. However, applying for a tender requires serious preparation and investing large amounts of money in the preparation. If you apply and are not selected, the company will suffer significant financial losses. According to expert estimates, your company has a 90% chance of being selected in a competition. You decided to apply for the tender, but you were not selected and because of that the company suffered very serious financial losses. How good was your decision to apply for this tender?

Sunk cost
1. You paid for a vacation in Greece. However, you played a lottery and won a free vacation in Spain, which is a more attractive opportunity for you, but which would be at the exact same time as a paid vacation in Greece. Unfortunately, you can no longer get a refund for paid vacation. You must opt for one of two vacations that are of the same duration. Which one would you choose?
(1 – Definitely the Greek one; 6 – Definitely the Spanish one)

2. As a director of the company, for the purpose of financial education of your three employees, you decided to send them to Course A. The total price you paid for Course A is 12,000 kuna. However, next week you also learned about Course B, which covers the same topics as Course A, and would be more useful for your employees than Course A. Course B is also cheaper - it costs only HRK 2,500. So you decided to afford your employees a course B. However, a few days after paying for both courses, you realized that they are held at exactly the same time. Both courses are strongly based on practical work. As a result, your employees would not be able to transfer the skills acquired in the

course to other employees. Also, you cannot get a refund for any of the two courses. Which course would you send your employees to?
(1 – Definitely the Course A; 6 – Definitely the Course B)

3. You stay alone in a hotel room. You paid 35 kn to watch the movie on TV. After 15 minutes you're bored, the movie is pretty bad, and you can watch other things on regular TV that might be more interesting. Would you continue to watch the movie or not?
(1 – Definitely continue watching; 6 – Definitely stop watching)

4. You went to the movies with a friend, but after the first 20 minutes of the movie, both you and a friend find the movie to be terribly boring. While you're sorry for the money you paid for tickets, you both feel you'd have a better time at a nearby cafe. You could sneak out of the cinema without anyone noticing. Would you go or stay and watch the movie until the end?
(1 – Definitely stay; 6 – Definitely go)

Risk framing

1.A. Imagine that recent evidence has shown that a pesticide is threatening the lives of 1,200 endangered animals. Two response options have been suggested:
If Option A is used, 600 animals will be saved for sure.
If Option B is used, there is a 75% chance that 800 animals will be saved, and a 25% chance that no animals will be saved.
Which option do you recommend to use?
(1 – Definitely would choose A 6 - Definitely would choose B)

1. B. Imagine that recent evidence has shown that a pesticide is threatening the lives of 1,200 endangered animals. Two response options have been suggested:
If Option A is used, 600 animals will be lost for sure.
If Option B is used, there is a 75% chance that 400 animals will be lost, and a 25% chance that 1,200 animals will be lost.
Which option do you recommend to use?
(1 – Definitely would choose A 6 - Definitely would choose B)

2.A. You unexpectedly received an inheritance of HRK 60,000, but you are not the only heir. You can choose between two options.

If you choose option A, you will lose HRK 24,000 out of HRK 60,000.
If you choose option B, you participate in a lottery in which you have a 40% chance of losing nothing and a 60% chance of losing all 60,000 kuna.
Which option would you choose?
(1 – Definitely would choose A 6 - Definitely would choose B)
2.A. You unexpectedly received an inheritance of HRK 60,000, but you are not the only heir. You can choose between two options.
If you choose option A, you will get 36,000 out of 60,000 kuna.
If you choose option B, you participate in a lottery in which you have a 40% chance of winning all 60,000 kuna and a 60% chance of winning nothing.
Which option would you choose?
(1 – Definitely would choose A 6 - Definitely would choose B)

3.A. Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
If Program A is adopted, 200 people will be saved.
If Program B is adopted, there is a 33% chance that 600 people will be saved, and a 67% chance that no people will be saved.
Which program do you recommend to use?
(1 – Definitely would choose A 6 - Definitely would choose B)
3.B. Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
If Program A is adopted, 400 people will die.
If Program B is adopted, there is a 33% chance that nobody will die, and a 67% chance that 600 people will die.
Which program do you recommend to use?
(1 – Definitely would choose A 6 - Definitely would choose B)

4.A. The company you work for has decided to give you a bonus of 20,000 kuna. Still, your boss thinks that money should be shared somehow with the rest of the workgroup and offers you to choose between one of two options.

If you accept option A, you lose 4000 of the possible 20,000 kuna.

If you accept option B, you participate in a lottery in which you have an 80% chance of losing nothing and a 20% chance of losing all 20,000 kuna.

Which option would you choose?

(1 – Definitely would choose A 6 - Definitely would choose B)

4.B. The company you work for has decided to give you a bonus of 20,000 kuna. Still, your boss thinks that money should be shared somehow with the rest of the workgroup and offers you to choose between one of two options.

If you choose option A, you get 16,000 out of a possible 20,000 kuna.

If you choose option B, you participate in a lottery in which you have an 80% chance of winning all of the 20,000 kuna and a 20% chance of winning nothing.

Which option would you choose?

(1 – Definitely would choose A 6 - Definitely would choose B)

ICAR

a) Verbal reasoning items

1. What number is one fifth of one fourth of one ninth of 900?

a. 2
b. 3
c. 4
d. 5
e. 6
f. 7

2. Zach is taller than Matt and Richard is shorter than Zach. Which of the following statements would be most accurate?

a. Richard is taller than Matt
b. Richard is shorter than Matt
c. Richard is as tall as Matt

d. It's impossible to tell

3. Joshua is 12 years old and his sister is three times as old as he. When Joshua is 23 years old, how old will his sister be?

a. 35
b. 39
c. 44
d. 47
e. 53
f. 57

4. If the day after tomorrow is two days before Thursday then what day is it today?

a. Friday
b. Monday
c. Wednesday
d. Saturday
e. Tuesday
f. Sunday

b) Letters and numbers series

5. In the following aplhanumeris series, what letter comes next? K N P S U

a. S
b. T
c. U
d. V
e. W
f. X

6. In the following aplhanumeris series, what letter comes next? V Q M J H

a. E
b. F
c. G
d. H
e. I
f. J

7. In the following aplhanumeris series, what letter comes next? I J L O S

a. T
b. U
c. V
d. X
e. Y
f. Z

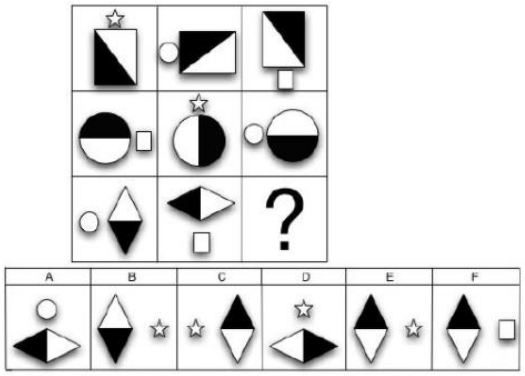
8. In the following aplhanumeris series, what letter comes next? Q S N P L

a. J
b. H
c. I
d. N
e. M
f. L

c) Matrix reasoning

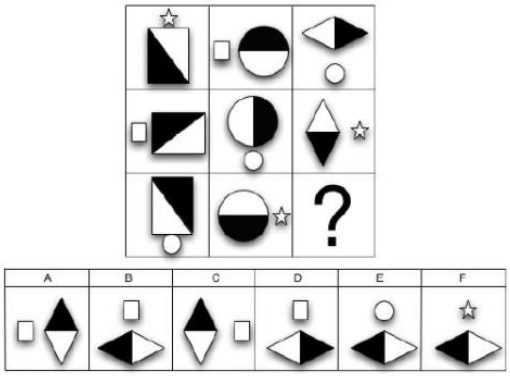
Please indicate which of the six options provided best completes the picture.

9.



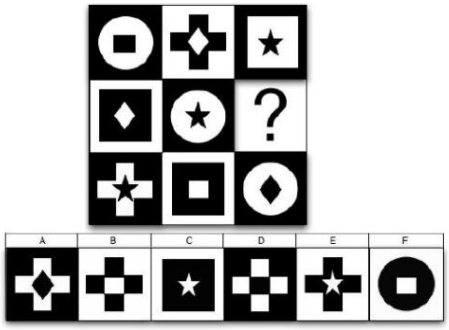
Correct answer: E

10.



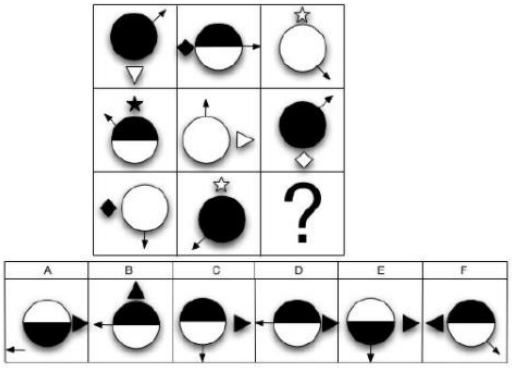
Correct answer: B

11.



Correct answer: B

12.

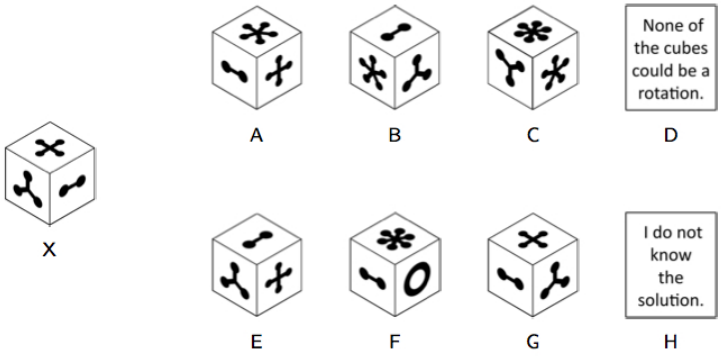


Correct answer: D

e) Three-dimensional rotations

13.

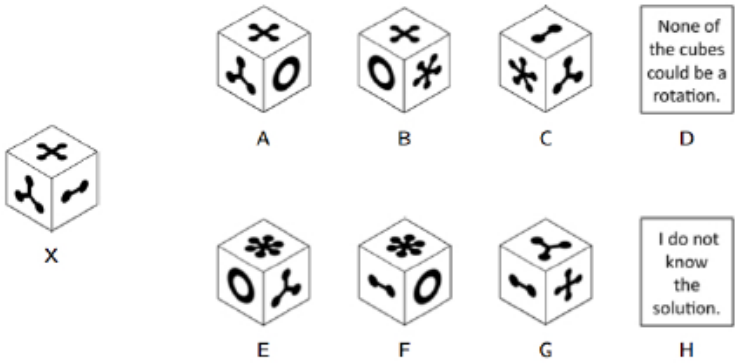
All the cubes below have a different image on each side.
Select the choice that represents a rotation of the cube labeled X.



Correct answer: C

14.

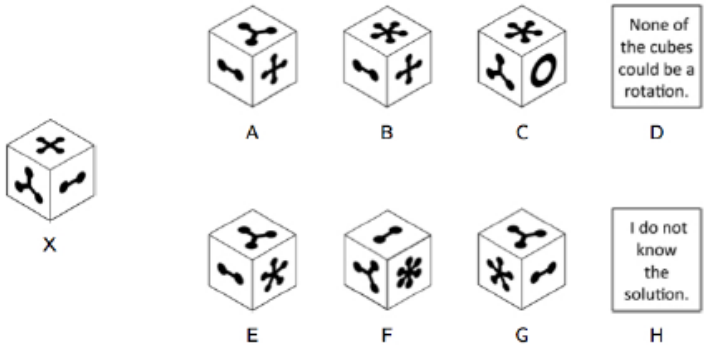
All the cubes below have a different image on each side.
Select the choice that represents a rotation of the cube labeled X.



Correct answer: B

15.

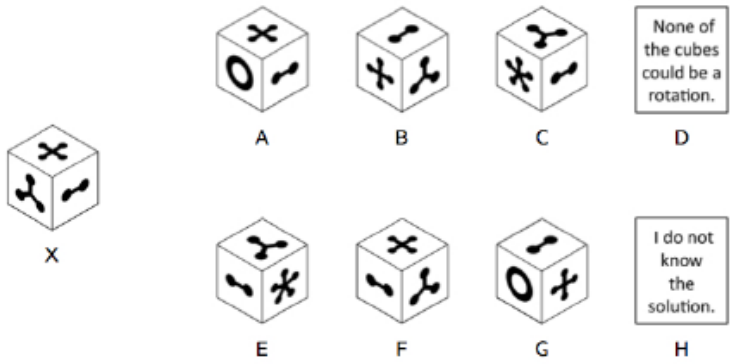
All the cubes below have a different image on each side.
Select the choice that represents a rotation of the cube labeled X.



Correct answer: F

16.

All the cubes below have a different image on each side.
Select the choice that represents a rotation of the cube labeled X.



Correct answer: G

Berlin numeracy test

1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent. Correct response: **25 %**
2. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? Correct response: **30** out of 50 throws.
3. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6? Correct response: **20** out of 70 throws.
4. In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? Correct response: **50**

Cognitive reflection test

1. A bat and a ball together cost 110 kunas. The bat costs 100 kunas more than the ball. How much does the ball cost? Correct: 5; Lure: 10.
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? Correct: 5; Lure: 100.

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 the lake? Correct: 47; Lure: 24.
4. Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has:
    a. broken even in the stock market,
    b. is ahead of where he began, (lure)
    c. has lost money (correct)
5. In an athletic team, tall athletes are three times more likely to win a medal than short athletes. This year the team has won 60 medals so far. How many of those medals were won by short athletes? Correct: 15; Lure: 20.
6. A square shaped garage roof with 6 meters long edge is covered with 100 tiles. How many tiles of the same size are covering a neighbouring roof, which is also square shaped, but with a 3 meters long edge? Correct: 25; Lure: 50.

Actively open-minded thinking

1. There are two kinds of people in this world: those who are for the truth and those who are against the truth. *
2. Changing your mind is a sign of weakness. *
3. I believe we should look to our religious authorities for decisions on moral issues. *
4. No one can talk me out of something I know is right.
5. Basically, I know everything I need to know about the important things in life.
6. Considering too many different opinions often leads to bad decisions. *
7. There are basically two kinds of people in this world, good and bad. *
8. Most people just don't know what's good for them.
9. It is a noble thing when someone holds the same beliefs as their parents.*
10. I believe that loyalty to one's ideals and principles is more important than "open-mindedness." *
11. Of all the different philosophies which exist in the world there is probably only one which is correct. *
12. One should disregard evidence that conflicts with your established beliefs. *
13. I think that if people don't know what they believe in by the time they're 25, there's something wrong with them.

14. I believe letting students hear controversial speakers can only confuse and mislead them.
15. Intuition is the best guide in making decisions.
16. It is important to be loyal to your beliefs even when evidence is brought to bear against them.*
17. People should revise their conclusions in response to relevant new information. *
18. People should take into consideration evidence that goes against conclusions they favor. *
19. Certain beliefs are simply too important for us to renounce them no matter how good the evidence is against them. *

    * means that the items is used in Study 2

Conspiracy thinking

1. The power held by heads of states is second to that of small unknown groups who really control the world's politics.
2. The spread of certain viruses and/or diseases is the result of the deliberate, concealed efforts of some organization.
3. Groups of scientists manipulate, fabricate or suppress evidence in order to deceive public.
4. The government permits of perpetrates acts of terrorism on its own soil, disguising its involvement.
5. A small, secret group of people is responsible for making all major world decisions, such as going to war.
6. Evidence of alien contact is being concealed from the public.
7. Technology with mind-control capacities is used on people without their knowledge.
8. New and advanced technology which would harm current industry is being suppressed.
9. Certain significant events have been the result of the activity of small group who secretly manipulate world events.
10. The government uses people as patsies to hide its involvement in criminal activity.
11. Experiments involving new drugs or technologies are routinely carried out on the public without their knowledge.
12. A lot of important information is deliberately concealed from the public out of self-interest.

Superstitious thinking

1. I have found that talking about successes that I am looking forward to can keep them from happening.
2. I do not believe in any superstitions.

3. When something good happens to me, I believe it is likely to be balanced by something bad.
4. I have personal possessions that bring me luck at times.
5. The number 13 is unlucky.
6. It is bad luck to have a black cat cross your path.
7. Opening an umbrella indoors will increase one's chances of misfortune in the near future.
8. It is advisable to consult your horoscope daily.
9. Astrology can be useful in making personality judgments
10. Some people have the ability to predict the future.
11. Mind reading is not possible.
12. Dreams can provide information about the future.
13. A person's thoughts can influence the movement of a physical object.

Actively open-minded thinking situational judgment test

1. You were recently promoted to the position of HR Manager of a large company. Management expects you to make some changes to motivate employees. An older colleague, a long-term employee of the human resources department, believes based on his practice and experience that rewarding employees according to performance is the best way to motivate them. This sounds like a good idea to you too - it makes sense to you that people will work harder if they are paid according to the work they did and you don't see any objective disadvantage of this method. What will you do?

a) You can't see any major drawbacks to this approach, so you'll be introducing a performance-based reward system as soon as possible. This will increase employee motivation and at the same time meet the requirements of management.
b) You will talk to an older colleague who advocates this system and knows more about it than you do. If his reasons for introducing this approach are reasonable and good, you will implement it immediately.
c) You will engage and try to find on the internet what other experts think about why such a system should be introduced.
d) Although it seems that this approach is generally supported, you will do your best to find key arguments against it or identify possible problems with the introduction of this human resource management practice.

2. You are the owner of a tourist agency that makes a smaller part of its bookings online, and a larger one through the branch office it has in the city center. Your business went very well last season and

now you are considering renting another space in the city center and turning it into another branch office to further increase the number of customers. You are convinced that this move would double your customer numbers and greatly increase your earnings. However, the spaces in the city center are very expensive (around 10,000 Euros per month), so if the new branch fails, you could find yourself in serious trouble. You currently have 15 employees in the company some of whom have expressed support for the idea while others are reserved. Also, you have an acquaintance in a similar industry who has already decided on a similar move and is convinced he is right. What are you going to do?

a) You are convinced that renting additional space is the right decision, so you are going to rent it. You feel that there is no point in procrastinating too much with the decision.
b) You will consult with an acquaintance who already rented additional space and did not regret it and ask him to explain why he decided to make such a decision.
c) You will try to do a little research on your own about the predictions on the incoming tourists in the next year, put all available figures on paper and make a decision.
d) You know that some of your employees disagree about renting a new branch. Before making a decision, you should talk to your employees and you are especially interested in the arguments of employees who are skeptical about renting a new space.

3. You are the director of a large state-owned company with over 3,000 employees, and the country is facing an economic crisis. The association of private business owners and certain liberal economic circles are suggesting that you should lay off a large number of "unnecessary" workers in your company in order to relieve taxpayers and increase the productivity of other workers. At the same time, the unions strongly oppose the layoffs, believing that in this way the state will end up in an even bigger economic crisis. You have your relatively firm position on this issue and if you decided to do it your way tomorrow, you would have ready arguments for your position. In a week, you must announce your decision at a press conference. What are you going to do?

a) You have a clear position on this issue and strong enough arguments to defend it. That is why you will simply announce at the press conference that you have decided to act in accordance with your position.
b) You believe business owners and liberal economists know best what's good for the economy. You will meet with them and listen to their arguments before making a decision.

c) You think that the unions care about the well-being of the largest number of people in this situation and that their arguments should be listened to. You will meet with them before making a decision.
d) While this may significantly postpone your decision, you will meet with both groups to hear arguments for and against your own position.

Mini IPIP

Extraversion

1. Am the life of the party.
2. Talk to a lot of different people at parties.
3. Don't talk a lot.
4. Keep in the background.

Agreeableness

5. Sympathize with others' feelings.
6. Feel others' emotions.
7. Am not really interested in others.
8. Am not interested in other people's problems.

Conscientiousness

9. Get chores done right away.
10. Like order.
11. Often forget to put things back in their proper place.
12. Make a mess of things.

Emotional stability

13. Have frequent mood swings.
14. Get upset easily.
15. Am relaxed most of the time.
16. Seldom feel blue.

Openness

17. Have a vivid imagination.
18. Have difficulty understanding abstract ideas.
19. Am not interested in abstract ideas.
20. Do not have a good imagination.

Job satisfaction

Think about your current job. Weigh all its advantages and disadvantages and then assess how satisfied you are, on the whole, with your job.

Career satisfaction

1. I am satisfied with the success I have achieved in my career.
2. I am satisfied with the progress I have made in achieving my overall career goals.
3. I am satisfied with the progress I have made in achieving the income I would like to have.
4. I am satisfied with the progress I have made in achieving my goals related to promotions.
5. I am satisfied with the progress in meeting my own goals related to the development of new skills.

Peer-rated decision-making quality

1. The decisions my friend makes are quality ones.
2. The decisions my friend makes end up working out well.
3. The decisions my friend makes are good ones.
4. The decisions my friend makes are regretted later.

Decision-outcome inventory

In the last five years, have you ever…

1. Returned a book you borrowed from the library without reading it? *
2. Bought new clothes or shoes you never wore? *
3. Threw out food or groceries you had bought because they went bad? *
4. Ruined your clothes because you didn’t follow the laundry instructions on the label? *
5. Had your driver’s license taken away from you by the police? *
6. Been accused of causing a car accident while driving?
7. Gotten a parking ticket?
8. Missed a flight train or a bus? *
9. Taken the wrong train or bus?
10. Had your ID, driver's licence or student's ID replaced because you lost it?
11. Had the key to your home replaced because you lost it?

12. Been kicked out of a bar, restaurant, or hotel by someone who works there?
13. Loaned more than 100 HRK to someone and never got it back?
14. Cheated on your romantic partner?
15. Consumed so much alcohol you vomited?
16. Got blisters from sunburn?
17. Been in a public fight or screaming argument? *
18. Forgotten a birthday of someone close to you and did not realize until the next day or later? *
19. Broke a bone because you fell, slipped, or misstepped?
20. Missed an exam because you fell asleep?
21. Missed an exam because you did not learn enough for it?
22. Had to borrow money because you irrationaly spent yours? *
23. Been late with turning in seminars, reports or assignments?
24. Forgot that you had to meet with your friend or partner? *
25. Started going to gym or fitness center but giving up very soon?
26. Unsuccessfully tried to change your diet or go on a diet? *
27. Lost more than 200 HRK betting?
28. Bought bad birthday present because you proxrastinated with it until last minute? *
29. Had to do additional assignments because you missed too many classes?
30. Fell asleep during classes multiple times?
31. Got kicked out of the class for whatever reason?
32. Been sick because you ate too much? *
33. Lost a contact with a person you liked because you continuosly forgot to get in touch? *
34. Took an unfavorable short-term loan? *
35. Spent more money in a month than you had available? *
36. Been paying a subscription for a product or service for a long time because you forgot to cancel it? *
37. Failed to pay the loan installment for an apartment, house or car on time. *

* Items used in Study 2.

Study 4 instruments

General decision making scale (if the instruments are used in the study, but not reported here, it means that they are the same as the ones already reported above)

Rational decision-making style

1. I explore all of my options before making a decision.
2. I double-check my information sources to be sure I have the right facts before making decisions.
3. I make my decisions in a logical and systematic way.
4. My decision-making requires careful thought.
5. When making a decision, I consider various options in terms of a specific goal.

Intuitive decision-making style

1. When making decisions, I rely upon my instincts.
2. When I make decisions, I tend to rely on my intuition.
3. I generally make decisions that feel right to me.
4. When I make a decision, it is more important for me to feel the decision is right than to have a rational reason for it.
5. When I make a decision, I trust my inner feelings and reactions.

Dependent decision-making style

1. I often need the assistance of other people when making important decisions.
2. I rarely make important decisions without consulting other people.
3. If I have the support of others, it is easier for me to make important decisions.
4. I use the advice of other people in making my important decisions.
5. I like to have someone to steer me in the right direction when when I am faced with important decisions.

Avoidant decision-making style

1. I avoid making important decisions until the pressure is on.
2. I postpone decision making whenever possible.
3. I often procrastinate when it comes to making important decisions.
4. I generally make important decisions at the last minute.
5. I put off making many decisions because thinking about them makes me uneasy.

Spontaneous decision-making style

1. I generally make snap decisions.
2. I often make decisions on the spur of the moment.
3. I make quick decisions.
4. I often make impulsive decisions.

5. When making decisions, I do what seems natural at the moment.

Counterproductive academic behavior

1. Missed exam because you overslept.
2. Missed an exam because you failed to study for it.
3. Missed the deadline for project assignments.
4. Had to do an additional assignment because you missed too many lectures.
5. Overslept classes multiple times.
6. Been kicked out of a classroom for whatever reason.

In-role performance

1. Adequately completes assigned duties.
2. Fulfills responsibilities specified in job description.
3. Performs tasks that are expected of him/her.
4. Meets formal performance requirements of the job.
5. Engages in activities that will directly affect his/her performance evaluation.
6. Neglects aspects of the job he/she is obliged to perform.
7. Fails to perform essential duties.

Counterproductive work behavior

1. Purposely wasted your employer's materials/supplies
2. Complained about insignificant things at work
3. Told people outside the job what a lousy place you work for
4. Came to work late without permission
5. Stayed home from work and said you were sick when you weren't
6. Insulted someone about their job performance
7. Made fun of someone's personal life
8. Ignored someone at work
9. Started an argument with someone at work
10. Insulted or made fun of someone at work

Peer-rated general decision-making quality

1. The decisions my friend makes are quality ones.
2. The decisions my friend makes end up working out well.
3. The decisions my friend makes are good ones.
4. The decisions my friend makes are regretted later.

Need for achievement

1. Maintaining high standards for the quality of my work.
2. Personally producing work of high quality.
3. Projects that challenge me to the limits of my ability.
4. Continuously improve myself.
5. Continuously engage in new, exciting, and challenging goals and projects.
6. Opportunities to take on more difficult and challenging goals and responsibilities.

Perceived entrepreneurial efficacy

How would you say your employer compares to others in terms of following tasks?

1. Searching for opportunities
2. Creating new products
3. Thinking creatively
4. Commercializing ideas and new products
5. Fund raising
6. Selling new products or services
7. Solving other people's problems
8. Finding new ways to solve problems
9. Imagining different ways of thinking and doing

Perceived organizational support

1. The organization values my contribution to its well-being.
2. The organization fails to appreciate any extra effort from me.
3. The organization would ignore any complaint from me.
4. The organization really cares about my well-being.
5. Even if I did the best job possible, the organization would fail to notice.
6. The organization cares about my general satisfaction at work.

7. The organization shows very little concern for me.
8. The organization takes pride in my accomplishments at work.

Job satisfaction index

1. I feel fairly well satisfied with my present job.
2. Most days I am enthusiastic about my work.
3. Each day of work seems like it will never end.
4. I find real enjoyment in my work.
5. I definitely dislike my work.

Intentions of leaving organization

1. In the next year, I intend to look for a job outside the organization in which I am currently employed.
2. I intend to remain in this organization indefinitely.
3. I often think about quitting.

APPENDIX B

Appendix B consists of supplementary analyses and additional explanations accompanying our main manuscript.

Study 1 supplementary material

Table B1. Fit indices of CFA analyses test appropriateness of one-factor solutions of our measure

	$\chi 2$	df	CFI	TLI	RMSEA	SRMR	N	Estimator
CRT	36.35	35	1	1	.01	.03	506	DWLS
BBS	10.54**	2	.99	.98	.09	.04	506	DWLS
NUM	0.27	2	1	1	.00	.01	506	DWLS
VR	1.43	2	1	1	.00	.02	506	DWLS
AOT	261.34**	90	.87	.85	.06	.05	469	ML
BRN	6.19	5	1	1	.03	.04	253	DWLS
FCS	5.21	5	1	1	.01	.04	253	DWLS
CBR	Just 3 variables, i.e. perfect fit							
GF	5.16	5	1	.99	.01	.05	253	DWLS
AV	12.00**	2	.92	.77	.14	.09	253	DWLS
AV +	0.07	1	1	1	.00	.00	253	DWLS

[+] after allowing the first two items to covary as they are both related to diabetes

Table B2. “Lureness” of our CRT items

Item	Lureness
CRT1	.86
CRT2	.64
CRT3	.73
CRT4	.57
CRT5	.81
CRT6	.84
CRT7	.78
CRT8	.81
CRT9	.78
CRT10	.70

Study 2 supplementary material

Study 1

Multilevel regression Analyses using alternative conflict detection indicators and cut-off points for conflict detection

CRT analyses

In addition to the analysis reported in the manuscript (response time as a conflict-detection indicator and 20% of the fastest responders categorized as non-detectors), we have also conducted three additional analyses: a) with response time as a conflict-detection indicator and **10%** of the fastest responders categorized as non-detectors; b) with response-time difference between lure and no-lure tasks with **+/-2 seconds** being the cut-off point for non-detection; c) with response-time difference between lure and no-lure tasks with **+/-3 seconds** being the cut-off point for non-detection.

*1) Response time as a conflict-detection indicator and **10%** of the fastest responders categorized as non-detectors*

The results of this analysis are shown in the Table B3.

Table B3. Results of the multilevel logistic regression analyses for the Cognitive reflection test tasks with 10% of the fastest respondents categorized as non-detectors

	Correct non-detection vs. correct detections			Correct detections vs. incorrect detections			Incorrect detections vs. incorrect non-detections		
	B	SD	OR	B	SD	OR	B	SD	OR
ICAR	0.06	0.05	1.06	0.38***	0.09	1.46	0.08	0.11	1.08
NUM	0.32*	0.13	1.38	0.70***	0.002	2.01	0.10	0.31	1.11
AOT	0.25	0.24	1.28	0.56	0.46	1.75	0.13	0.48	1.14
Matura	-0.02	0.10	0.98	0.39 .	0.20	1.48	0.12	0.21	1.13

Note. Outcome variables are coded such that first category (e.g. correct non-detections) is coded as 1 and second category (e.g. correct detections) is coded as 0.

SD = Standard deviation; OR = Odds ratio; ICAR = International Cognitive Ability Resource; NUM = Numeracy; AOT = Actively open-minded thinking. p < .10; * p < .05; ** p < .01; *** p > .001

The results that are the same as ones presented in the manuscript (in terms of effects being significant or non-significant) are highlighted in yellow.

2) Response time difference as a conflict detection index

Response time difference is calculated as a difference between the response time on original item and response time on the control item. Positive values here should indicate a conflict detection, while no difference should indicate lack or very weak conflict detection. The problem here is that, expectedly, there were virtually none of the cases where the difference between the response times on original and control item was exactly 0 seconds. Therefore, we again faced a need to make an arbitrary decision on where to draw the line for conflict detection, or more precisely, how wide should our interval around 0 seconds be. For example, a person that generally solved the original and the control items at the same speed and with the same confidence, therefore not showing any conflict detection signs, will probably sometimes just by chance spend somewhat more time on control than on the original items, having a negative response time difference. However, this negative value should not be too large as there is no rational reason for why someone would take much more time to solve control than original item. In other words, large negative values are hard to explain as they are probably result of issues such as technical problems or laps of attention. Therefore, large negative values are best to be discarded from further analyses. This brings us back again to our arbitrary decision. With this decision, the logic should be the same as with response time – the narrower the interval, the higher chance that it will mostly contain trials on which participants showed very little signs of conflict detection, especially in conjunction with no confidence decrease in responses from control to original item. We have again decided to go with two different cut-off points, or in this case intervals, to see whether and how this decision affects our results. These intervals are +/- 2 and +/- 3 seconds. As we noted earlier, these are arbitrary decisions, but we believe that these intervals are conservative. For example, the median response time for our control CRT items ranged between 13.8 seconds and 44.2 seconds. The two or three seconds difference in response time between the original and control item represent only a small fraction of time that it took to read and solve easy control items. Therefore, it is plausible that someone who solved these items roughly at the same speed would sometimes end up solving one item up to two or three seconds faster or slower compared to the other item.

Thus, those trials which did not show confidence difference in responses between control and original items AND whose response time differences were between -2 and 2 seconds (more strict classification)

and between -3 and 3 seconds (less strict classification) respectively were classified as conflict non-detection trials. In the Table B4, we are showing the frequencies of trials in each of our four categories, both for +/- 2 and +/-3 seconds cut-off interval.

Table B4. Frequencies of the CRT trials based on accuracy and conflict detection for two different cut-off intervals for conflict detection, +/- 2 seconds and +/- 3 seconds.

	N (+/- 2 seconds)	N (+/- 3 seconds)
Correct non-detection	122	184
Correct detection	730	705
Incorrect non-detection	49	90
Incorrect detection	478	456
Total	1151	1435

*2a) Results of multilevel logistic regression analyses with response-time difference as conflict-detection indicator and +/-**2 seconds** being the cut-off point for non-detection*

The results of this analysis are shown in the Table B5.

Table B5. Results of the multilevel logistic regression analyses for the Cognitive reflection test tasks with response-time difference as conflict-detection indicator and **+/-2 seconds** being the cut-off point for non-detection

	Correct non-detection vs. correct detections			Correct detections vs. incorrect detections			Incorrect detections vs. incorrect non-detections		
	B	SD	OR	B	SD	OR	B	SD	OR
ICAR	0.01	0.04	1.01	0.29***	0.05	1.34	0.11*	0.05	1.12
NUM	0.14	0.09	1.15	0.76***	0.004	2.14	0.12	0.15	1.13
AOT	0.001	0.15	1.001	0.27	0.71	1.31	0.13	0.24	1.14
Matura	0.02	0.06	1.02	0.56***	0.003	1.75	0.05	0.09	1.05

Note. Outcome variables are coded such that first category (e.g. correct non-detections) is coded as 1 and second category (e.g. correct detections) is coded as 0.

SD = Standard deviation; OR = Odds ratio; ICAR = International Cognitive Ability Resource; NUM = Numeracy; AOT = Actively open-minded thinking. p < .10; * p < .05; ** p < .01; *** p > .001

The results that are the same as ones presented in the manuscript (in terms of effects being significant or non-significant) are highlighted in yellow.

*2b) Results of multilevel logistic regression analyses with response-time difference as conflict-detection indicator and +/-**3 seconds** being the cut-off point for non-detection*

The results of this analysis are shown in the Table B6.

Table B6. Results of the multilevel logistic regression analyses for the Cognitive reflection test tasks with response-time difference as conflict-detection indicator and **+/-3 seconds** being the cut-off point for non-detection

	Correct non-detection vs. correct detections			Correct detections vs. incorrect detections			Incorrect detections vs. incorrect non-detections		
	B	SD	OR	B	SD	OR	B	SD	OR
ICAR	0.01	0.03	1.01	0.26***	0.05	1.30	0.07 .	0.04	1.07
NUM	0.24**	0.08	1.27	0.79***	0.15	2.20	0.18	0.12	1.20
AOT	0.08	0.13	1.08	0.40	0.26	1.49	0.06	0.18	1.06
Matura	0.05	0.05	1.05	0.28**	0.11	1.32	0.02	0.07	1.02

Note. Outcome variables are coded such that first category (e.g. correct non-detections) is coded as 1 and second category (e.g. correct detections) is coded as 0.

SD = Standard deviation; OR = Odds ratio; ICAR = International Cognitive Ability Resource; NUM = Numeracy; AOT = Actively open-minded thinking. p < .10; * p < .05; ** p < .01; *** p > .001

The results that are the same as ones presented in the manuscript (in terms of effects being significant or non-significant) are highlighted in yellow.

BBS analyses

In addition to analyses that we report in the manuscript (response time as conflict-detection indicators with 20% of the fastest respondents categorized as non-detectors), we conducted additional analyses with 10% of the fastest respondents categorized as non-detectors. The results of these multilevel logistic regression analyses are presented in the Table B7.

Table B7. Results of the multilevel logistic regression analyses for the Belief bias syllogism tasks with response time as conflict-detection indicator and 10% of the fastest respondents categorized as non-detectors

	Correct non-detection vs. correct detections	Correct detections vs. incorrect detections	Incorrect detections vs. incorrect non-detections

	B	SD	OR	B	SD	OR	B	SD	OR
ICAR	-0.03	0.06	0.97	0.35*	0.14	1.42	0.09*	0.04	1.09
NUM	0.12	0.16	1.13	0.94*	0.41	2.56	0.02	0.13	1.02
AOT	-0.20	0.29	0.82	0.83	0.58	2.29	0.01	0.19	1.01
Matura	0.16	0.12	1.17	0.39	0.27	1.48	0.12	0.08	1.13

Note. Outcome variables are coded such that first category (e.g. correct non-detections) is coded as 1 and second category (e.g. correct detections) is coded as 0.

SD = Standard deviation; OR = Odds ratio; ICAR = International Cognitive Ability Resource; NUM = Numeracy; AOT = Actively open-minded thinking. p < .10; * p < .05; ** p < .01; *** p > .001

The results that are the same as ones presented in the manuscript (in terms of effects being significant or non-significant) are highlighted in yellow.

Study 2

a) General instructions

Please read the following information carefully.

You will be solving several types of reasoning tasks. Before each set of tasks, you will be given two exercise tasks to that will prepare you for the actual tasks.

For each of the tasks we want to know: a) what is your first, intuitive answer and b) what is your answer to that same task after you have had enough time to think about it.

To find that out, you will be solving the same tasks twice:

The first time, we will ask you to answer as quickly as possible and without thinking, that is, to check the response that intuitively came to your mind without thinking about the problem.

Right after that, we will show you the same task, but this time you will have as much time as you want to think and check the response that you think is correct.

In addition, after both answers, both fast and slow, we will ask you how confident you are in your responses.

In short, make sure that the first answer you give is always quick, intuitive and without thinking, because right after you will be able to think further about the problem and change your response, if necessary.

The response time for the first problem will always be very short and limited, and when it expires, the program will automatically move on to the next task. So try to give your first response as quickly as possible so that you manage to solve the task without the time running out.

b) CRT specific instructions

You will now solve several tasks consisting of a problem and four response options. Your task will be to choose the response that you think is correct. The tasks will look something like this:

Marko has 40 kuna in his pocket. If he gives 10 kunas to his friend Tomislav, how many kunas will Marko have left in his pocket?

a) 0

b) 10

c) 20

d) 30

The correct answer is obviously d) 30 kuna.

Some tasks will be a bit easier, and some will be a bit harder, but the structure will always be the same: problem with four response options. You will solve each task twice. The first time quickly, intuitively and without thinking, and the second time with enough time to think and change answers.

To make sure that the first answer is truly intuitive, each task is very limited in time - you will have just enough to read the task and choose the answer.

In addition, we will show you one image that you will need to remember and that you will need to recognize among the similar images shown after giving a quick intuitive response. It is very important that you take this task seriously - although memory tasks will not be very difficult, if you answer them incorrectly, your responses to problem tasks will not count. We are interested in intuitive responses only from those people who are able to memorize and recognize the image.

At any time, you can see how much time you have left to respond so you can hurry up and answer every question. Answer as quickly as possible so that you manage to answer each task.

To recap, after each quick, intuitive response, we will ask you how confident you are in your response and to recognize the image that was shown to you at the beginning. After that, you will solve the same

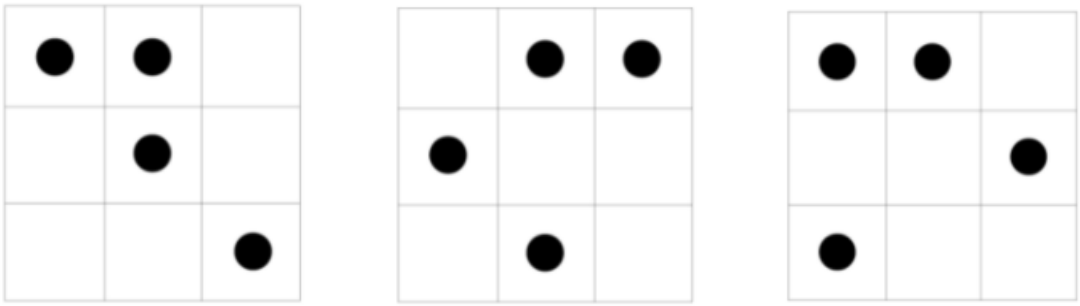
task again, only this time with no time limit and no memory task. Once you have chosen your final answer, we will ask you again how confident you are in it.

To make the whole process more understandable, before you start solving the real tasks, you will first solve two exercise tasks. By solving these tasks, you will get a sense of how little time you have on first response and what the images you need to remember look like.
If you are ready to continue, click on the "Next" butt and you will be presented with the first exercise task.

c) BBS specific instructions
In this part of the survey, you will solve several logical tasks. Tasks will always be similar in structure to this one:
Premise 1: All dogs have four legs.
Premise 2: Puppies are dogs.
Conclusion: Puppies have four legs.
Does the conclusion follow logically from the premises?
a) Yes
b) No.

In the previous case, the answer is Yes - the conclusion logically follows from the premises. So, your task is to indicate whether the conclusion is logically and unambiguously derived from the premises, whether or not it sounds convincing.
Similar to the previous tasks, you will respond to each task twice. For the first response, we want you to give an answer as quickly as possible and without thinking. During the first response, time will be limited and quite short, and at the same time you will have to solve memory task accurately. During the second response, you will be able to take time to think before giving the final response and you no memory task will be presented. Like the first time, both after a quick and after a slow response, we will ask you how confident you are in it.
If you are ready, you can start solving two exercise tasks by clicking "Next". This will help you to familiarize with the tasks and the time limit before moving on to the real tasks.

Memory task matrix examples



Study 3 supplementary material

Scree plots and parallel-analysis results

a) Study 1

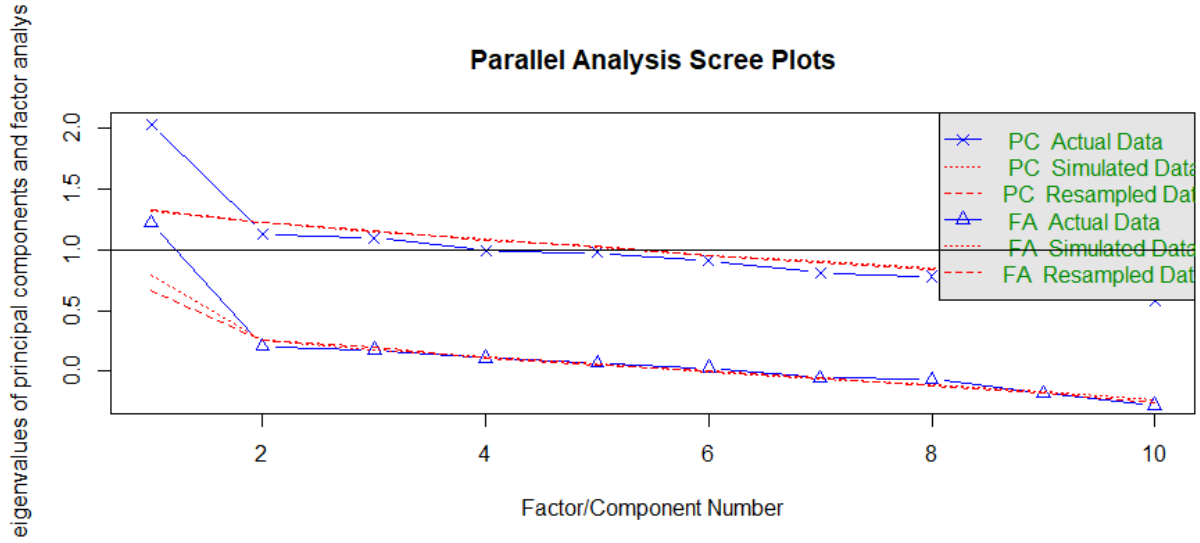


Figure B1. Output from the parallel analysis on 10 cognitive biases from Study 1, done using the fa.parallel () function from the "psych" R package (Revelle, 2021). Both scree plot and parallel analysis indicate a one-factor solution.

b) Study 2

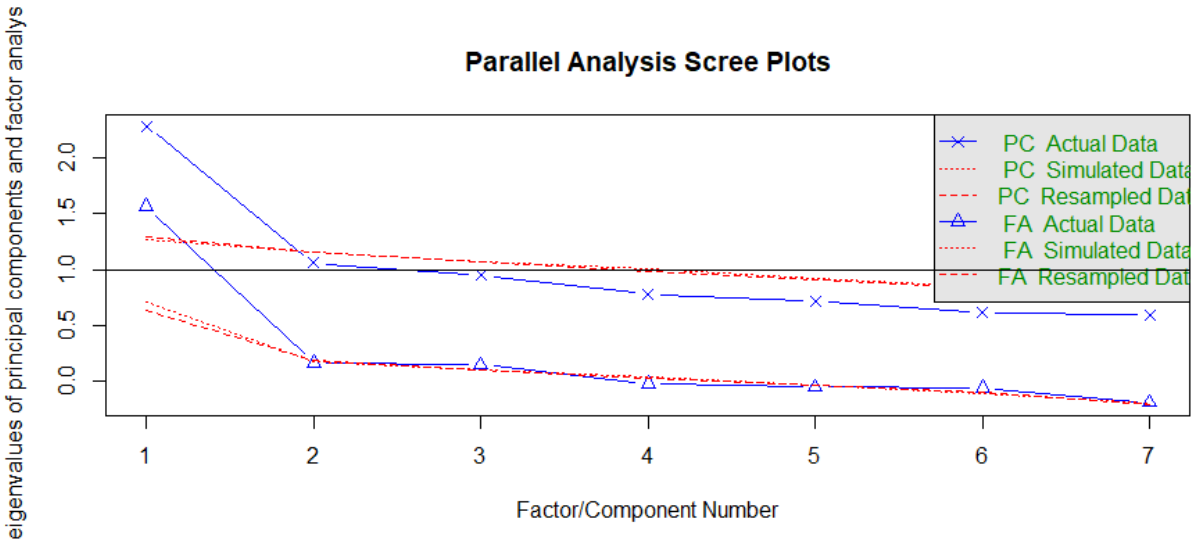


Figure B2. Output from the parallel analysis on seven cognitive biases from Study 2, done using the fa.parallel () function from the "psych" R package (Revelle, 2021). Both scree plot and parallel analysis indicate a one-factor solution.

SEM regression analyses

a) Study 1

Before conducting regression analysis using SEM, we checked the fit of our measurement model. Specifically, we defined each of our four latent variables (Rationality, CRT, ICAR and AOT) by their respective manifest variables: Rationality factor was defined as a second order factor of ten cognitive bias factors that were each defined by their corresponding manifest variables, while CRT, ICAR numeracy and AOT factors were defined as first order factors by their respective manifest variables, i.e. six CRT items, 16 ICAR items and 15 AOT items.

The initial fit of the Rationality factor (with all the covariances between manifest variables set at 0) was mediocre, with some fit indices suggesting relatively good fit (e.g. *RMSEA* = 0.05, *SRMR* = 0.07) and others suggesting poor fit (e.g. *CFI* = 0.77, *TLI* = 0.75). Therefore, we redefined the model by allowing for some theoretically justifiable covariances to be freely estimated. First, as we had two different types of BBS tasks, those where conclusions were believable, but logically incorrect and those where conclusions were unbelievable, but logically correct, we allowed covariations among the four tasks of both kind. Second, as we has two somewhat different types of tasks within four-cards selection tasks, namely two deontic and two non-deontic tasks, we allowed the covariances between the two deontic and two non-deontic tasks to be freely estimated. Third, two of our availability tasks were related to diabetes, thus we allowed them to covariate. Finally, we also allowed the covariation

between our two latent variables related to framing, attribute framing and risk framing, to be freely estimated. These modifications substantially improved all fit indices, bringing them in the are of good or acceptable fit (*RMSEA* = 0.03, *SRMR* = 0.06, *CFI* = 0.89, *TLI* = 0.88).

The other three factors showed excellent fits. The fit indicators for the CRT factor defined by the six CRT variables with no covariations between them were *RMSEA* = 0.00, *SRMR* = 0.03, *CFI* = 1, *TLI* = 1, and for the ICAR factor (after allowing for the covariations among four tasks comprising each of the four ICAR subtests to be freely estimated) *RMSEA* = 0.00, *SRMR* = 0.04, *CFI* = 1, *TLI* = 1. AOT factor defined by the 15 manifest AOT variables with no covariations between them had somewhat worse, but still acceptable fit: *RMSEA* = 0.06, *SRMR* = 0.06, *CFI* = 0.88, *TLI* = 0.88.

After establishing our measurement model, we regressed the Rationality factor on CRT, ICAR and AOT factors using SEM. The beta ponders of this regression analysis are shown in the Table B8 below.

Table B8. SEM regression results with Rationality factor as an outcome and CRT, ICAR and AOT factors as predictors

Variable	Beta
CRT	0.66*
ICAR	-0.04
AOT	0.33**
R^2	0.61

Note. ** p < .01; * p < .05

b) Study 2

Similarly as in the Study 1, before doing the regression analysis, we checked the fit of our measurement model. We defined rationality factor as a second order factor of seven CB factors defined by their respective manifest variables (as we now had only one type of BBS tasks and did not measure risk framing and availability bias, the only theoretically justified covarion that we allowed to be freely estimated was between different types of four-card selection task, as described above). However, one attribute framing item turned out to be problematic (item number three) having negative estimated variance. Therefore, we removed it and defined the attribute framing latent variable with three, instead of four manifest variables. Defined in this way, rationality factor showed a satisfactory fit to the data (*RMSEA* = 0.04, *SRMR* = 0.06, *CFI* = 0.93, *TLI* = 0.92). CRT factor was again defined by its six corresponding manifest variables and showed a good fit to the data (*RMSEA* = 0.06, *SRMR* = 0.04,

CFI = 0.97, *TLI* = 0.95). AOT was first defined by its 13 manifest variables with no covariations among them, but this fit was poor (*RMSEA* = 0.10, *SRMR* = 0.07, *CFI* = 0.83, *TLI* = 0.79). Modification indices suggested that by allowing the items 12 and 13 and items 1 and 5 to freely covary, the fit of the model could be significantly improved. When looking into the content of these items, it became clear why this might be the case. Items 12 and 13 are specifically related to the proper treatment of new information and evidence (i.e. "People should revise their conclusions in response to relevant new information." and "People should take into consideration evidence that goes against conclusions they favor."), while items 1 and 5 both refer to dichotomous thinking about the world (i.e. "There are two kinds of people in this world: those who are for the truth and those who are against the truth." and "There are basically two kinds of people in this world, good and bad."). When allowing these two covariances to be freely estimated in the model, the model showed much better fit to the data (*RMSEA* = 0.06, *SRMR* = 0.05, *CFI* = 0.94, *TLI* = 0.93).

After defining our measurement model, we regressed the rationality factor on CRT and AOT factors. The output of this regression analysis is shown in the Table B9.

Table B9. SEM regression results with Rationality factor as an outcome and CRT and AOT factors as predictors

Variable	Beta
CRT	0.73**
AOT	0.22*
R^2	0.75

Note. ** p < .01; * p < .05

Supplemental material for the manuscript "Incremental validity of decision-making styles in predicting work-related outcomes"

Study 4 supplementary material

Study 1

Procedure description

In exchange for course credits, psychology students helped with recruitment for this study by forwarding the sign-up link to participants. Participants who agreed to participate then signed up and chose one of the available time slots for participation in the study. Participants solved our focal tasks

as a part of larger battery of tasks not all of which are reported in this study. Relevant for the current study, participants solved an intelligence test, numeracy test, decision-outcome inventory (DOI) and a questionnaire assessing five different decision-making styles. Students solved tests and questionnaires on computers, in groups of 20 to 25 participants under the supervision of the investigators. The whole testing lasted up to two hours, but was divided in two parts between which there was a 15 minutes break. In the first part, along with other instruments not reported here, participants solved numeracy test and DOI. After this, there was a 15 minutes break followed by a second part of testing. In the second part, participants first solved an intelligence test, followed by several other questionnaires, among which were decision-making styles. The tests and questionnaires relevant for this study were solved in a fixed order. Participation in this study, as well as in the other two, was voluntary, and participants were free to quit at any moment. All of the studies were approved by the Department of psychology, University of Zagreb ethical committee.

Description of instruments

DOI

As our participants were students, some of the original outcomes were extremely unlikely or totally impossible to have happened to them, and were thus removed. For example, we removed some of the most serious negative outcomes for this reason (e.g. been in a jail cell overnight for any reason or got divorced). We also added some outcomes that seems appropriate for college students such as "Overslept classes multiple times" or "Had to do an additional assignment because you missed too many lectures."

The total score was calculated in a following way: first, for some outcomes participants were asked whether they had the opportunity to experience them (for example, someone who does not have a driving license could not have had it seized from him/her – therefore, for these kinds of outcomes, we first asked participants whether they could or could not experience it). Next, to account for the severity of outcomes, possible outcomes were weighted by the proportion of participants who reported not experiencing them (thus, more severe outcomes or the ones that were experienced by less people were weighted more). Finally, a total score was calculating by averaging these weighted scores.

CRT

Probably the best-known item is a bat-and-ball item: "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" The immediate response that comes to mind is

10 cents which is, on a further reflection, incorrect and a correct response is 5 cents. Following the publication of an original three-item test, several studies were published that extended this short form test with additional items (e.g. Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; Thomson & Oppenheimer, 2016; Toplak, West & Stanovich, 2014). Recent validation studies of this test show that it mainly captures cognitive abilities and dispositions related with numerical reasoning (e.g. Attali & Bar Hillel, 2020; Erceg, Galić, & Ružojčić, 2020).

Study 2

Procedure description

Similarly as Study 1, psychology students assisted in the recruitment proces sin exchange for course credits. Their job was to recruit the participant and ensure that the participant understood everything about the study, measurements and data collection process. Unlike Study 1, Study 2 was not conducted in person, but online. It was conducted in two parts, only this time the two parts were on average a week apart. Again, the participants solved a battery of tests and questionnaires that were not all relevant for this study. Of the variables we report here, in a first part of the study, participants solved decision-making styles questionnaire. In the second part, they completed personality questionnaire, DOI, job and career satisfaction scale, in-role job performance and counterproductive work behavior scales. Additionally, we asked our participants to forward a link containing several other-report measures to their peers and ask them if they could rate them on these measures. Of relevance for this study is a peer-rated reason-based and overall decision-making quality (Wood, 2012). In total, 192 of our participants received peer ratings. They were mostly rated by two peers (N = 144), but some were rated by only one peer (N = 39) and several were rated by three peers (N = 9).

Instruments description

DOI

We dropped a number of outcomes that were appropriate for student population, but not so much for adults. We also added several finance-related items based on the ones from Toplak, West and Stanovich (2017), such as “Took out an unfavorable short-term loan.” and “Spent more money in a month than you could afford.” The items were again weighted by the percentage of participants who did not experience given outcome, thus giving more weight to more serious negative outcomes. The total score is average of item scores.

Bayesian Model Averaging analysis explanation

Bayesian Model Averaging (BMA) recognizes that, although the amount of variance explained in the outcome will always be the greatest when all the predictors are included in the model, this type of model will nevertheless overfit the data and generalize poorly to other datasets (van den Bergh et al., 2020). Therefore, BMA tries to recognize the most appropriate model given the data. It does so by calculating the probability of each candidate model given the data and comparing that probability with a model's prior probability. In this case, we set uniform priors for all of our models: as we are agnostic about the most appropriate models for our outcomes, we gave the all the same prior probability. By averaging these comparisons, BMA can inform us about the odds of each candidate model compared to all possible models averaged and these odds are expressed in the model Bayes factor (BFm). In this sense, the Bayes factor (BF) is the strength of evidence in favor of that particular model given the data compared to the averaged model. BFs ranging from 1 to 3 are often interpreted as anecdotal or insufficient evidence, BFs from 3 to 10 as moderate evidence, BFs from 10 to 30 as strong evidence, BFs from 30 to 100 as very strong evidence and BFs greater than 100 as extremely strong evidence.

Apart from informing about the most probable model given the data, BMA can give us the probability or the odds of including each of the candidate predictors in the model. It does so by summing all posterior model probabilities of all models that include a specific predictor and compares them with summed prior probabilities of all the models including that predictor (van den Bergh et al., 2020). For example, imagine we want to calculate the odds of including rational style as predictor in a model predicting DOI. We would sum all the prior probabilities of all the models that include rational style and compare it with summed posterior probabilities (probabilities given the data) of all the models that contain rational style as predictor. If we get that the summed prior probability is 0.5 and the summed posterior probability is 0.90, that the BF would be 18. This would mean that there is strong evidence that rational factor is important predictor of DOI, given the data and taking into account all other possible predictors.

Study 3

Procedure description

Psychology students approached entrepreneurs of small businesses (between 3 and 30 employees) with request for participation in our study. Entrepreneurs who agreed to participate agreed to solve a set of self-report scales and questionnaires (including decision-making styles and some other instruments

measuring entrepreneur's motives not reported in this study), but also to provide up to three contacts of their employees after asking them to participate in our study. Students than approached these employees and gave them a pencil-and-paper questionnaire in an envelope with scales assessing our outcome measures. After filling the questionnaires, employees put them back in the envelope, sealed an envelope and mailed it to investigators' University address. This way they were sure that their employer would not have access to their responses, as we reassured them earlier too. As an incentive for entrepreneurs, we gave them feedback about their own scores, but only on those instruments they solved themselves, and not those on which their employees rated them.

Study 5 supplementary material

Joining samples for the incremental validity analysis

As we had multiple items and variables that were the same across both studies, we were able to combine them and join our two samples. This larger sample then allowed us to conduct incremental validity analyses with greater statistical power. Specifically, we managed to combine samples for AOT measure, Big Five personality traits (after transforming Study 2 ratings that were originally on a seven-point scale to the five-point scale that was used in Study 1), subordinates' ratings of managers decision-making quality and intellectual humility, as well as perceptions of subordinates' job satisfaction and perceived organizational support. When joining sample, we were looking for "common denominator" of both samples, i.e. items that were the same in both samples. This resulted in decision-making quality measure consisting of only three items that were identical across studies (first three items in both studies), which was the only substantial deviation from the measures as they are commonly used. Other measures were the same as the appeared in Study 1, meaning that the joined sample consisted of a 10-item AOT version, Mini IPIP, and job satisfaction and perceived organizational support measures that were the same in both studies. This joint sample had between N = 214 and N = 250 cases, depending on the variable.

Results of the incremental validity analysis

To conduct the incremental validity analysis, we did a SEM regression, regressing the four outcomes on Big five factors and AOT factor simultaneously. SEM regression analysis is done on latent variables that are free from measurement error. This means that prior to calculating beta ponders, we specified a model where each of the latent variables (four outcomes, five personality factors and AOT) were

defined by their respective manifest variables (i.e. the scale items) and where these latent variables were allowed to freely covary. This model showed an acceptable fit to the data (CFI = .84, RMSEA = .06, SRMR = .07). There was one problematic AOT item ("There is nothing wrong with being undecided about many issues") whose loading on the AOT factor was negative. However, as removing this item when specifying the AOT factor did not have any effect on the results of regression analyses, here we report the results with this item included in the scale.

In total we conducted four SEM regression analyses with subordinate ratings of managers' decision making quality and intellectual humility, subordinates' job satisfaction and perceived organizational support as outcomes, and personality traits and AOT as predictors. The results of these analyses are shown in Table B10.

Table B10. Results of SEM regression analyses

	Manager's decision making quality			Manager's intellectual humility			Job satisfaction			Perceived organizational support		
	B	SE	β	B	SE	β	B	SE	β	B	SE	β
Open.	-0.13	0.10	-0.14	-0.07	0.08	-0.09	-0.15	0.09	-0.16*	-0.45	0.16	-0.29**
Consc.	0.01	0.08	0.01	-0.01	0.07	-0.01	-0.12	0.07	-0.14	-0.09	0.13	-0.06
Extra.	-0.12	0.10	-0.13	-0.10	0.09	-0.13	-0.09	0.09	-0.11	0.19	0.17	0.13
Agree.	0.43	0.16	0.33**	0.27	0.13	0.25*	0.30	0.14	0.24*	0.60	0.25	0.28**
Neuro.	-0.04	0.13	-0.03	-0.08	0.12	-0.07	0.04	0.12	0.03	-0.10	0.22	-0.04
AOT	0.19	0.14	0.16	0.23	0.12	0.23*	0.21	0.13	0.18*	0.37	0.22	0.19*
R^2	0.129**			0.122**			0.119**			0.171**		
ΔR^2	0.018			0.042			0.022			0.027		

Note. ** p < .01, * p < .05

Open. = Openness; Consc. = Conscientiousness; Extra. = Extraversion; Agree. = Agreeableness; Neuro. = Neuroticism; AOT = Actively open-minded thinking. R^2 – Total proportion of variance in outcomes explained by all predictors; ΔR^2 – Additional proportion of variance in outcomes explained by AOT after accounting for the effects of Big five factors.

Životopis autora

Nikola Erceg rođen je 14.04. u Ljubljani, Republika Slovenija. Osnovnu i srednju matematičko-prirodoslovnu gimnaziju završio je u Splitu. Preddiplomski studij psihologije upisao je na Sveučilištu u Zadru 2006. godine, a diplomski na Filozofskom fakultetu Sveučilišta u Zagrebu 2009. godine, gdje je 2012. godine stekao zvanje magistra psihologije. Doktorski studij psihologije na Sveučilištu u Zagrebu upisao je 2015. godine. Od 2017. godine radi kao asistent na Katedri za psihologiju rada pri Odsjeku za psihologiju Filozofskog fakulteta u Zagrebu.

Glavni istraživački interesi Nikole Ercega vezani su uz područje prosuđivanja i donošenja odluka te individualnih razlika u racionalnosti, kao i uz područje ekonomske psihologije i financijske pismenosti. Dosad je u koautorstvu objavio 20 znanstvenih članaka u inozemnim i domaćim časopisima te jednu znanstvenu monografiju. Recenzirao je preko 20 radova u inozemnim časopisima, a član je i Društva za prosuđivanje i donošenje odluka (eng. Society for judgment and decision making) te Europske asocijacije za psihologiju ličnosti (eng. Europaen association of personality psychology).

Popis objavljenih radova autora

Znanstveni radovi

- **Erceg, N.**, Galić, Z, & Buljan Šiber, A. (2023). Testing the Theory of Good Thinking and Deciding in Organizational Setting: Many Benefits of Leader's Actively Open-minded Thinking. *Studia Psychologica*.
- **Erceg, N.,** & Galić, Z. (2023). Incremental Validity of Decision-Making Styles in Predicting Real-Life and Work-Related Outcomes. *Journal of Individual Differences.* https://doi.org/10.1027/1614-0001/a000404
- Ruggeri, K., Panin, A., Vdovic, M., Većkalov, B., Abdul-Salaam, N., Achterberg, J., ..., **Erceg, N.,** … & Toscano, F. (2022). The globalizability of temporal discounting. *Nature Human Behaviour, 6*(10), 1386-1397. https://doi.org/10.1038/s41562-022-01392-w
- Šćeta, L., Sliško, J., & **Erceg, N.** (2022). Predicting the Most Common Incorrect Response: Metacognitive Advantage of Deliberative over Intuitive Responders on Cognitive Reflection Test. *Studia Psychologica, 64*(3), 256-267. https://doi.org/10.31577/sp.2022.03.852

- **Erceg, N**., Galić, Z., & Bubić, A. (2022). Normative responding on cognitive bias tasks: Some evidence for a weak rationality factor that is mostly explained by numeracy and actively open-minded thinking. *Intelligence, 90,* 101619. https://doi.org/10.1016/j.intell.2021.101619
- **Erceg, N.,** Galić, Z., Bubić, A., & Jelić, D. (2022). Who detects and why: how do individual differences in cognitive characteristics underpin different types of responses to reasoning tasks?. *Thinking & Reasoning*, 1-49. https://doi.org/10.1080/13546783.2022.2108897
- Tomljenovic, H., Bubic, A., & **Erceg, N.** (2022). Contribution of rationality to vaccine attitudes: Testing two hypotheses. *Journal of Behavioral Decision Making, 35*(2), e2260. https://doi.org/10.1002/bdm.2260
- Lučić, A., Barbić, D., **Erceg, N.,** Palić, I. & Uzelac, M. (2022). Financial Literacy and Financial Capability - What Is What*? Proceedings of the 39th International Business Information Management Association (IBIMA),* 30-31 May 2022, Granada, Spain, ISBN: 978-0-9998551-8-8, ISSN: 2767-9640
- **Erceg, N.,** Galić, Z., & Ružojčić, M. (2020). A reflection on cognitive reflection-testing convergent/divergent validity of two measures of cognitive reflection. *Judgment & Decision Making, 15*(5), 741-755. https://doi.org/10.1017/S1930297500007907
- **Erceg, N.,** Ružojčić, M., & Galić, Z. (2020). Misbehaving in the corona crisis: The role of anxiety and unfounded beliefs. *Current Psychology, 41*, 5621–5630. https://doi.org/10.1007/s12144-020-01040-4
- Tomljenovic, H., Bubic, A., & **Erceg, N.** (2020). It just doesn't feel right–the relevance of emotions and intuition for parental vaccine conspiracy beliefs and vaccination uptake. *Psychology & health, 35*(5), 538-554. https://doi.org/10.1080/08870446.2019.1673894
- **Erceg, N**., Galić, Z., & Bubić, A. (2019). "Dysrationalia" among university students: The role of cognitive abilities, different aspects of rational thought and self-control in explaining epistemically suspect beliefs. *Europe's Journal of Psychology, 15*(1), 159-175. https://doi.org/10.5964/ejop.v15i1.1696
- **Erceg, N.**, Galić, Z., & Vehovec, M. (2019). What Determines Financial Literacy? In Search of Relevant Determinants. *Revija za socijalnu politiku, 26*(3), 293-312. https://doi.org/10.3935/rsp.v26i3.1541

- **Erceg, N.,** & Bubić, A. (2019). Doprinos moralnih temelja, religioznosti i društvene ideologije u objašnjenju orijentacija prema sreći. *Suvremena Psihologija. 22*(1), 67-85. https://doi.org/10.21465/2019-SP-221-05
- **Erceg, N.,** Burghart, M., Cottone, A., Lorimer, J., Manku, K., Pütz, H., ... & Willems, M. (2018). The effect of moral congruence of calls to action and salient social norms on online charitable donations: A protocol study. *Frontiers in Psychology, 9*, 1913. https://doi.org/10.3389/fpsyg.2018.01913
- Bubic, A., & **Erceg, N.** (2018). Do we know what makes us happy? The relevance of lay theories of happiness and values for current happiness. *Primenjena Psihologija, 11*(3), 345-364. https://doi.org/10.19090/pp.2018.3.345-364
- Bubić, A., & **Erceg, N.** (2018). The role of decision making styles in explaining happiness. *Journal of Happiness Studies, 19*, 213-229. https://doi.org/10.1007/s10902-016-9816-z
- **Erceg, N.,** & Bubić, A. (2017). One test, five scoring procedures: different ways of approaching the cognitive reflection test. *Journal of Cognitive Psychology, 29*(3), 381-392. https://doi.org/10.1080/20445911.2016.1278004
- Rogić Vidaković, M., Jerković, A., Jurić, T., Vujović, I., Šoda, J., **Erceg, N.,** ... & Đogaš, Z. (2016). Neurophysiologic markers of primary motor cortex for laryngeal muscles and premotor cortex in caudal opercular part of inferior frontal gyrus investigated in motor speech disorder: a navigated transcranial magnetic stimulation (TMS) study. *Cognitive processing, 17*, 429-442. https://doi.org/10.1007/s10339-016-0766-5
- Bubić, A., & **Erceg, N.** (2015). The relevance of cognitive styles for understanding individuals' cognitive functioning. *Suvremena psihologija, 18*(2), 159-173. https://hrcak.srce.hr/169814
- Vidaković, M. R., Schönwald, M. Z., Jurić, T., **Erceg, N.,** Bubić, A., Vulević, Z., ... & Đogaš, Z. (2015). Evaluation of corticobulbar and corticospinal excitability in developmental stuttering: the navigated transcranial magnetic stimulation study. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation, 8*(2), 317-318. https://doi.org/10.1016/j.brs.2015.01.031

- **Erceg, N.,** & Galić, Z. (2014). Overconfidence bias and conjunction fallacy in predicting outcomes of football matches. *Journal of economic psychology, 42*, 52-62. https://doi.org/10.1016/j.joep.2013.12.003

Knjige

- Lučić, A., Barbić, D., & **Erceg, N.** (2020). *Financial socialization of children: Using education to encourage lifetime saving*. Faculty of Economics and Business, University of Zagreb.

Poglavlja u knjigama

- **Erceg, N.** (2014). Dva primjera pogreški zbog dvostrukog procesiranja. U Polšek, D. i Bovan, K. (Ur.), *Uvod u Bihevioralnu Ekonomiju*. Zagreb: Institut društvenih znanosti, Ivo Pilar.