

Judgements of Solvability: Elucidating the First Stage of Meta-Reasoning

By

Olivia Rose Burton

*Thesis
Submitted to Flinders University
for the degree of*

Doctor of Philosophy

College of Education, Psychology, and Social Work

16th of November 2022

Table of Contents

Summary	iv
Declaration	i
Acknowledgement of Country	ii
Acknowledgements	iii
Author Note	v
List of Conference Proceedings	vi
List of Publications	vii
List of Figures	viii
List of Tables	ix
Chapter 1: General Introduction	1
Reasoning and Meta-reasoning	2
Judgements of Solvability	7
JOS Measurement Factors	11
Are JOSs Trainable?	12
JOS Study Design Factors	14
Self-Regulated versus Time-Limited Solving Trials	15
The Role of Cognitive Reflection in Judgements of Solvability.....	17
Thesis Overview.....	19
Chapter 2: How Accurate and Predictive are Judgements of Solvability? Explorations in a Two-Phase Anagram Solving Paradigm	23
Abstract	23
Introduction	25
Are JOSs Discriminating?	25
Do JOSs Predict Later Problem Solving?.....	27
Approaches to Measuring JOS Discrimination and Predictiveness	28
Is JOS Discrimination and Predictiveness Trainable?.....	29
Overview	30
Experiment 1	31
Method.....	32
Results	34
Discussion.....	45
Experiment 2	45
Method.....	46
Results	47
Discussion.....	56
Experiment 3	57

Method.....	58
Results	59
Discussion.....	69
General Discussion.....	71
JOS Discrimination.....	72
JOS Predictiveness	74
Paradigms for Measuring JOSs	78
Implications for Learners.....	78
Conclusion	78
Supplementary Materials.....	80
Chapter 3: Linking Judgements of Solvability, Solving Success, and Cognitive Reflection.....	87
Abstract	87
Introduction	88
Cognitive Reflection and Meta-Reasoning.....	89
Is Cognitive Reflection Related to JOS Discrimination and Predictiveness?	90
The Effect of Task Factors on the Relationship Between JOSs and Cognitive Reflection.....	91
Overview	92
Method	93
Participants	93
Stimuli	93
Procedure	94
Results	94
Effects of CRT on JOS Phase Discrimination.....	95
Solving Phase	98
S JOS discrimination versus S JOS predictiveness	104
Discussion	106
Cognitive Reflection Predicted Anagram Solving Ability, but Not Solvability Intuition.....	107
Inclusion of Longer-Duration Anagrams Increased Solving for Reflective Thinkers	110
Chapter 4: Unpacking the Relationship Between Initial Judgements of Solvability and Problem Solving: Interleaving Impacts Meta-Reasoning	114
Abstract	114
Introduction	115
JOS Measurement in a Blocked Design	116
The Influence of Design on JOS Discrimination and Predictiveness.....	117
Effects of Problem Duration in Blocked versus Interleaved Designs	120
Method	120
Participants	120
Stimuli	121
Design.....	121

Procedure	121
Results	122
JOS Trials	123
Solving Trials	126
S JOS discrimination versus S JOS predictiveness	134
Discussion	136
Interleaving Influences JOS Discrimination and Predictiveness.....	137
Longer-Duration Anagrams Increased Solving During JOS Trials but did not Moderate the Effects of Blocking vs. Interleaving	138
Conclusion	139
Supplementary Materials.....	141
Does JOS Predictiveness of Solving Response Times Vary Depending on Duration and Design?	141
Chapter 5: General Discussion	149
Summary of Experiment Findings	149
JOSs Can be Discriminating, but not Predictive of Problem-Solving Success.....	153
Do JOSs Predict Feelings of Rightness?	156
Pass Responses During Self-Regulated Solving	158
Individual Differences in Meta-Reasoning and Insight Reasoning.....	159
Cognitive Reflection Predicts Anagram Solving.....	160
Effects of Open-Minded Thinking on JOS Discrimination and Predictiveness.....	161
Optimising JOS Discrimination and Predictiveness	163
Other Limitations and Future Directions.....	164
Conclusions	166
References	168
Appendices.....	189
Appendix A – Anagram stimuli	189
Appendix B – Experiment 1 & 2: Training group instructions.....	190
Appendix C – Experiment 2: No-training group instructions	192
Appendix D – Experiment 3: Training group instructions.....	193
Appendix E – Experiment 3: No-training group instructions	195
Appendix F – Experiment 4: Blocked design instructions.....	197
Appendix G – Experiment 4: Interleaved design instructions	198
Appendix H – 7-item Cognitive Reflection Test (CRT; Toplak et al., 2014).....	199
Appendix I – Example Judgement of Solvability trial.....	200
Appendix J – Example solving trials.....	201
Appendix K – Debriefing form (all experiments).....	202
Appendix L – Mechanical Turk Virtual Task Description (all experiments)	203

Summary

Meta-reasoning involves monitoring and control of one's reasoning processes and begins with a Judgement of Solvability (JOS). Reasoners use the JOS to decide whether a problem is solvable prior to a solving attempt, and whether to regulate solving effort. Thus, misjudging problem solvability can have several consequences such as wasting time on futile attempts to solve unsolvable problems, or erroneously abandoning solvable problems.

Most of the current research using JOSs has focussed on the problem-solving stimulus features that bias or inform JOSs. My thesis provides a new and original contribution to JOS research by examining how experiment, measurement, and individual difference factors influence this initial stage of meta-reasoning. To this end, my thesis clarifies how well JOSs about anagrams discriminate between solvable and unsolvable problems (termed JOS discrimination) and predict later problem-solving outcomes (termed JOS predictiveness) under different experimental task conditions. It also clarifies whether individual differences in cognitive reflection enabled more discriminating and predictive JOSs, given that some research has shown that a more reflective thinking style is related to better meta-reasoning.

Anagrams are sometimes solved spontaneously, thus each of my thesis experiments separated problems solved during the JOS (i.e., 'already solved' JOSs) from intuitions about problem solvability (i.e., 'solvable' or 'not solvable' JOSs), to avoid confounding intuitions with solutions found during the JOS. In each of four experiments, participants were briefly shown an anagram and then made a JOS about the anagram, and then later attempted to solve the anagram. Experiments 1-3 examined the influence of anagram presentation duration prior to making one's JOS to determine whether allowing participants more time to develop their intuitions about solvability led to more discriminating and more predictive JOSs. To do this, Experiments 1-3 presented anagrams in 4 blocks. In the training groups, anagrams were presented for 16 s at first, which halved across blocks. In the no-training groups anagram duration was always 2 s. After participants completed the blocks of JOSs, they then attempted

to solve each anagram in a single solving block. Thus, Experiments 1-3 had an additional focus of whether participants could be trained (via practice with initial blocks of longer-duration anagrams) to develop more accurate and predictive JOSs. In Experiment 4, I examined whether JOS discrimination and predictiveness are influenced by whether solving attempts follow each JOS (interleaved design) or occur after all JOSs are made (blocked design).

Each experiment revealed that ‘solvable’ JOSs were less discriminating than ‘already solved’ JOSs, but were often (though not always) discriminating. Furthermore, when anagram duration was manipulated within-subjects, ‘solvable’ JOSs were more discriminating at longer versus shorter durations. However, training with initial blocks of longer-duration anagrams did not generate more discriminating JOSs. A more reflective cognitive style led to a higher likelihood of an anagram being reported as ‘already solved’ during the JOS, but interestingly, did not produce more discriminating or predictive JOSs.

Although ‘solvable’ JOSs were discriminating, they generally did not predict problem-solving success, except when the study design was interleaved. When the solving trials were self-regulated (i.e., solving trials presented solvable and unsolvable anagrams and participants could choose whether to attempt problem-solving or disengage), ‘solvable’ JOSs led to greater solving effort expenditure on unsolvable items. Thus, simply judging a problem as ‘solvable’ was generally not predictive of problem-solving success, and it also misled effort regulation on unsolvable problems.

In sum, my findings demonstrate that study design, anagram presentation duration, and self-regulated solving influence intuitive JOSs. Given that these early judgements inform later effort regulation, it is important to understand what drives accurate and predictive solvability judgements. My thesis contributes to this goal.

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signed: Olivia R. Burton

Date: 1/09/2022

Acknowledgement of Country

I acknowledge that this work was produced on the unceded lands of the Kurna and Gunaikurnai people. I recognise the Kurna and Gunaikurnai people are the Traditional Custodians of the land where my research was conducted. Always was, always will be Aboriginal land.

Acknowledgements

First, I would like to thank my principal supervisor, Associate Professor Glen Bodner. Thank you for agreeing to supervise me when I approached you as an orphaned PhD student! Your creativity, expertise, and insights from a different research background have provided a new and interesting perspective on this project, and I am extremely grateful for the time that you have dedicated towards this research. Most importantly though, thank you teaching me to really *think* like a researcher by encouraging my curiosity and critical thinking. I have learned so much from being under your wing, and I am excited to carry all my knowledge forward.

To my associate supervisor, Dr Paul Williamson – thank you for your support and valuable guidance, and for always finding the time (no matter how busy your schedule) to provide careful feedback on my work. You have been an excellent teacher in data analysis and interpretation, and I appreciate the time you took to teach me these important skills.

To my original principal supervisor, Dr Michelle Arnold, thank you for inspiring my interest in meta-reasoning, for fostering my critical thinking skills, and for pushing me to self-regulate my learning – a skill which has proven to be invaluable during my PhD journey. The lab culture you established, and your knowledge and expertise, were certainly missed after you left Flinders.

To my peers in the Memory and Metacognition lab, Matt, Henry, and Georgie – you have been integral parts of this journey. Matt and Henry – thank you for your supportive feedback, and for helping to demystify my pesky CRT results. I am sure that I would still be procrastinating Chapter 3 if it were not for our group discussions about cognitive reflection. Georgie – thank you for the momentary distractions during coffee and/or lunch breaks, and for providing a sounding board when I needed one. I have no doubt that all three of you will all be incredible researchers and/or academics and I am so excited to hear about all your achievements.

I also want to express my gratitude to the broader Flinders University postgraduate psychology community, both past and present. I have made some lifelong friends here – it feels bittersweet that my journey at Flinders is coming to an end. You are an incredible and supportive group of people. I will miss common room lunches, hallway conversations, and trips to Bon Voyage and the Tavern. I wish you all every success on your journeys.

My friends outside of Flinders have been an incredible support, too, but particularly my best friend Madi. Thank you for always and unreservedly believing in me. Your interest in my research (even when we no longer saw each other every day) has meant so much to me.

To Ryan – thank you for truly understanding what it is like to carry out such a huge project to completion, for cooking me meals, and for making me laugh when I needed it. Your efforts to make the last few months of my PhD as smooth as possible were so appreciated, and I will always be grateful for your patience, generosity, kindness, and understanding during this time.

To my uncle Mike and my dear γιαγιά, thank you both for ensuring I was fed, watered, and generally surviving every week. Thank you for providing a comfortable space for me to retreat to when I needed it. Γιαγιά, σε ευχαριστώ για όλα. Σας ευχαριστώ που με ταΐσατε και με φροντίσατε. Σας αγαπώ και σας αφιερώνω αυτή τη διατριβή.

To my mum and dad – thank you for literally everything. I have been so privileged to have you both provide me with every opportunity you possibly could. I am grateful for your unwavering support despite how outlandish my ambitions are (who would have thought I would complete a PhD?), and for pushing me to keep going when things were difficult. Although writing this thesis was a huge personal goal for myself, my other goal was to make you both proud of me. I hope I succeeded. I dedicate this thesis to you both.

Finally, I acknowledge I was supported by an Australian Government Research Training Program Scholarship.

Author Note

In the empirical chapters of this thesis, I have used the pronoun “we” to reflect the collaboration between myself and the co-authors. Although I took the primary role in conceptualising, programming, data analysis, and writing each experiment, I had valuable help and insights from co-authors. In the General Introduction and General Discussion, I have used the pronouns “I” and “my” to reflect that these sections were done independently.

I have also chosen to use four acronyms in this thesis:

Judgement of Solvability	JOS
‘Already solved’	AS
‘Solvable’	S
‘Not solvable’	NS

These acronyms are used repeatedly throughout this thesis, and they aid in ease of locating a particular Judgement of Solvability. That is, it is easier for the reader to spot “AS JOS” than “Already Solved Judgement of Solvability” in the text.

List of Conference Proceedings

- Burton, O.R., Bodner, G.E., Williamson, P. (2022, Jul). *Unpacking the Relationship Between Initial Judgements of Solvability and Problem Solving: Interleaving Impacts Metacognition*. Paper presented at the Australasian Brain and Psychological Sciences Meeting, Brisbane, Australia.
- Burton, O.R., Bodner, G.E., & Williamson, P. (2021, Nov). *Training improves discrimination of judgements of solvability, but not how well they predict later problem-solving success*. Poster presented at the 62nd Annual Meeting of the Psychonomic Society, Boston, United States of America
- Burton, O.R., Bodner, G.E., & Williamson, P. (2021, Jun). *Training can improve the accuracy of judgements of solvability*. Paper presented at the Virtual Society of Applied Research in Memory and Cognition Conference.
- Burton, O.R., Bodner, G.E., & Williamson, P. (2021, May). *The effects of cognitive reflection on judgements of problem solvability and problem-solving success*. Paper presented at the Virtual European Association for Research on Learning and Instruction: Special Interest Group 16 Meeting, Maastricht, the Netherlands
- Burton, O.R., Bodner, G.E., & Williamson, P. (2021, Apr). *Initial Judgements of Solvability: Are they accurate, trainable, and do they predict problem-solving success?* Paper presented at the Virtual Australasian Experimental Psychology Society Annual Meeting, Brisbane, Australia
- Burton, O.R., Bodner, G.E., & Williamson, P. (2020, Dec). *Initial judgements of a problem's solvability: Are they accurate & trainable?* Poster presented at Virtual KiwiCAM2020, Auckland, New Zealand

List of Publications

- Burton, O.R., Bodner, G.E., Williamson, P., & Arnold, M.M. (2022). How accurate and predictive are judgments of solvability? Explorations in a two-phase anagram solving paradigm. *Metacognition and Learning*. <https://doi.org/10.1007/s11409-022-09313-y>
- Burton, O.R., Bodner, G.E., & Williamson, P. (2022). *Linking Judgements of Solvability, Solving Success, and Cognitive Reflection* [Manuscript under review at *Thinking and Reasoning*]
- Burton, O.R., Bodner, G.E., & Williamson, P. (2022). *Unpacking the Relationship Between Initial Judgements of Solvability and Problem Solving: Interleaving Impacts Meta-Reasoning* [Manuscript under review at *Thinking and Reasoning*]

List of Figures

Figure 1.1 <i>The Time Course of Reasoning and Meta-Reasoning Processes</i>	4
Figure 2.1 <i>Experiment 1: Mean Proportions of Hits and False Alarms for AS+S JOSs in the JOS Phase in Experiment 1</i>	35
Figure 2.2 <i>Experiments 1-3: Mean Proportions of Hits and False Alarms for AS JOSs in the JOS Phase</i>	36
Figure 2.3 <i>Experiments 1-3: Mean Proportions of Hits and False Alarms for S JOSs in the JOS Phase</i>	37
Figure 2.4 <i>Experiments 1-3: Mean Proportions of Solvable Anagrams Solved in the Solving Phase</i>	42
Figure 2.5 <i>Experiments 1-3: Mean Proportion of Solved Versus Not Solved Outcomes for Solvable Anagrams</i>	44
Figure 2.6 <i>Experiment 3: Mean Proportions of Not-Solvable Outcomes to Solvable Anagrams</i>	64
Figure 2.7 <i>Experiment 3: Mean Solving Times for Anagrams</i>	67
Figure 2.8 <i>Experiment 3: Mean Response Time for Not-Solvable Responses</i>	68
Supplementary Figure 2.1 <i>Experiments 2 and 3: Mean Proportions of Hits and False Alarms for AS+S JOSs in the JOS Phase</i>	85
Supplementary Figure 2.2 <i>Mean Proportion of Anagrams Solved in Each Block as a Function of JOS</i>	86
Figure 3.1 <i>AS JOSs: Mean Discrimination as a Function of Experiment, Group, and CRT Score</i>	97
Figure 3.2 <i>Solving Phase: Mean Proportion Solved Collapsed Across JOS as a Function of Experiment, Group, and CRT Score</i>	101
Figure 3.3 <i>Solving Phase: Solving Outcomes Among AS and S JOSs as a Function of CRT Score</i>	104
Figure 4.1 <i>JOS Trials: Mean Proportions of Hits and False Alarms</i>	124
Figure 4.2 <i>Solving Trials: Mean Proportion Solved</i>	127
Figure 4.3 <i>Solving Trials: Proportion of Solved Versus Not Solved Outcomes for Solvable Anagrams</i>	129
Figure 4.4 <i>Solving Trials: Mean Proportion of Not-Solvable Responses</i>	133
Supplementary Figure 4.1 <i>Solving Trials: Mean Response Times for Solved Anagrams</i>	142
Supplementary Figure 4.2 <i>Solving Trials: Mean Response Times for Not-Solvable Responses</i>	145

List of Tables

Table 2.1 <i>Experiment 1: JOS Phase Results</i>	38
Table 2.2 <i>Experiment 1: JOS Phase Linear Contrast ANOVAs</i>	40
Table 2.3 <i>Experiment 2: JOS Phase ANOVAs Results</i>	48
Table 2.4 <i>Experiment 2 and 3: JOS Phase Linear Contrast ANOVAs</i>	49
Table 2.5 <i>Experiment 2: JOS Phase Discrimination ANOVAs in Blocks 1-3</i>	51
Table 2.6 <i>Experiment 2 and 3: JOS Phase Discrimination ANOVAs in Block 4</i>	53
Table 2.7 <i>Experiments 2 and 3: Solved versus Not Solved Outcomes ANOVAs</i>	55
Table 2.8 <i>Experiment 3: JOS Phase ANOVAs Results</i>	60
Table 2.9 <i>Experiment 3: JOS Phase Discrimination ANOVAs in Blocks 1-3</i>	62
Table 2.10 <i>Experiment 3: Proportion of Not-Solvable Responses ANOVA Results</i>	65
Table 2.11 <i>Experiment 3: Interaction Contrasts for Proportion of Not-Solvable Responses ANOVA Results</i>	66
Table 2.12 <i>Experiment 3: Mean Response Time for Not-Solvable Responses ANOVA Results</i>	69
Supplementary Table 2.1 <i>Experiment 2: JOS Phase ANOVA Results for AS+S JOSs</i>	80
Supplementary Table 2.2 <i>Experiment 2: JOS Phase Discrimination ANOVAs in Blocks 1-3 for AS+S JOSs</i>	81
Supplementary Table 2.3 <i>Experiment 2 and 3: JOS Phase Discrimination ANOVAs in Block 4 for AS+S JOSs</i>	82
Supplementary Table 2.4 <i>Experiment 3: JOS Phase ANOVA Results for AS+S JOSs</i>	82
Supplementary Table 2.5 <i>Experiment 3: JOS Phase Discrimination ANOVAs in Blocks 1-3 for AS+S JOSs</i>	83
Supplementary Table 2.6 <i>Experiments 1-3: Mean Proportions and Standard Deviations for Solving Phase Solutions</i>	83
Supplementary Table 2.7 <i>Experiment 3 Solving Phase: Mean Proportions and Standard Deviations for Final Not-Solvable Responses to Unsolvable Anagrams</i>	83
Supplementary Table 2.8 <i>Example of how Each Solving Phase Measure was Calculated for a Participant</i>	84
Table 3.1 <i>JOS Discrimination: CRT ANCOVA Results by JOS</i>	96
Table 3.2 <i>AS JOS Discrimination: CRT × Group Interaction Contrasts</i>	97
Table 3.3 <i>Solving Phase: Proportion Solved ANCOVA Result</i>	100
Table 3.4 <i>Mean Proportion Solved: CRT × Group Interaction Contrasts</i>	101
Table 3.5 <i>Solving Phase: Solving Outcome ANOVA Results by JOS</i>	103
Table 3.6 <i>Proportion Solved for S JOSs: ANCOVA with Mean-Centred S JOS Discrimination and Mean-Centred CRT</i>	106
Table 4.1 <i>JOS Trial Discrimination: ANOVA Results by JOS</i>	125
Table 4.2 <i>Solving Trials: Proportion Solved ANOVA Results</i>	128
Table 4.3 <i>Solving Trials: Solved vs. Not Solved Outcomes ANOVA Results by JOS</i>	131

Table 4.4 <i>Solving Trials: Proportion of Not-Solvable Responses ANOVA Results</i>	134
Table 4.5 <i>Proportion Solved for S JOSs: ANCOVA with Mean-Centred S JOS Discrimination</i>	136
Supplementary Table 4.1 <i>Mean Solving Time on Solving Trials: ANOVA Results</i>	142
Supplementary Table 4.2 <i>Mean ‘Not Solvable’ Response Times on Solving Trials: ANOVA Results</i>	144
Supplementary Table 4.3 <i>Mean ‘Not Solvable’ Response Times by Duration on Solving Trials: ANOVA Results</i>	147
Supplementary Table 4.4 <i>Mean Proportions and Standard Deviations for Solving Rates</i> ..	148
Supplementary Table 4.5 <i>Mean Proportions and Standard Deviations for Final Not- Solvable Responses to Unsolvable Anagrams</i>	148

Chapter 1: General Introduction

Some of the most common pieces of advice imparted upon learners taking exams are to “use the time wisely” (Deakin University, 2020; UNSW Sydney, 2022) and to “start with what you know” (Lyness, 2016). Such advice aims to encourage learners to skew their efforts toward the exam items they think they can answer to maximise their performance. The advice is sound – learners have nothing to gain from expending effort on exam items they cannot answer, especially when time is critical (Ackerman & Thompson, 2017; Toplak et al., 2014). However, benefitting from this approach is contingent on making a *Judgement of Solvability* (JOS; Ackerman & Thompson, 2017) that accurately reflects whether the learner can solve the problem. Making a ‘solvable’ JOS for an unsolvable problem may cost the individual valuable time and making an ‘unsolvable’ JOS for a solvable problem may cost them marks if they skip an item that they could have solved. Therefore, the ability to make accurate and predictive JOSs is critical to the strategic regulation of time and effort for problem solving.

Research on metacognition and its various stages (e.g., judgements of learning, retrospective decision confidence ratings) has mainly focussed on how well people understand their cognitive processes about memory (i.e., meta-memory) and general knowledge (Karpicke, 2009; Koriat, 1997; Livingston, 1997). Metacognition for more complex cognitive tasks such as reasoning and problem-solving, or *meta-reasoning*, has only recently become a research focus. Analogous to ease-of-learning meta-memory judgements (Nelson & Narens, 1990), the JOS is the first stage of metacognitive monitoring during reasoning. JOSs are theorised to be made intuitively before formal reasoning processes commence (Ackerman & Beller, 2017; Ackerman & Thompson, 2017; Bolte & Goschke, 2005; Markovits et al., 2015; Novick & Sherman, 2003). Hence, JOSs may relate to problem-solving effort-investment (e.g., Lauterman & Ackerman, 2019; Payne & Duggan, 2011), and may predict successful reasoning outcomes and failures (e.g., Markovits et al., 2015; Siedlecka et al., 2016).

To date, meta-reasoning research has mainly focussed on measuring reasoners' confidence in their final reasoning outcomes; the initial stages of meta-reasoning have been relatively less explored. To close this knowledge gap, my thesis provides an in-depth exploration of JOSs, establishing whether they are sensitive to problem solvability and whether they predict later problem-solving success and effort regulation. To this end, I investigated how JOSs are impacted by different experimental task conditions (e.g., longer vs. shorter problem presentation durations) and study designs (e.g., blocking vs. interleaving JOSs and solving attempts). I also examined whether JOSs are trainable in ways that increase their accuracy and predictiveness, and whether individual differences in meta-reasoning (specifically with respect to cognitive reflection) influence how well people make JOSs.

Reasoning and Meta-reasoning

People constantly engage in reasoning to complete tasks, make decisions, and perceive the world around them (Rips, 1990). Reasoning is engaged when methodical problem-solving is required to solve a problem (Metcalf & Wiebe, 1987; Salvi et al., 2016). Reasoning can be deductive or inductive (Johnson-Laird, 1999; Markovits & Nantel, 1989; Thompson et al., 2013; Toplak et al., 2014), can vary from abstract to concrete (Markovits et al., 2002; Markovits et al., 2015), and can be analytical (e.g., Ravens Matrices, Latin Squares; Keedwell & Dénes, 2015; Raven, 2003) or intuitive (Kahneman, 2011).

However, solutions to problems can also be derived via insight, without methodical problem solving. *Insight reasoning* refers to the phenomenon of uncovering solutions to problems with little awareness of how the answer was produced. Examples of insight reasoning problems include anagrams (e.g., Bowden & Jung-Beeman, 2003; Novick & Côté, 1992; Novick & Sherman, 2003; Novick & Sherman, 2008; Salvi et al., 2016), rebus puzzles (e.g., Chu & MacGregor, 2011) and remote associates (e.g., Chuderski & Jastrzębski, 2018; Gilhooly & Fioratou, 2009; Zedelius & Schooler, 2015). Successful reasoning via insight is

commonly known as an “Aha!” or “Eureka!” experience (Chuderski & Jastrzębski, 2018; Topolinski & Reber, 2010). Solutions to insight reasoning problems are often generated differently from analytic reasoning problems such as algebraic problems. For instance, Metcalfe and Wiebe (1987) had participants solve algebraic problems or insight problems, and collected ratings of how close reasoners believed they were to solution retrieval at various intervals. For algebraic problems, these ratings increased incrementally until a solution was reached, but for insight problems, these ratings were stable then jumped sharply when a solution was found.

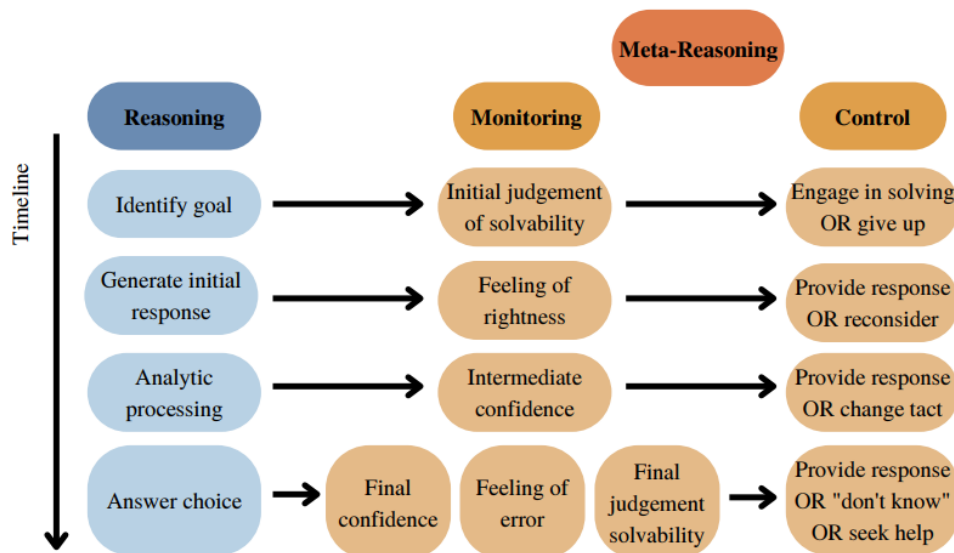
Although research on reasoning has spanned decades, research has only recently begun to examine meta-reasoning, that is, the awareness of one’s cognitive processes during reasoning (Ackerman & Thompson, 2017). The term meta-reasoning was developed to distinguish it from metacognition which more broadly involves assessing the cognitive processes concerned with learning, remembering, and comprehension (Ackerman & Thompson, 2017; Berardi-Coletta, 1995). Dual-process accounts of reasoning suggest that two distinct systems can be engaged when reasoning: a faster System 1, and a slower System 2. System 1 is an autonomous process that relies on mental shortcuts and generates answers based on intuition, whereas System 2 involves a more methodical and analytical approach. Whether reasoners engage in System 1 or System 2 depends on the type of reasoning problem (Alter et al., 2007), time constraints (Evans & Curtis-Holmes, 2005; Finucane et al., 2000), and whether the reasoner has the motivation and capacity to engage in analytical thinking (Stanovich, 1999; Stanovich & West, 1998).

As with metacognition, meta-reasoning also involves both monitoring and control processes. Monitoring allows the reasoner to assess how well their reasoning process is tracking, and control determines decisions to change reasoning strategies, report an answer, or to seek help (Ackerman & Thompson, 2017; Evans & Fisher, 2011; Koriat & Goldsmith,

1996; Koriat et al., 2006). Ackerman and Thompson (2017) proposed a meta-reasoning framework to illustrate the monitoring and control processes in a dual-process theory (see Figure 1.1).

Figure 1.1

The Time Course of Reasoning and Meta-Reasoning Processes



Note. Adapted from “Meta-Reasoning: Monitoring and Control of Thinking and Reasoning” by R. Ackerman & V.A. Thompson, 2017, *Trends in Cognitive Sciences*, 21(8), 607-617.

According to Ackerman and Thompson’s (2017) meta-reasoning framework, the reasoner first makes an initial judgement about whether the problem is solvable. This initial judgement informs the decision about whether to pursue problem-solving efforts. If the reasoner chooses to engage in solving, their reasoning processes may first generate a rapid, intuitive response, which they monitor by evaluating their feeling of rightness. If this feeling is strong, the reasoner may choose to stop reasoning and put forth their found solution, whereas if it is weak, they may try a more analytic reasoning strategy. Finally, reasoners will then assess their final confidence or feeling of error for their answer, and then control their response by providing the chosen answer, seeking help, or giving up. After completing a

solving attempt, reasoners may also form a final judgement of solvability regarding whether the problem was solvable.

To date, meta-reasoning research has mainly focussed on calibration—that is, whether reasoners' confidence in their final answers calibrates with their objective performance (Borracci & Arribalzaga, 2018; Burson et al., 2006; Lingel et al., 2019). Additionally, much of this research has been performed on heuristics-and-biases tasks (Pennycook et al., 2017; Primi et al., 2018; Thompson & Morsanyi, 2012; Thompson, 2009). Take, for example, the following item from De Neys and Glumicic (2008):

“In a study 1000 people were tested. Among the participants, there were 997 nurses and 3 doctors. Paul is a randomly chosen participant of this study. Paul is 34 years old. He lives in a beautiful home in a posh suburb. He is well-spoken and very interested in politics. He invests a lot of time in his career.

What is most likely?

a) Paul is a nurse

b) Paul is a doctor” (p. 1282).

Because the description of Paul is a more salient representation of a doctor, reasoners often ignore the base rate and decide that Paul is a doctor, even though it is far more likely Paul is one of the 997 nurses in the sample. Decisions on these tasks tend to be made easily and intuitively (Thompson, 2009; Thompson et al., 2013). Easy retrieval typically leads reasoners to report higher confidence in their reasoning decisions, even when those decisions are incorrect (Kelley & Lindsay, 1993; Oppenheimer, 2008; Thompson et al., 2013), resulting in poor calibration of meta-reasoning and accuracy. Indeed, research using heuristics-and-biases tasks have found reasoners to be poorly calibrated due to overconfidence in incorrect responses (Coutinho et al., 2021; Pennycook et al., 2017) because reasoners used misleading metacognitive cues to judge their decision accuracy (such as answer fluency; De

keersmaecker et al., 2019). Heuristics-and-biases tasks are useful for measuring feelings of rightness, intermediate confidence, and final confidence because some reasoning must occur before reasoners can monitor errors (or *conflict*) in their reasoning outcomes, and in turn, be able to control their responses (De Neys & Pennycook, 2019). For instance, a low confidence rating on an incorrect, intuitive response would indicate that a reasoner has identified conflict in their reasoning process, suggesting their meta-reasoning is well-calibrated. Thus, dual-process approaches to measuring meta-reasoning are informative about how well people can identify and overcome errors and biases in their reasoning.

Relative to the final confidence stage in Ackerman and Thompson's (2017) meta-reasoning framework, the initial JOS stage has received less study. JOSs are intuitive judgements about solvability made before formal reasoning commences on a task. Importantly, measuring JOSs using typical dual-process reasoning tasks (e.g., syllogisms, base rate problems, conjunction fallacy problems) can present some issues because the intuitive answer is often generated very rapidly after reasoners process the problem (Bago & De Neys, 2017; Strudwicke et al., 2022). For example, simply reading the given problem might elicit an incorrect heuristic response before the reasoner can even generate a JOS. Thus, although the initial stage of meta-reasoning warrants investigation, typical dual-process reasoning tasks may obstruct the measurement of JOSs. Therefore, one of the aims of my thesis was to establish suitable means of measuring JOSs to help elucidate the first stage of meta-reasoning.

In dual-process reasoning, intuitive answers are generated rapidly and without much conscious deliberation (Bago & De Neys, 2019). Solutions to insight reasoning problems are generated in a similar way. Therefore, in this thesis, I distinguish *intuition* about solvability as a "gut feeling" that a solution to a problem exists that is made before reasoning commences

(Stanovich & West, 2000; Topolinski & Strack, 2009a) from *insight* that involves retrieval of the solution (Metcalf & Wiebe, 1987; Zhang et al., 2016).

Judgements of Solvability

Insight reasoning problems are useful for research on JOSs because they can be processed quickly, allowing researchers to capture a large number of JOSs while minimising participant fatigue (Healy et al., 2004). Solutions to insight problems can arise spontaneously (Topolinski et al., 2016), or after reasoners “restructure” the problem following a solving impasse (Ash & Wiley, 2006; Gilhooly & Murphy, 2005; Ohlsson, 2011). Restructuring may occur consciously (Gilhooly & Fioratou, 2009). For example, when anagram solving, one might decide to rearrange the letters in an anagram after failed attempts to decipher the solution which may lead to the sudden instantiation of the solution. However, restructuring may also occur unconsciously, for example, via unconsciously activating different representations of possible solutions (Ohlsson et al., 1992; Weisberg, 2015). Although insight problems are not necessarily subject to effortful and conscious processing, meta-reasoning processes can still arise during solution retrieval. For example, a reasoner might make a JOS about a problem, which then determines whether they choose to solve it. They might generate a solution straight away or may sit “lost in thought” (West et al., 2012, p. 506) trying to decipher a solution. A reasoner may monitor their reasoning strategy, decide it is unsuccessful, and in turn control their reasoning strategy by consciously restructuring the given problem, which in turn results in a solution or a decision to give up. Thus, insight reasoning problems can still capture meta-reasoning monitoring and control processes analogous to analytical reasoning, making them well-suited for studying the initial meta-reasoning stages.

To date, only a few studies have investigated whether JOSs can accurately distinguish between solvable and unsolvable insight problems (which I term *JOS discrimination*) and whether JOSs predict problems-solving successes or failures, and problem-solving effort regulation (which I term *JOS predictiveness*). Using the remote associates task, some studies found that participants were sensitive to whether word triads (e.g., playing, credit, report) were each related to a fourth word (e.g., card) or not (Balas et al., 2011; Bolte & Goschke, 2005; Bolte et al., 2003; Topolinski & Strack, 2009a; Undorf & Zander, 2017). Other studies have found that reasoners can accurately discriminate between solvable anagrams (e.g., RFADU – FRAUD) and unsolvable anagrams (e.g., ZEREB) (Novick & Sherman, 2003; Topolinski et al., 2016). Furthermore, Siedlecka et al. (2016) found that prospective confidence judgements about anagrams (i.e., “How confident are you that you will choose the correct solution?”) reliably predicted anagram solution decision accuracy. Together, these studies suggest that reasoners can be sensitive to problem solvability.

Other research has found that reasoners are not sensitive to problem solvability. Metcalfe (1986) measured JOSs by having participants rank different reasoning problems from “most likely to solve” to “least likely to solve”. Participants’ rankings did not correspond to their objective likelihood of solving each problem. Using remote associates, Ackerman and Beller (2017) compared “general” initial JOSs (i.e., whether reasoners believed the problem was objectively solvable) to “personal” initial JOSs (i.e., whether reasoners believed they could solve the problem). Neither general nor personal JOSs were discriminating – for both judgements, participants’ rate of “solvable” JOSs exceeded the actual rate of solvable problems, and these judgements did not reliably predict problem-solving performance. These studies suggest that reasoners’ JOSs can be unreliable and unrelated to problem solvability.

A third set of studies suggests that JOSs can discriminate and predict problem-solving success and effort regulation, but only under specific reasoning task conditions. For example, Valerjev and Dujmović (2020) found that JOSs for anagrams were discriminating for shorter anagrams (i.e., two syllables) but not for longer anagrams (i.e., three syllables). Reasoning task conditions have also been found to influence JOS discrimination and predictiveness with analytic problems. For example, Markovits et al. (2015) found that JOSs about syllogisms (i.e., rating confidence that their response will be logical) predicted successful problem-solving, but only when abstract premises were used (e.g., “If someone glebs, then they are brandup.”) rather than concrete premises (e.g., “If a candle is lit, the room will be illuminated.”). Lauterman and Ackerman (2019) examined whether initial JOSs discriminated between solvable and unsolvable Raven’s matrices and whether JOS predicted problem-solving success and effort regulation. JOSs were not discriminating when the unsolvable matrices violated more of the matrix rules (by switching more of the locations of the matrix elements) but were discriminating when there were fewer violations of the matrix rules. However, participants invested more time in solving matrices they judged as solvable regardless of whether the matrix was solvable. Participants’ time investment on unsolvable matrices suggested that JOSs misled effort regulation, as reasoners took longer to abandon problems that they judged to be solvable. Taken together, the literature on JOSs for insight reasoning problems provides mixed evidence regarding whether intuitive JOSs are sensitive to problem solvability and predict problem-solving outcomes.

With the exception of Metcalfe (1986), the JOS research described above has almost exclusively focussed on the stimulus features that bias or inform JOSs, such as problem length and problem difficulty (Ackerman & Beller, 2017; Balas et al., 2011; Markovits et al., 2015; Topolinski et al., 2016; Topolinski & Strack, 2009a; Valerjev & Dujmović, 2020). Because a reasoner’s ability to discriminate solvable from unsolvable problems is easily

biased by problem-solving stimulus features, this focus may explain, at least in part, the inconsistent findings. For example, Payne and Duggan (2011) and Ackerman and Beller (2017) both manipulated the accessibility of potential answers for insight reasoning problems; reasoners judged problems that had numerous potential answers as more solvable than problems with fewer potential answers, even though this was an invalid cue of solvability. Topolinski and Strack (2009b) found that JOSs for remote associates were also biased by semantic affect; words with positive valence were judged as having a solution word more often than words with negative valence. Topolinski et al. (2019) found that, among unsolvable anagrams, anagrams that were easy to pronounce (e.g. LAPNUK) were consistently judged as solvable more often than those that were difficult to pronounce (e.g. UNKLPA), even though easy-to-pronounce anagrams are actually *more* difficult to solve. Perhaps some stimulus features are more misleading than others, leading some studies to find that JOSs were sensitive to problem solvability despite biasing stimulus features (e.g., processing fluency of anagrams; Topolinski et al., 2016), and leading other studies to find that JOSs were not discriminating in the presence of biasing stimulus features (e.g., answer accessibility for remote associates; Ackerman & Beller, 2017). The focus on stimulus features that bias JOSs has been useful in informing our understanding of *how* people may judge solvability prior to a solving attempt (Lauterman & Ackerman, 2019), but our understanding of meta-reasoning may be strengthened by investigating JOSs in reasoning tasks that do not systematically elicit biased JOSs (i.e., a neutral task, e.g., de Chantal et al., 2020; Goel & Dolan, 2003).

Based on this literature review, a key research question my thesis aimed to answer was: Are JOSs generally predictive and discriminating when reasoners are not deliberately misled about solvability? To this end, I used anagrams as the problem-solving stimuli. Anagrams were all 5-letters long to control for misleading effects of problem length, required 3 letter moves to solve, and were all selected from the 5000 most used words in a corpus of

contemporary English (*Word frequency: based on 450 million word COCA corpus*, 2016).

Although I did not control for anagram processing fluency (some anagrams were pronounceable whereas others were not), Topolinski et al. (2016) found that participants were sensitive to problem solvability regardless of pronounceability. As detailed below, I also examined several experimental design factors that may have influenced how well JOSs discriminated and predicted problem-solving success in prior studies.

JOS Measurement Factors

When faced with a problem-solving task, reasoners spend some time processing the problem to get an intuitive sense of its solvability. Reasoners then make a JOS based on a threshold they set for accumulating information about problem solvability (Payne & Duggan, 2011). Thus, when reasoners are given more time to accumulate information about solvability and to develop their intuition, their JOSs should be more discriminating and predictive. On the other hand, given that insight reasoning problems can be solved spontaneously (Topolinski et al., 2016), some researchers have expressed concern that setting longer problem durations might lead to spontaneous solution retrieval (Balas et al., 2011; Bolte & Goschke, 2005; Lauterman & Ackerman, 2019; Siedlecka et al., 2016; Valerjev & Dujmović, 2020). To mitigate this issue, participants usually make JOSs for verbal insight reasoning problems presented under very short time constraints (e.g., 500 ms; Novick & Sherman, 2003). However, even though intuitive judgements can be made quickly and accurately (Kahneman, 2003; Lieberman, 2000), setting very brief problem presentation durations might underestimate JOS discrimination and predictiveness if participants begin relying on irrelevant cues to make their JOSs (Ackerman, 2019; Benjamin, 2005; Kahneman et al., 1982), or perhaps even resort to random responding.

Surprisingly, only one study has measured whether participants' JOSs were still discriminating after excluding problems solved during or before the JOS. Topolinski et al.

(2016, Experiment 7), found that JOS discrimination was marginally significant after discarding trials where participants had found the solution. In my studies, participants reported when they found a solution during the JOS task. Moreover, some participants (those in the “training” groups) were given longer-duration anagrams at first. Longer-duration anagrams may enable participants to develop more accurate intuitions about solvability, , to determine the conditions that maximise JOS discrimination and predictiveness. The experiments reported in Chapter 2 manipulated anagram presentation duration within subjects, starting with longer durations that halved across subsequent blocks. The experiment reported in Chapter 4 manipulated anagram duration between-subjects.

Are JOSs Trainable?

Generally, the goal of metacognitive training strategies is to help participants learn to self-regulate on future tasks to improve their ability to monitor and control decision making (Boekaerts, 1999; Schuster et al., 2020). Some prior work has found that explicit feedback about metacognitive decisions helps people to calibrate their confidence on a decision-making task (Novick & Sherman, 2003). Other studies have investigated whether individuals can strategically regulate their metacognitive confidence, by requiring them to either “report” their answer (and in turn, receive points if they are correct, or a point deduction if they are incorrect) or “withhold” their answers (Arnold et al., 2016; Higham, 2007). These studies have shown that participants can adjust their metacognitive decisions (i.e., the decision to report or withhold their answers) based on the costs/benefits of reporting or withholding in ways that improve their metacognitive judgement calibration on future trials. Other (more applied) studies have investigated the efficacy of training learners to self-regulate their learning processes by instructing, or modelling, different self-regulated learning strategies (Kostons et al., 2012; Leopold & Leutner, 2015; Zimmerman & Schunk, 2012). For example, Leopold and Luetner found that instructing students to self-question whether they

successfully distinguished between relevant and irrelevant information improved discrimination on a learning task. In sum, metacognitive training strategies can benefit how well participants regulate reasoning strategies in a future task.

Research on whether training influences how well people regulate their initial metacognitive processes is minimal. Baars et al. (2014) found that metacognitive training via modelling use of a strategy did not improve how calibrated students' prospective judgements were about items on a biology quiz (i.e., whether they correctly evaluated if they would answer correctly). To assess whether training influences initial meta-reasoning, the experiments reported in Chapter 2 took a different approach. Here, participants were given longer-duration anagrams at first (16 s), which halved over subsequent blocks of trials (8s, 4 s, 2 s; *training group*). I expected that experience with longer-duration problems would enable participants to better regulate their JOSs at the shorter durations, compared to when anagram duration was always brief (*no-training group*), thus improving JOS discrimination and predictiveness.

In addition, I expected that participants would solve some of the longer-duration problems, which would also inform their JOSs on future trials. Given that intuition precedes solving, a 'solvable' JOS should precede the solution on these trials (Ackerman & Thompson, 2017). Thus, solving the anagram during the JOS task provides the reasoner with feedback that their intuition was correct, which participants can then use to regulate their future solvability intuition (Finn & Metcalfe, 2007; Griffin et al., 2009). When reasoners are given more time to make JOSs, their decisions should be more accurate because they have more time for cues about solvability to develop, such as possible representations of anagram solutions (Arnold et al., 2013; Engeler & Gilbert, 2020; Lichtenstein & Fischhoff, 1980), which they might regulate on future JOS trials. In sum, Chapter 2 investigated the impact of

longer-duration anagrams on meta-reasoning performance to identify possible processes underlying self-regulated JOS decision-making.

JOS Study Design Factors

Chapter 2 also examined whether JOSs discriminate and predict problem-solving success in a blocked design, as used by Ackerman and Beller (2017) and Lauterman and Ackerman (2019). In a blocked design, participants make JOSs for the entire set of reasoning problems in one block (*JOS phase*) and then attempt to solve each of the problems in a second block (*solving phase*). The JOS phase captures solvability intuition, and the solving phase captures how well solving outcomes are predicted by those JOSs.

In contrast, other studies have used an interleaved design, such that after reporting their JOS, participants are immediately prompted to solve the problem (e.g., Balas et al., 2011; Bolte & Goschke, 2005; Markovits et al., 2015; Novick & Sherman, 2003; Siedlecka et al., 2016). In an interleaved design, JOSs and solving attempts may impact each other in ways that influence how discriminating and predictive JOSs are. Memory for the JOS is readily available in this design, given that participants make their JOS right before their solving attempt. Greater availability of the JOS in an interleaved design may motivate participants to regulate their solving effort to align with their JOSs (i.e., they might exert more effort on a problem they judged as “solvable”, and less effort on a problem they judged as “not solvable”), which would generate more predictive JOSs if more effort leads to more solving successes (Pennycook et al., 2015a). Moreover, interleaved designs provide participants with solving experiences after each JOS. These solving experiences may increase a reasoner’s sensitivity to problem solvability and provide feedback about their JOS accuracy. For example, when a participant judges an anagram as “solvable”, solving the anagram on the following solving trial would confirm that their intuition was correct (Griffin et al., 2009). In line with these theories, studies that have interleaved JOSs with solving have tended to find

that JOSs are discriminating and predictive. Thus, solvability intuition may be influenced by the type of study design used to measure JOSs.

The experiments in Chapter 2 separated the influence of solving attempts from JOSs using a blocked design, to capture reasoners' naïve intuitive judgements. Using the blocked design, I was able to examine whether JOSs are discriminating and predictive when fewer cues are available to make a JOS (such as feedback from deliberate solving attempts).

Anagram solving was expected to occur often at longer durations in the training group. Even so, JOSs would still capture intuition on trials that did not lead to a solution during the JOS task. At shorter durations, the JOS would capture intuition before a solving attempt has begun. Thus, the blocked design allowed me to test whether JOSs were discriminating and predictive in the absence of purposeful anagram-solving experience. However, my use of a blocked design was not able to test the assumption that intuitive JOSs would be more naïve relative to an interleaved design. Therefore, the experiment reported in Chapter 4 compared JOS discrimination and predictiveness in interleaved versus blocked designs.

Self-Regulated versus Time-Limited Solving Trials

A fourth experiment factor that may influence how well JOSs predict problem-solving outcomes is whether the solving trials are time-limited or self-regulated. In this thesis, I measured JOS discrimination using two different solving phases. Experiments 1 and 2 (Chapter 2) used a *time-limited* solving phase, such that the participants were told that the solving phase would re-present only the solvable anagrams from the JOS phase, and participants would have a maximum of 45 s to solve each one – hence, solving attempts were regulated by the experimenter. Experiment 3 (Chapter 2) and Experiment 4 (Chapter 4) used a *self-regulated* solving phase, in which participants attempted to solve both solvable and unsolvable anagrams from the JOS phase, and could spend as much time solving each anagram as they wished. Participants could also decide to pass on their solving attempt if they

believed the anagram was solvable but could not solve it or submit a final “not solvable” response if they believed the anagram was unsolvable. Hence, solving attempts were regulated by the participant. Furthermore, the final ‘pass’ and ‘not solvable’ responses in the self-regulated solving phase served as a final JOS (Ackerman & Beller, 2017; Ackerman & Thompson, 2017; Lauterman & Ackerman, 2019) and provided more information about how well JOSs predicted self-regulated solving attempts. For example, using the self-regulated solving trials, I was able to examine whether ‘not solvable’ JOSs were more likely to result in final ‘not solvable’ responses on solving trials than ‘solvable’ JOSs.

Time-limited and self-regulated solving trials were both expected to provide useful information about how well JOSs predict successful problem-solving. Payne and Duggan (2011) found that participants persisted for longer on certain problems when they were informed that the probability of solving them was high. Therefore, when participants are aware that each anagram in the solving phase is solvable, they may persist more on these anagrams, which may generate more solving successes. Consequently, there may be fewer differences in predictiveness between each JOS, as solving efforts will be informed less by what JOS the anagram was assigned and more by participants’ perceived likelihood of solving each anagram (which may be high given that each anagram is solvable).

Self-regulated solving trials should lower rates of successful solving and may sponsor more predictive JOSs due to the added uncertainty regarding the solvability of each item. In the self-regulated solving phase, participants are aware that not all their solving attempts will be successful. Given that reasoners are unlikely to invest effort in problem-solving when the probability of success is low (De Neys et al., 2013; Payne & Duggan, 2011), a self-regulated solving phase should produce more solving failures (that is, pass or not-solvable responses) for more of the solvable anagrams, particularly if a reasoner intuitively felt the anagram is unsolvable. Rather than using external information regarding the probability of solving each

item, participants in Experiment 3 of Chapter 2 (and in Experiment 4 in Chapter 4) received self-regulated solving trials, where it was expected that their JOS intuitions would influence how they regulated their solving effort.

The Role of Cognitive Reflection in Judgements of Solvability

Individual differences in *cognitive reflection* might also contribute to meta-reasoning ability. From a dual-process account, to reason accurately one may need to identify when System 1 reasoning has generated errors (or *conflict*) and then override the intuitively generated response with analytic thinking by using System 2. The term “cognitive reflection” refers to a disposition to override the intuitively generated prevailing response (i.e., System 1 thinking), and instead use analytic thinking to reason to the correct response (i.e., System 2 thinking).

The 7-item Cognitive Reflection Test (CRT; Frederick, 2005; Toplak et al., 2014) is a verbal reasoning test designed to measure an individual’s propensity to engage in reflective thinking. What differentiates the CRT from typical cognitive ability tasks (e.g., IQ tests) is its use of “trick” questions which are designed to cue an initial incorrect response (Frederick, 2005; Pennycook et al., 2015b). For example, the CRT item, “*In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days to cover the entire lake, how long would it take for the patch to cover half of the lake?*” cues the incorrect response of 24 days. To arrive at the correct response of 47 days, reasoners need to be able to use *smart deliberation* (Raoelison et al., 2020) to deliberate on their initial incorrect response and detect errors, and instead use analytic thinking to calculate the correct answer (De Neys & Pennycook, 2019).

There is some evidence that cognitive reflection ability is related to more accurate meta-reasoning calibration for final confidence judgements (Mata et al., 2013; Pennycook et al., 2017). For example, Pennycook et al. (2017) found that individuals lower in cognitive

reflection were more likely to overestimate their accuracy on a reasoning task, whereas those higher in cognitive reflection were more likely to correctly evaluate their accuracy. Thus, more reflective thinkers may possess a better innate error-detection ability, which involves the ability to detect biases in reasoning (Pennycook et al., 2014; Pennycook et al., 2015a). Some researchers have found the CRT to be a reliable predictor of metacognitive ability because of its ability to separate those who can detect errors in their reasoning from those who rely on biased intuition (Bialek & Pennycook, 2017). Although some studies have found that cognitive reflection predicts how well people meta-reason about their reasoning outcomes (Coutinho et al., 2021; Duttler, 2016; Mata et al., 2013; Noori, 2016; Pennycook et al., 2017), research has yet to investigate whether more reflective thinkers can also make more accurate and discriminating JOSs. To this end, I re-analysed JOS discrimination and predictiveness from Experiments 2 and 3 with consideration of participants' CRT scores.

In Chapter 2, I suggest that reasoners may regulate their future JOS accuracy by using instances of solutions found during the JOS to test whether their intuitions were correct. Perhaps then, those with a stronger disposition to “decouple” from erroneous intuition, and instead engage analytic thinking, are more responsive to this feedback. For example, a reasoner might recognise that their intuitions about solvability were incorrect after noticing a miscalibration between their intuitions and when a solution was spontaneously retrieved on a JOS trial. Reflective thinkers may be more inclined to reflect on their JOSs, and to realise that they were relying on biased intuitions about solvability. Hence, they may then be better able to readjust their intuition to meet the goals of the task, resulting in more accurate and predictive JOSs. If individual differences in cognitive reflection determine reasoners' ability to make discriminating and predictive JOSs, this might be more likely to occur when anagram durations are *shorter* (and thus more susceptible to biased responding). Thus, in addition to investigating whether cognitive reflection is related to more discriminating and predictive

JOSs, Chapter 3 also sought to determine whether longer versus shorter duration anagrams influenced how strongly cognitive reflection was related to JOS discrimination and predictiveness.

Thesis Overview

The overarching goal of my thesis was to contribute to our understanding of Ackerman and Thompson's (2017) first stage of meta-reasoning—the Judgement of Solvability (JOS). To this end, I conducted an in-depth investigation of JOSs for anagrams, focusing on whether they discriminate between solvable and unsolvable anagrams and whether they predict later problem-solving outcomes and effort regulation. To clarify the mixed findings for JOS discrimination and predictiveness in the literature, I considered several JOS task factors: (1) whether solutions found during the JOS process potentially exaggerate JOS discrimination and predictiveness, (2) whether JOSs are more discriminating and predictive when reasoners are given more time to develop their intuition, (3) whether JOSs discrimination and predictiveness can improve with training, (4) whether JOSs are more predictive of problem-solving outcomes when solving trials are self-regulated, (5) whether study design influences how well JOSs discriminate and predict problem-solving success and (6) whether cognitive reflection ability facilitates better initial meta-reasoning.

Using a blocked design, Experiment 1 sought to clarify whether intuitive JOSs were discriminating and predictive after excluding anagrams that had been solved during the JOS process, and whether JOSs were more discriminating and predictive with longer (versus shorter) anagram durations. In the JOS phase, participants were presented with a series of solvable and unsolvable anagrams, one at a time, and judged each anagram as *solvable*, *not solvable*, or *already solved* (*S*, *NS*, *AS*) over 4 blocks. Participants then completed a solving phase in which they attempted to solve some of the anagrams, providing our measure of whether JOSs predicted problem-solving successes or failures.

In the JOS phase of Experiment 1, anagram duration was manipulated within-subjects. Participants made JOSs for anagrams presented at longer durations initially, which halved over subsequent blocks (16 s, 8 s, 4 s, 2 s). The longer-duration blocks were intended to provide participants with some solving successes, to motivate them to provide rational JOSs (without reverting to guesses or unreliable heuristics). Experiment 1 used a time-limited solving phase – the solving phase presented each of the *solvable* anagrams again, and participants were given 45 s to attempt to solve each one.

However, Experiment 1 was not able to determine whether more accurate JOS discrimination in the final 2 s block occurred due to training from the longer duration blocks. Therefore, Experiment 2 included a training versus no-training between-subjects factor. The training group was identical to Experiment 1; anagram durations in the JOS phase started at 16 s in Block 1 and halved across each subsequent block. In the no-training group, anagram duration was just 2 s in each of 4 blocks. I expected that JOSs would be more discriminating and predictive when participants were given longer durations to assess solvability. Moreover, I anticipated that participants would more accurately regulate their JOSs in the final 2 s block of the training group by having more experience/practice with longer-duration anagrams, compared to when anagram duration was just 2 s across blocks. Thus, Experiment 2 addressed whether JOS discrimination and predictiveness were trainable, as well as establishing the replicability of the results of the training group in Experiment 1.

Experiment 3 used a self-regulated solving phase to examine how well JOSs predicted problem-solving outcomes across training and no-training groups. Here, participants were presented with both solvable and unsolvable anagrams and were informed that they could spend as much time as they liked solving each anagram. Participants also had the option of skipping a solving trial if they believed the anagram was unsolvable. These modifications to the blocked design in Experiment 3 also provided new measures about the relationship

between JOSs and problem-solving effort regulation. Specifically, I was able to measure final “not solvable” responses to anagrams in the solving phase (and whether these were predicted by JOSs), as well as response times for solutions or final “not solvable” responses. To this end, Experiment 3 addressed how well JOSs are predictive of problem-solving outcomes using a self-regulated solving phase, and the impact of training on JOS discrimination and predictiveness when the solving phase was self-regulated.

Chapter 3 investigated the potential impacts of cognitive reflection on JOS discrimination and/or predictiveness. Given prior evidence that individual differences in cognitive reflection can impact reasoning and meta-reasoning, here the data from Experiments 2 and 3 were further analysed using participants’ CRT scores, to test whether cognitive reflection facilitates accurate and predictive JOSs. Additionally, I also investigated the effect of cognitive reflection on JOS predictiveness varied depending on whether solving is self-regulated versus time limited.

Chapter 3 also examined whether the inclusion of longer-duration anagrams in the training group mitigated the effects of cognitive reflection on JOS discrimination and predictiveness. If individuals with greater cognitive reflection are innately better at insight reasoning and are less susceptible to biased intuitions about solvability, then they should show more accurate and predictive JOSs when anagram duration is just 2 s in each block. However, inclusion of longer-duration anagrams should provide less-reflective participants valuable practice and experience in making meta-reasoning judgements about solvability, which may diminish the impact of cognitive reflection on JOS discrimination and predictiveness.

Finally, Chapter 4 examined how study design contributes to how well JOSs discriminate and predict later self-regulated problem-solving. Studies measuring JOS discrimination and predictiveness have either interleaved JOSs with problem-solving attempts or have blocked JOSs and solving attempts into two separate phases. Considering that the

studies that find JOSs to be discriminating and predictive interleaved JOSs with solving, I expected that study design might influence JOS discrimination and effort regulation, and in turn whether JOSs predict problem-solving performance.

Experiment 4 manipulated study design (blocked vs. interleaved) between groups. I also included anagram duration (2 s vs. 4 s) as a second between-groups manipulation, to examine the generality of any findings between interleaved and blocked designs. This manipulation also allowed investigation of whether JOS discrimination and predictiveness in each study design varied as a function of having more (versus less) time for participants to develop their intuition. To this end, Experiment 4 addressed my research question about how JOS study design influences JOS discrimination and predictiveness.

Chapter 2: How Accurate and Predictive are Judgements of Solvability?

Explorations in a Two-Phase Anagram Solving Paradigm

Author contributions: GEB, MMA, and I conceptualised the study design. I programmed the experiment and collected the data, cleaned the data for analysis, and performed the data analyses. GEB and PW both advised me on which data analyses to carry out and I interpreted the data under their guidance. I drafted the manuscript and GEB provided critical revisions. PW provided critical revisions on the Results sections. GEB approved the final version of the manuscript for submission.

Abstract

Meta-reasoning requires monitoring and controlling one's reasoning processes, and it often begins with an assessment of problem solvability. We explored whether *Judgements of Solvability (JOS)* for solvable and unsolvable anagrams discriminate and predict later problem-solving outcomes once anagrams solved during the JOS task are excluded. We also examined whether providing training via longer-duration anagrams improves JOS discrimination and predictiveness. In a two-phase paradigm, participants judged each anagram as *solvable*, *not solvable*, or *already solved* (*S*, *NS*, *AS*; *JOS phase*) then later attempted to solve the anagrams within 45 s (*solving phase*). Anagrams were presented in 4 blocks. In the *training groups*, anagram duration started at 16 s and halved across blocks, whereas in the *no-training groups* anagram duration was always 2 s. Participants' S JOSs typically were discriminating after excluding anagrams that received AS JOSs, but training did not lead to better discrimination in the final block. Training improved AS JOS predictiveness, but not S JOS predictiveness. Thus, training increased solving during the JOS process rather than increasing JOS predictiveness. In Experiment 3 these findings replicated when both solvable and unsolvable anagrams were presented in the solving phase and no response deadline was set. Here, problem-solving outcomes and effort regulation (i.e., response times) were

predicted by AS and NS JOSs, but not by S JOSs. Overall, although S JOSs were discriminating, they were not predictive of later problem solving or effort regulation—and this was true even after training with longer-duration anagrams.

Introduction

Reasoning refers to the cognitive processes engaged during methodical problem solving. An important component of reasoning is meta-reasoning, which involves assessing the quality of one's judgements and cognitive processes (Metcalf & Wiebe, 1987; Salvi et al., 2016). Though metacognition and meta-reasoning are both concerned with awareness of one's cognitive processes, metacognition research has focussed on memory and general knowledge (Berardi-Coletta, 1995), whereas meta-reasoning research has focussed on monitoring and control processes during reasoning and problem solving (Ackerman & Thompson, 2017). One's ability to meta-reason is important because it informs decisions about whether to engage in problem solving, about effort investment during problem solving, and retrospective confidence about solving outcomes (Ackerman, 2014; Ackerman & Thompson, 2017; Payne & Duggan, 2011).

According to Ackerman and Thompson (2017), the first stage of meta-reasoning involves making a *Judgement of Solvability* (JOS). A JOS indicates one's beliefs about whether a problem is solvable and/or whether one can solve it (Ackerman & Beller, 2017; Metcalfe & Wiebe, 1987). JOSs inform decisions to either engage in the problem or to give up on it. Misjudging problem solvability can have negative consequences such as wasting time attempting to solve unsolvable problems, or prematurely abandoning solvable problems (Ackerman & Thompson, 2017; Payne & Duggan, 2011; Toplak et al., 2014). To further our understanding of this first stage of meta-reasoning, our study sought to provide a detailed investigation of JOSs, focussing on whether they are discriminating and predictive of later problem-solving outcomes.

Are JOSs Discriminating?

Intuitive judgements can be made quickly and accurately and without analytic engagement (Kahneman, 2003; Lieberman, 2000). However, the evidence regarding whether

intuitive JOSs are sensitive to a problem's actual solvability has been mixed. Several studies have found that JOSs can discriminate between solvable problems (e.g., Balas et al., 2011; Bolte & Goschke, 2005; Novick & Sherman, 2003; Topolinski et al., 2016; Topolinski & Strack, 2009a; Undorf & Zander, 2017). Other studies have failed to find such effects (e.g., Ackerman & Beller, 2017), or have reported that JOSs are discriminating only when certain problem-solving task conditions are met (e.g., Lauterman & Ackerman, 2019; Valerjev & Dujmović, 2020). Whether JOSs are found to be discriminating may depend in part on how investigators treat problems that are spontaneously solved during the JOS task. Often, researchers use “insight” problems to measure JOSs—these are short, verbal problems (such as anagram solving or the remote associates task) which can be solved in a sudden, non-incremental way (Bowden & Jung-Beeman, 2003; Weisberg, 1992). Solutions to insight problems are usually found without much deliberate analytic engagement (Metcalf, 1986; Metcalf & Wiebe, 1987). Where insight differs from intuition is that insight involves retrieval of the solution, whereas intuition is based on a “gut feeling” that a solution to the problem exists (Stanovich & West, 2000; Topolinski & Strack, 2009a). An advantage of using insight problems to measure JOSs is that such problems can be processed and solved rapidly. Use of analytic problems typically requires more solving time (De Neys, 2006), which limits the number of JOSs that can be captured in a single experiment without fatiguing participants (Healy et al., 2004). Thus, we used anagrams as our problem-solving task, which allowed us to capture more JOSs than would be possible in an experiment that used analytic problems.

Despite the merits of using insight problems to measure JOS discrimination, solutions to insight problems can arise spontaneously during the JOS task (e.g., Novick & Sherman, 2003). Consequently, significant JOS discrimination may be attributable at least in part to participants spontaneously solving some of the problems during their presentation, rather than

because they had accurate intuitions about their solvability. JOSs are intended to capture participants' intuitions before reasoning occurs (Ackerman & Thompson, 2017), thus spontaneous solutions arising before/during the JOS would confound the measurement of JOS discrimination. For instance, Topolinski et al. (2016, Experiment 7) found that JOS discrimination was only marginally significant when anagrams that participants spontaneously solved during the JOS task were excluded from analysis. Therefore, our study reassessed whether JOSs are discriminating when problems solved during the JOS process are excluded.

Do JOSs Predict Later Problem Solving?

People generally avoid expending cognitive effort on problems they deem themselves unlikely to solve (De Neys et al., 2013; Payne & Duggan, 2011)—a process known as effort regulation. If intuition about problem solvability guides decisions about effort regulation, then people should exert more time and effort solving problems that they deem to be solvable (Ackerman & Thompson, 2017). Because longer processing time is associated with better reasoning performance (e.g., Pennycook et al., 2015), greater effort expenditure should lead to more successful problem solving.

Surprisingly, the few studies that have evaluated whether JOSs predict problem-solving success and effort regulation have yielded mixed findings. Judging a problem as solvable (vs. unsolvable) has been found to predict successful problem solving in some studies (Markovits et al., 2015; Siedlecka et al., 2016), but not others (Ackerman & Beller, 2017; Lauterman & Ackerman, 2019). Moreover, Lauterman and Ackerman found that participants who judged a problem as solvable later spent more time attempting to solve it, regardless of its actual solvability.

The same methodological issues noted above for measuring JOS discrimination apply equally to measuring JOS predictiveness. Specifically, spontaneous problem solving during

the JOS task will exaggerate how well JOSs predict later problem solving. This issue may contribute to the mixed findings regarding JOS predictiveness. Thus, another aim of our study was to evaluate whether JOSs are predictive of effort regulation and problem-solving success after accounting for spontaneously solved items.

Approaches to Measuring JOS Discrimination and Predictiveness

To measure JOSs, researchers typically aim to choose a problem duration that will limit spontaneous solving during the JOS task. However, if the problems are presented too briefly, participants may revert to using unreliable heuristic cues to make their decisions (Ackerman, 2019; Benjamin, 2005; Kahneman et al., 1982), or may even engage in random responding that would reduce the accuracy and predictiveness of JOSs. An alternative is to provide problems for longer but to allow participants to report whenever they have solved a problem during its presentation. This method enables the researcher to examine whether JOS discrimination and predictiveness is limited to solved problems or extends to unsolved problems. A second advantage of this method is that participants can be given more time to make their JOSs without them outsourcing their cognitive efforts to unreliable heuristics.

Some studies have ignored the possibility of spontaneous solutions or have merely assumed that the selected problem duration prevents them (e.g., Balas et al., 2011; Bolte & Goschke, 2005; Siedlecka et al., 2016; Valerjev & Dujmović, 2020). Topolinski et al. (2016; Experiments and 7) instructed participants to report any spontaneous solutions to the anagram after each JOS trial, and then discarded trials where participants had solved the anagram. However, their participants only provided solutions to spontaneously solved problems; JOS predictiveness was not measured. Thus, participants' ability to solve the problems they judged as 'solvable' (but did not report having spontaneously solved) was not assessed.

Most studies that have not measured spontaneous solving have interleaved the JOS and problem-solving tasks (e.g., Balas et al., 2011; Topolinski & Strack, 2009; Valerjev &

Dujmović, 2020), such that on each trial participants made a JOS and then immediately attempted to solve the problem. In an interleaved paradigm, JOSs may be influenced by solving attempts, and vice versa. For instance, if a participant judges a problem as solvable, and then solves the problem, that serves as metacognitive feedback that the JOS was well calibrated. As a result, interleaved paradigms may lead to higher levels of JOS discrimination because participants can adjust their JOS calibration in light of their solving outcomes. Additionally, participants may exert more effort solving problems they have judged to be solvable, leading to more success and thus also rendering JOSs more predictive. In short, interleaved paradigms may allow reasoners to bootstrap their JOS intuitions, which in turn might improve JOS discrimination and predictiveness.

In contrast, other studies have used a two-phase paradigm. In a *JOS phase*, participants make JOSs for the entire set of reasoning problems. The JOS phase is then followed by a *solving phase*, in which participants attempt to solve some or all of these problems (e.g., Ackerman & Beller, 2017; Lauterman & Ackerman, 2019). The JOS phase is intended to capture intuitive judgements, and the solving phase is intended to capture solving outcomes and how well they are predicted by JOSs. Our study used the two-phase paradigm.

Is JOS Discrimination and Predictiveness Trainable?

To date, studies examining JOSs have focussed on identifying factors that may influence or bias JOSs, such as problem length, difficulty, and fluency (e.g., Balas et al., 2011; Lauterman & Ackerman, 2019; Topolinski et al., 2016; Valerjev & Dujmović, 2020). Research has yet to examine whether JOSs are trainable in ways that increase how discriminating they are, and how predictive they are of later problem solving. The influence of training on metacognition has largely occurred in the metamemory area (e.g., Dunlosky & Rawson, 2012; Koriat et al., 2002; West & Mulligan, 2019) . Our study examined the impact

of training on meta-reasoning, by measuring whether practice with longer-duration anagrams in the JOS phase enhances JOS discrimination and/or predictiveness.

Overview

We examined JOS discrimination and JOS predictiveness using a two-phase paradigm. Anagrams were used to allow the collection of brief assessments of solvability. In the JOS phase, equal numbers of solvable and unsolvable anagrams were presented in each of four blocks. In the training groups, the first block presented each anagram for 16 s, and anagram presentation duration was then halved across the three subsequent blocks. The training group allowed us to parametrically examine the effect of duration on JOS discrimination and predictiveness. This design resulted in the briefest blocks using 2 s and 4 s anagram durations, consistent with the durations used in prior studies (e.g., Lauterman & Ackerman, 2019; Novick & Sherman, 2003; Topolinski et al., 2016). After the anagram disappeared, participants quickly judged the anagram as either solvable, unsolvable, or already solved.

For the blocks with longer-duration anagrams, participants are likely to move from simply making a JOS to attempting to solve the anagrams. This should result in a higher rate of ‘already solved’ JOSs. Nonetheless, ‘solvable’ JOSs should accurately capture intuition regardless of whether solving efforts have not yet occurred (shorter-duration anagrams) or have occurred but have not yielded solutions (longer-duration anagrams). Starting with longer-duration anagrams was expected to provide the training groups with more solving successes that might increase participants’ motivation to provide rational JOSs, help them generate a better intuitive sense of an anagram’s solvability (Schuster et al., 2020), and help them to regulate their JOSs (Leopold & Leutner, 2015; Leutner et al., 2007). Examining JOS discrimination and predictiveness across a range of durations, rather than choosing an arbitrary “gold standard” duration, also served to increase generality. In the solving phase,

participants then attempted to solve each of the solvable anagrams within 45 s (Experiments 1 and 2), or they received both solvable and unsolvable anagrams and solving time was self-regulated (Experiment 3).

We report three experiments. Experiment 1 determined whether JOSs in a training group were discriminating and predictive after excluding anagrams classified by the participant as ‘already solved’ during the JOS task. We also examined how anagram duration affects JOS discrimination and the rate of ‘already solved’ JOSs. Experiment 2 compared the training group to a no-training group that consistently received short (2 s) duration anagrams, to allow us to measure the effect of longer-duration training. In Experiment 3, we modified the two-phase paradigm to allow effort-regulation and solving performance to vary in the solving phase, to examine whether JOSs predict self-regulation of effort investment in anagram solving. Here, we included both solvable and unsolvable anagrams in the solving phase, and no time limit was imposed on solving. Our experiments build on Topolinski et al.’s (2016) initial explorations of JOSs by measuring and considering ‘already solved’ JOSs, by examining the links between JOSs and later solving outcomes, and by exploring the effects of training using longer-duration anagrams.

Experiment 1

Experiment 1 explored whether already-solved (*AS*) and solvable (*S*) JOSs in the JOS phase discriminate between solvable and unsolvable anagrams. The subsequent solving phase allowed us to explore whether these JOSs predicted successful problem solving, as well as whether not solvable (*NS*) JOSs predicted problem-solving failures. During the JOS phase, anagrams were presented for 16 s in block 1, 8 s in block 2, 4 s in block 3, and 2 s in block 4. In the solving phase, participants attempted to solve each solvable anagram within 45 s.

Both *AS* and *S* JOS were expected discriminate solvable from unsolvable anagrams, and JOS discrimination was expected to decrease across blocks as anagram duration

decreased. We also expected that JOSs would be more discriminating when anagrams receiving AS JOSs were included in the discrimination measures than when they were excluded. In turn, we expected that anagrams receiving AS JOSs typically would be solved in the solving phase—indeed, this creates a manipulation check that participants used the AS JOS response option appropriately. We also evaluated whether anagrams receiving S JOSs were associated with greater solving-phase success, and whether anagrams receiving NS JOSs were associated with lower solving-phase success.

Method

The experiment was pre-registered on Open Science Framework (OSF) at <https://osf.io/zuqnw>.

Participants

Participants ($N = 122$) were recruited through Amazon's Mechanical Turk (MTurk) via TurkPrime (Litman et al., 2017) and each received USD \$2.25. We excluded 22 participants who met more than one pre-registered exclusion criterion (correctly solved less than 10% of anagrams, did not complete the study, failed an attention check, more than 2 SD outside the mean study completion time). The final sample was 100 participants (55 female, 44 male, 1 other; mean age = 39.76, $SD = 12.35$), in line with our pre-registration.

Stimuli

Because anagram solving depends to some degree on how frequently the solution word appears in the English language (Johnson, 1966; Mayzner & Tresselt, 1958), a set of 75 solvable 5-letter anagrams was selected from a corpus of frequently used words (*Word Frequency Data*, 2016). The anagrams were subject to Gilhooly's (1978) bigram analysis indicating each word had a single anagram solution. The anagrams were piloted online ($N = 94$) and 40 were selected to be used in the study. The anagrams were then sorted into 4 sets of 10 roughly equated on solvability (each block had a mean solving rate of roughly 77% and

solving rates ranged from 50% to 100%). To create the set of 40 unsolvable anagrams, the letters in pseudowords created using Wuggy (Keuleers & Brysbaert, 2010) were randomly shuffled using an online character randomizer (*Shuffle Characters in Text*, 2010), and were then randomly and evenly allocated to the 4 sets. Assignment of sets to blocks was counterbalanced across participants via Latin square.

Procedure

The experiment was conducted online using Qualtrics software (Qualtrics, 2019). For the JOS phase, participants were instructed that they would be presented with a sequence of letters on each trial (i.e., an anagram), some of which could be rearranged to spell a word (e.g., DSTMI - MIDST) and hence were ‘solvable’, and others of which did not have a solution word (e.g., ZEREB) and hence were “unsolvable”. Their task was to make one of three solvability judgements for each anagram in the allotted time: “YES it is solvable”, “NO it is not solvable”, or “I have already solved it”. They were told that the anagram duration would decrease across four blocks as follows: 16 s, 8 s, 4 s, 2 s. Participants were also forewarned that they would later have 45 s to attempt to solve each solvable anagram.

On each of the 80 JOS phase trials, an anagram was presented for the duration specified for that block. Once the anagram disappeared, the 3 JOS options appeared as response boxes, and participants had 3 s to click on a response. If they failed to make their JOS within 3 s, a message appeared asking them to respond within 3 s. This message remained on the screen for 4 s to discourage participants from continuing to try to solve anagrams after they disappeared. After making their JOS, participants pressed an arrow button to submit their response, and then the next trial began. If participants made a JOS but did not submit it within 3 s it was still recorded; this occurred on an average of 0.4% of trials in Experiment 1, 1.3% in Experiment 2, and 0.7% in Experiment 3. Before commencing the task, participants

completed 10 practice JOS trials (5 solvable, 5 unsolvable) at the 16 s duration. They then attempted to solve the 5 solvable anagrams, each within 45 s.

The solving phase immediately followed the JOS phase. The solvable anagrams from the JOS phase were presented sequentially in a random order, each for 45 s (due to a programming error, only 39 of the 40 solvable anagrams were presented). Participants had 45 s to type the solution into a response box (minimum allowed was 3 s) and to then press the “Next” button to proceed. The 45 s time limit was selected based on a pilot study with a 60 s time limit in place; here the mean response time plus 2 SD was roughly 45 s, so this time limit ensured adequate solving time for the majority of trials/participants. On average, responses to anagrams were made within 45 s on 93% of trials in Experiment 1 and 94% of trials in Experiment 2 (among retained participants). If participants did not respond within 45 s, any response they entered was recorded and the solving phase progressed.

Results

JOS Phase

Participants’ ability to distinguish solvable anagrams from unsolvable anagrams (i.e., JOS discrimination) was assessed by measuring whether their hit rate (i.e., judging a solvable anagram to be solvable) exceeded their false alarm rate (i.e., judging an unsolvable anagram to be solvable). Hits and false alarms were converted to proportions by dividing them by the total number of JOS phase trials in which participants entered a response within the 3 s time limit following anagram presentation. as would be the case if participants were not offered the option of reporting spontaneous solving, (Figure 2.2a) and S JOSs (Figure 2.3a).

Figure 2.1

Experiment 1: Mean Proportions of Hits and False Alarms for AS+S JOSs in the JOS Phase

in Experiment 1 (Bars show 95% CI of each mean)

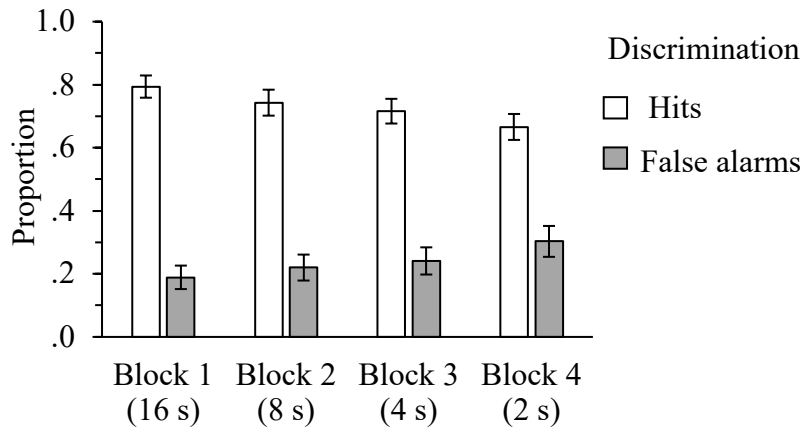


Figure 2.2

Experiments 1-3: Mean Proportions of Hits and False Alarms for AS JOSs in the JOS Phase

(Bars show 95% CI of each mean)

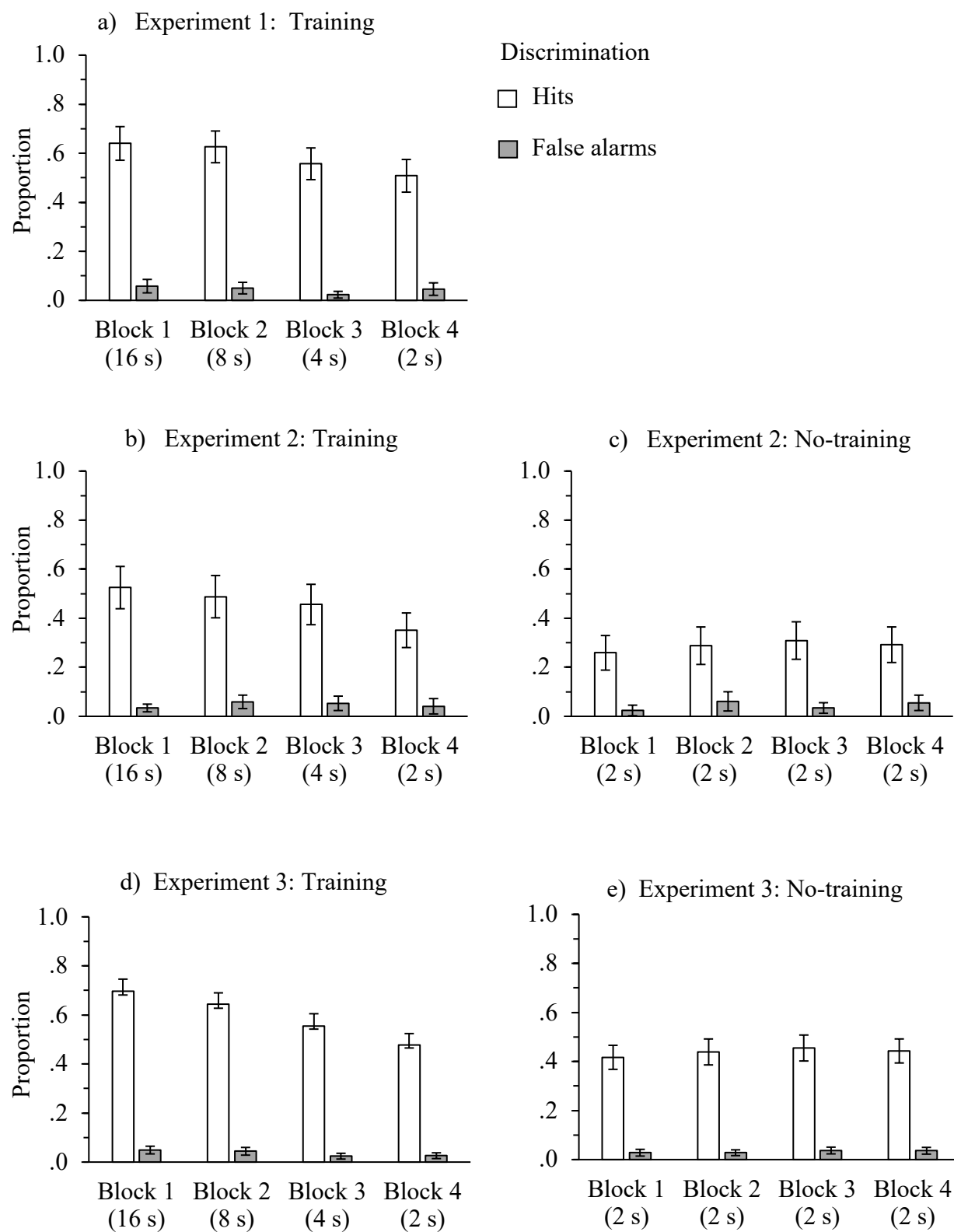
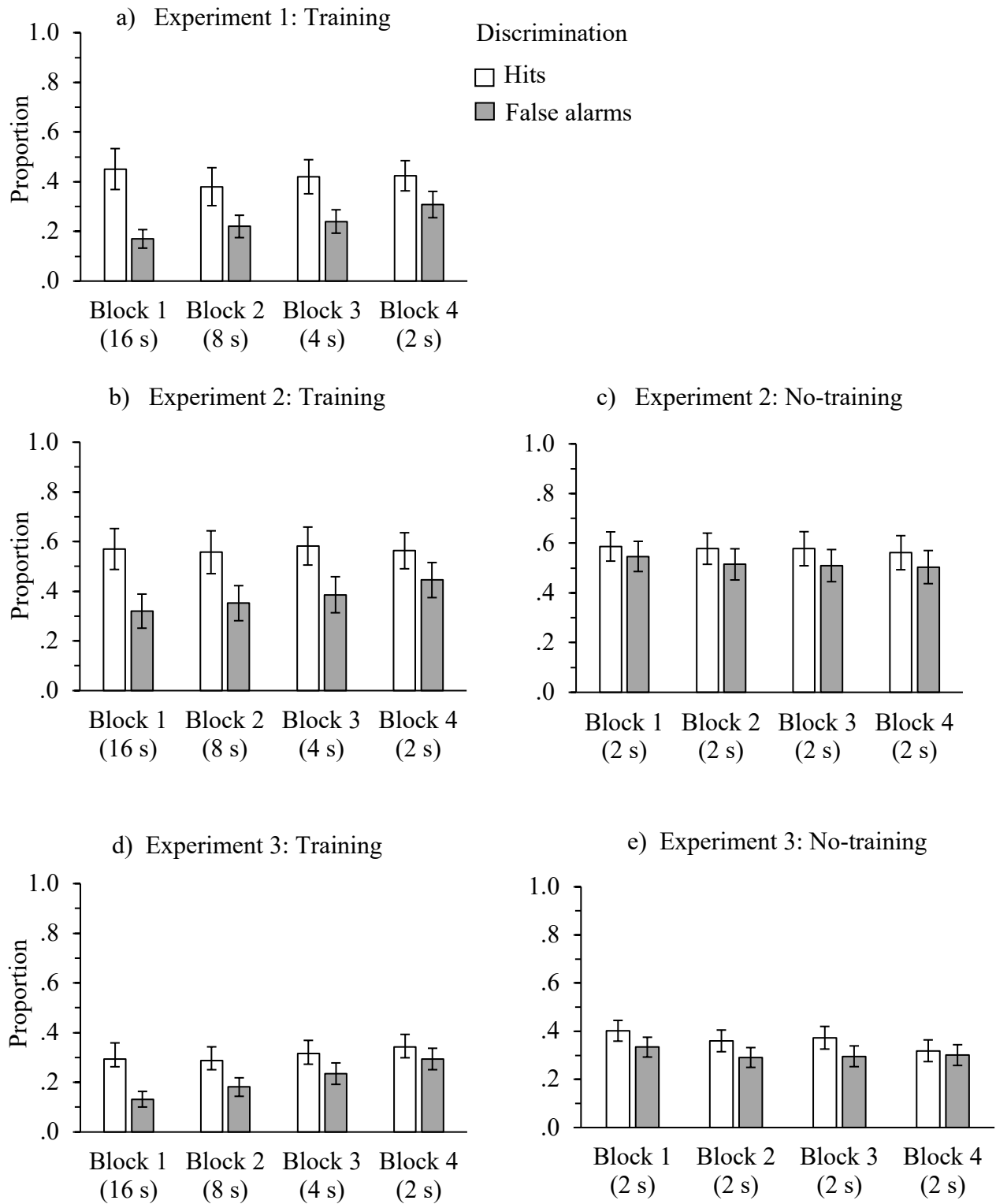


Figure 2.3

Experiments 1-3: Mean Proportions of Hits and False Alarms for S JOSs in the JOS Phase

(Bars show 95% CI of each mean)



Each dependent variable was analysed using a 2(discrimination: hits, false alarms) \times 4(block: 1-4) repeated-measures ANOVA. Table 2.1 provides the complete ANOVA results. The two key effects reviewed below are whether each JOS discriminated solvable from unsolvable anagrams (i.e., the main effect of discrimination) and whether JOS discrimination decreased across blocks (i.e., the interaction between discrimination and block).

Table 2.1

Experiment 1: JOS Phase ANOVAs Results

JOS(s)/Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS+S JOSs					
Discrimination	1, 99	48.26	430.63	< .001	.81
Block*	2.79, 275.98	0.01	0.33	.79	.003
Discrimination \times Block*	2.83, 280.11	0.54	24.18	< .001	.20
AS JOSs					
Discrimination	1, 91	53.41	331.15	< .001	.78
Block	2.52, 229.39	0.27	10.84	< .001	.11
Discrimination \times Block	3, 273	0.14	9.47	< .001	.09
S JOSs					
Discrimination	1, 82	5.58	46.79	< .001	.36
Block*	2.36, 193.66	0.17	3.55	.02	.04
Discrimination \times Block	3, 246	0.20	5.78	.001	.07

Note. * Huynh-Feldt correction was applied because assumption of sphericity was violated.

Participants' AS+S JOSs were highly discriminating; averaged across blocks, hits ($M = .73$, $SD = .16$) were significantly greater than false alarms ($M = .24$, $SD = .18$). JOS discrimination also interacted with block, reflecting reduced discrimination across blocks as anagram durations were reduced (see Figure 2.1); however, pairwise comparisons showed that discrimination was significant in each block ($ps < .001$). The interaction was followed up

using linear contrasts, given our parametric manipulation of anagram duration. The results are shown in Table 2.2. The significant interaction between block and discrimination indicated that the linear effect across anagram durations differed for hits versus false alarms (see Figure 1). Hits decreased about .04 per block as anagram duration was halved, whereas false alarms increased about .04 per block (both linear effects were significant).

When AS JOSs and S JOSs were analysed separately, the same patterns occurred: a significant main effect of discrimination and an interaction with block. Each JOS was again discriminating at each duration ($ps < .001$). For both AS JOSs and S JOSs, the linear contrast analyses showed that linear effect of block was significant, as was the linear interaction between block and discrimination. For AS JOSs, the decrease in discrimination across blocks was due to a linear decrease in hits, whereas false alarms did not increase across blocks. For S JOSs, the reverse pattern was found: the decrease in discrimination across blocks was due to a linear increase in false alarms, whereas hits did not decrease across blocks. We discuss this novel pattern further in the General Discussion.

Solving Phase

We devised two measures to assess how well JOSs predict later problem-solving success. The solving rates for each JOS were similar across block, therefore we averaged across JOS phase blocks in our solving phase analyses (our Supplementary Materials provide the block-wise means).

Our first measure, *proportion solved*, was calculated as the number of anagrams solved during the solving phase that had received a given JOS during the JOS phase, divided by the total number of anagrams that had received that JOS during the JOS phase. For example, if a participant went on to solve 6 out of 10 anagrams to which they had made S JOSs, their proportion solved in the solving phase would be .6 for S JOSs. The proportion solved in the

Table 2.2*Experiment 1: JOS Phase Linear Contrast ANOVAs*

JOS(s)/Linear contrast	<i>F</i>	<i>MSE</i>	<i>p</i>	η^2_p	Contrast Coefficient
AS+S JOSs					
Block	0.22	0.014	.64	.00	-.002
Block × Discrimination	51.01	0.059	< .001	.34	
Hit Rate	33.73	0.025	< .001	.25	-.041
False Alarm Rate	21.06	0.031	< .001	.18	.036
AS JOSs					
Block	18.37	0.319	< .001	.17	-.023
Block × Discrimination	19.74	0.756	< .001	.18	
Hit Rate	26.39	0.998	< .001	.23	-.051
False Alarm Rate	1.05	0.015	.31	.01	-.006
S JOSs					
Block	4.74	0.159	.03	.06	.02
Block × Discrimination	9.34	0.936	.003	.10	
Hit Rate	0.08	0.007	.78	.00	-.007
False Alarm Rate	30.23	0.866	< .001	.23	.042

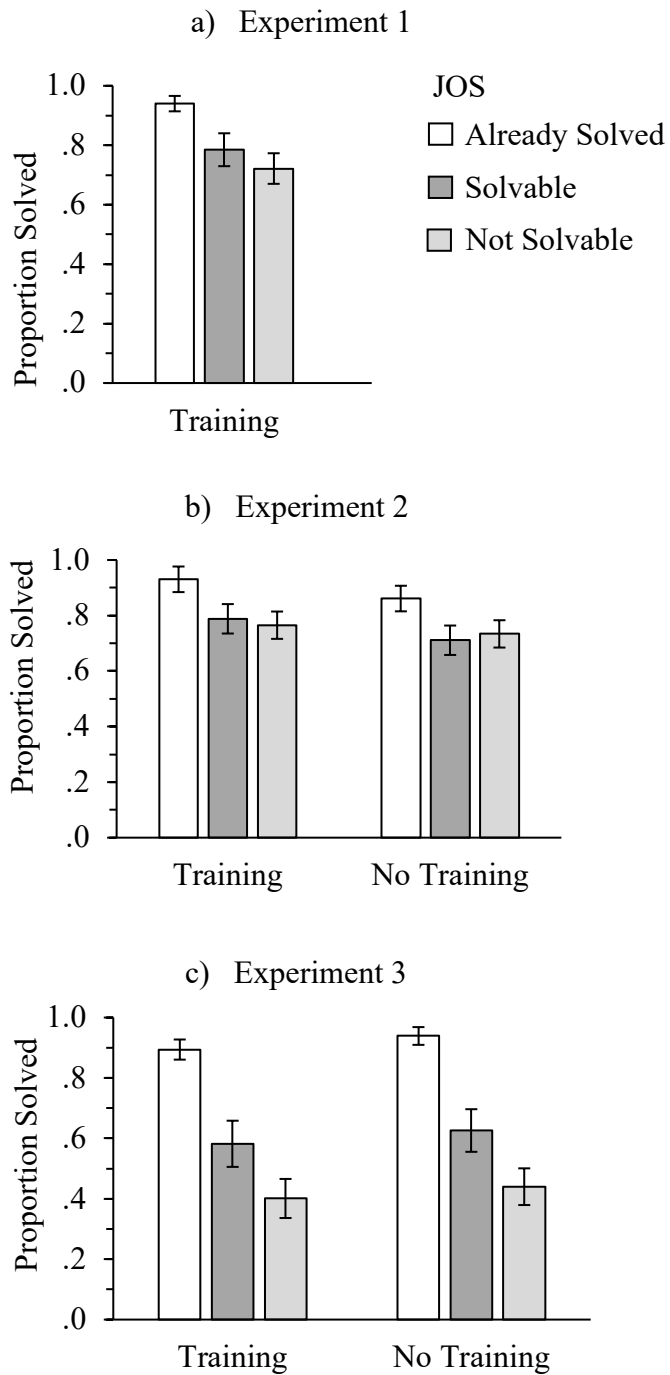
solving phase for each JOS were thus independent and each could range from 0 to 1. Our Supplementary Materials provide a full illustrative example.

The mean proportion solved as a function of JOS (AS vs. S vs. NS) was analysed using a repeated-measures ANOVA, which was significant, $F(2, 150) = 27.41$, $MSE = 1.01$, $p < .001$, $\eta^2_p = .27$ (see Figure 2.4a for the means). Pairwise multiple comparisons established that participants solved a greater proportion of anagrams that had received AS JOSs

compared to either S JOSs ($p < .001$) or NS JOSs ($p < .001$), confirming that AS JOSs were predictive of later solving. In sharp contrast, solving rates were not significantly higher for anagrams that had received S JOSs versus NS JOSs ($p = .67$), thus S JOSs were discriminating during the JOS phase but were not predictive of later solving.

Figure 2.4

Experiments 1-3: Mean Proportions of Solvable Anagrams Solved in the Solving Phase (Bars show 95% CI of each mean)



Our second measure for assessing how well JOSs predict later problem-solving success was *solved versus not solved outcomes*. This measure was calculated as the proportion of anagrams that were *solved* versus *not solved* in the solving phase that had received a given JOS. For each JOS, the proportion *solved* was calculated by dividing the total number of solved anagrams given that JOS by the total number of anagrams solved, and the proportion *not solved* was calculated by dividing the total number of not solved anagrams given that JOS by the total number of anagrams not solved. For example, if a participant solved 10 anagrams, and 6 of those anagrams had received S JOSs, the proportion solved in the solving phase for S JOSs would be .60. Likewise, if a participant failed to solve 10 anagrams, and 2 of those unsolved anagrams had received S JOSs, the proportion not solved in the solving phase for S JOSs would be .20. Thus, the two proportions were independent and could each range from 0 to 1 for each JOS, allowing us to compare the rates of anagrams solved versus the rates of anagrams not-solved directly (see Figure 2.5a for the means). Our Supplementary Materials provide a full illustrative example.

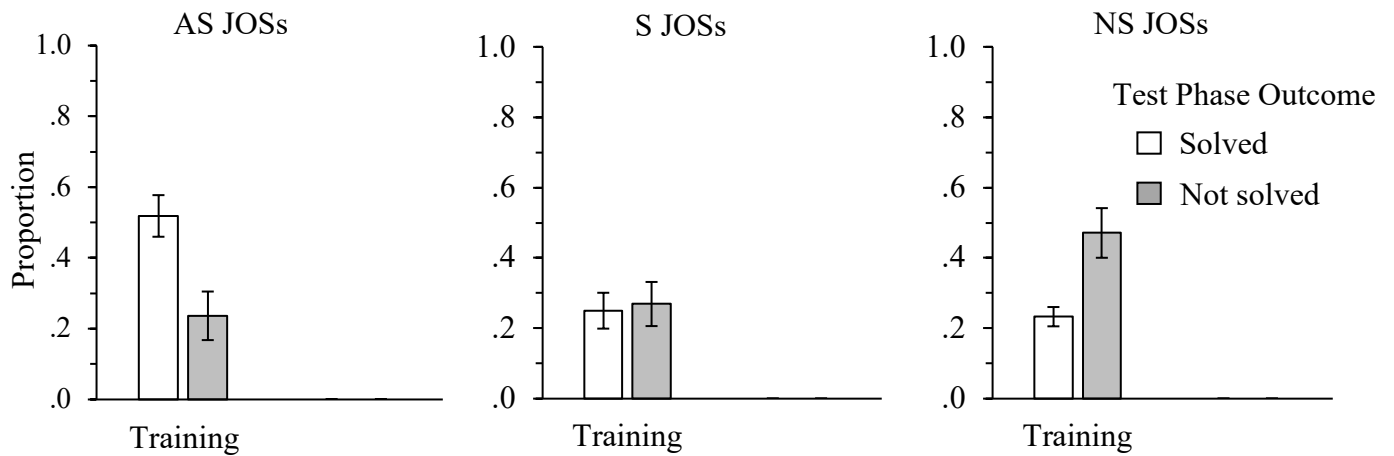
AS JOSs were more frequent among solved anagrams than among not-solved anagrams in the solving phase, $F(1, 93) = 53.70$, $MSE = 3.75$, $p < .001$, $\eta^2_p = .37$, whereas this was not the case for anagrams that received S JOSs, $F(1, 93) = 0.37$, $MSE = 0.02$, $p = .55$, $\eta^2_p = .004$. On the other hand, NS JOSs were more frequent among not-solved than among solved anagrams, $F(1, 93) = 56.40$, $MSE = 2.90$, $p < .001$, $\eta^2_p = .38$. In sum, AS and NS JOSs reliably predicted later solving outcomes, but S JOSs did not.

Figure 2.5

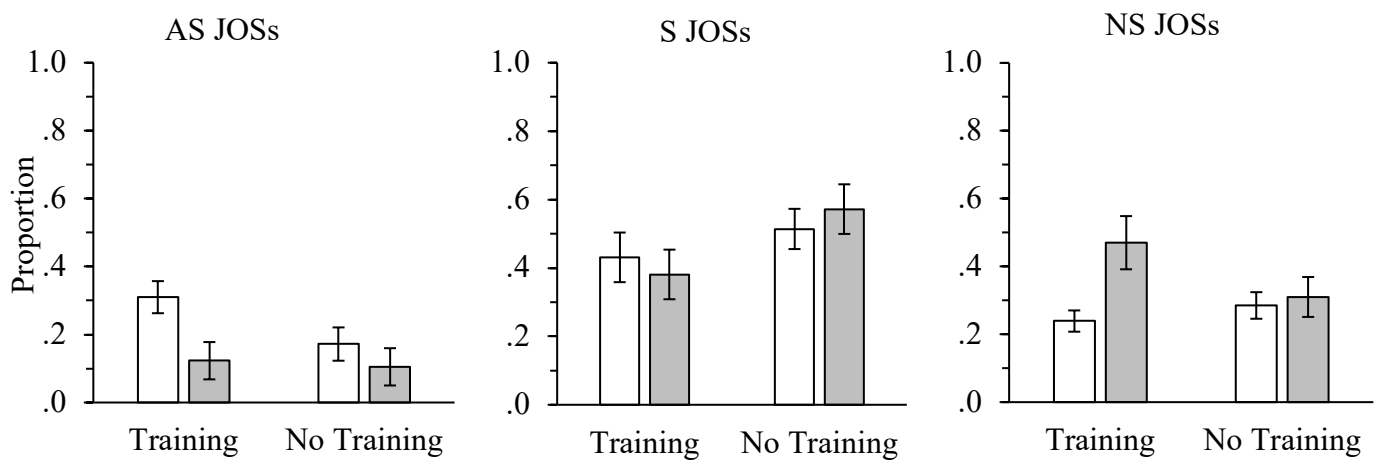
Experiments 1-3: Mean Proportion of Solved Versus Not Solved Outcomes for Solvable

Anagrams (Bars show 95% CI of each mean)

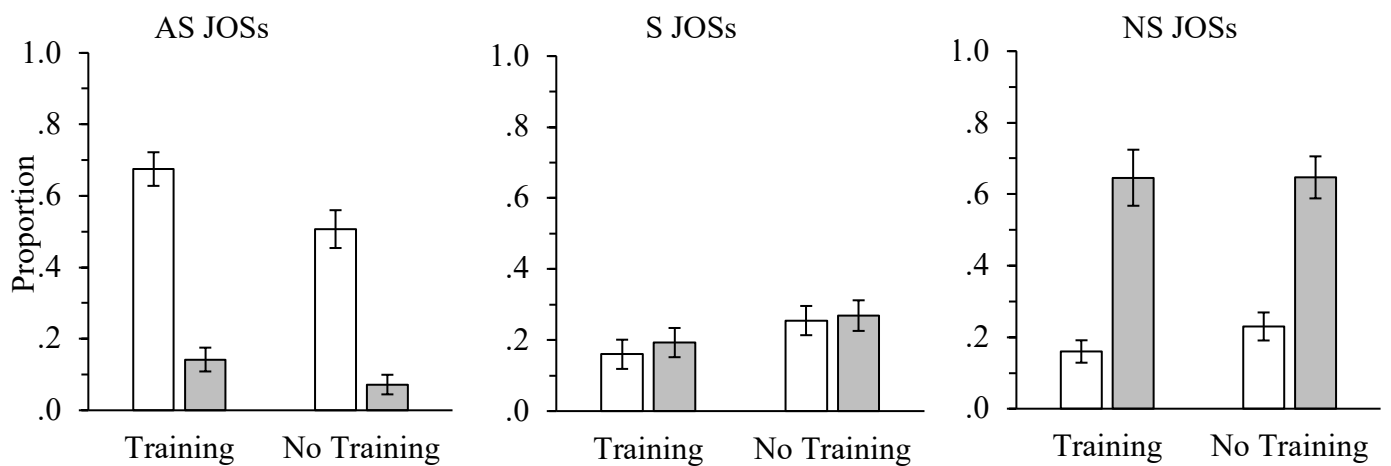
a) Experiment 1



b) Experiment 2



c) Experiment 3



Discussion

In the JOS phase, participants' JOSs accurately discriminated between solvable and unsolvable anagrams, even at our briefest anagram duration of 2 s. Importantly, this pattern held for S JOSs after excluding trials that led to AS JOSs. The S JOS pattern establishes that above-chance discrimination of JOSs can occur apart from trials in which participants have spontaneously solved problems during the JOS task. In contrast, Topolinski et al. (2016, Experiments 6 and 7) found that JOS discrimination was only marginally significant after excluding already-solved anagrams.

In the solving phase, AS JOSs were predictive of later solving success (and NS JOSs were predictive of later solving failure). Most surprisingly, we found that S JOSs were not predictive of later solving success. Participants solved more anagrams given AS JOSs than anagrams given either S or NS JOSs, but solving rates were not higher for anagrams given S JOSs rather than NS JOSs. In addition, solved outcomes were not more frequent than not-solved outcomes for anagrams that received S JOSs, unlike for AS JOSs. Thus, although S JOSs were discriminating, they were not associated with later problem-solving success.

Experiment 2

Although Experiment 1 provided new insights about JOS discrimination and predictiveness, its design did not allow us to gauge whether the 'training' we provided through the inclusion of longer duration anagrams increased S JOS discrimination. Therefore, in Experiment 2 we manipulated the presence versus absence of longer-duration anagrams across groups. The *training group* was identical to Experiment 1 and in the *no-training group* we presented anagrams for 2 s in all 4 blocks. This design allowed us to test whether the training group showed greater JOS discrimination than the no-training group in blocks 1-3 after the same amount of task experience. It also enabled us to test whether training improved JOS discrimination in block 4. The solving phase was the same as Experiment 1, thus the

Experiment 2 design again allowed us to test whether training with longer-duration anagrams in the JOS phase influences later solving performance.

We expected the training group to show greater AS JOS and S JOS discrimination than the no-training group. Because longer duration anagrams provide problem-solving successes during the JOS task, participants may use these successes to better regulate their JOSs on trials where they are less likely to solve the anagram during its presentation, resulting in improved discrimination. In turn, training with longer-duration anagrams was expected to result in AS and S JOSs being more predictive of anagram solving during the solving phase.

Method

The experiment was pre-registered on OSF at <https://osf.io/cq2kb>.

Participants

We tested another 238 MTurk workers, as per Experiment 1. Data for the training and no-training groups were collected in turn (back-to-back). We excluded 56 participants who met more than one pre-registered exclusion criteria. The final sample consisted of 182 participants (101 female, 81 male; mean age = 41.54, $SD = 13.39$), 91 per group, in line with our pre-registration.

Stimuli

The Experiment 1 stimuli were used.

Procedure

The Experiment 1 procedure was used, except the anagrams were presented for 2 s in each block in the no-training group; participants were informed of this duration. All 40 solvable anagrams were shown in the solving phase.

Results

JOS Phase

JOS discrimination was measured as in Experiment 1. The combined AS+S JOS results replicated Experiment 1 and are presented in our Supplementary Materials. The means for AS JOSs are provided in Figures 2.2b-2.2c, and for S JOSs in Figures 2.3b-2.3c. The measures were computed and analysed as per Experiment 1. Table 2.3 provides the ANOVA results, and Table 2.4 provides the linear contrast results.

Training Group. *In the training group*, the discrimination pattern for AS JOSs and S JOSs fully replicated Experiment 1. In both cases, JOSs distinguished between solvable and unsolvable anagrams, and discrimination decreased across blocks but was significant at each duration ($ps \leq .001$). The linear effect of block was significant for both AS JOSs and S JOSs, as was the linear interaction between block and discrimination. The decrease in discrimination across blocks was again due to a linear decrease in hits (rather than an increase in false alarms) for AS JOSs, and to a linear increase in false alarms (rather than a decrease in hits) for S JOSs.

No-Training Group. In the no-training group, JOS discrimination was significant for both AS JOSs and S JOSs. However, unlike in the training group, here the discrimination by block interactions were not significant. Discrimination was significant in each block for each measure, except in block 1 for S JOSs ($p = .09$).

Table 2.3*Experiment 2: JOS Phase ANOVAs Results*

JOS(s)/Effect	Training					No-training				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs										
Discrimination	1, 70	23.70	111.64	< .001	.62	1, 65	7.81	62.76	< .001	.49
Block	2.75, 192.70*	0.22	10.81	< .001	.13	2.45, 159.54*	0.03	1.15	.33	.02
Discrimination \times Block	3, 210	0.20	10.31	< .001	.13	3, 195	0.02	1.24	.30	.02
S JOSs										
Discrimination	1, 76	5.68	44.79	< .001	.37	1, 88	0.58	10.77	.001	.11
Block	3, 228	0.11	3.01	.03	.04	2.82, 248.46*	0.04	1.16	.33	.01
Discrimination \times Block	3, 228	0.12	3.83	.01	.05	3, 264	0.01	0.31	.82	.003

Note. * Huynh-Feldt correction was applied because assumption of sphericity was violated.

Table 2.4*Experiment 2 and 3: JOS Phase Linear Contrast ANOVAs*

JOS(s)/Linear contrast	Experiment 2					Experiment 3				
	<i>F</i>	<i>MSE</i>	<i>p</i>	η^2_p	Contrast Coefficient	<i>F</i>	<i>MSE</i>	<i>p</i>	η^2_p	Contrast Coefficient
AS JOSs										
Block	21.27	0.258	< .001	.23	-.021	122.40	1.274	< .001	.46	-.039
Block \times Discrimination	25.86	1.140	< .001	.27		85.61	3.164	< .001	.37	
Hit Rate	31.09	1.071	< .001	.31	-.054	115.44	3.93	< .001	.44	-.075
False Alarm Rate	0.00	0.000	.97	.00	.001	9.57	0.06	.002	.06	-.008
S JOSs										
Block	7.05	0.164	.01	.09	.025	29.17	0.786	< .001	.19	.034
Block \times Discrimination	8.91	0.635	.004	.11		11.14	0.835	.001	.08	
Hit Rate	0.001	0.00	.98	.00	-.007	2.86	0.185	.09	.02	.006
False Alarm Rate	23.13	0.852	< .001	.20	.043	86.05	2.279	< .001	.37	.055

Did Longer-Duration Anagrams Improve JOS Discrimination in Blocks 1-3? We next gauged whether training enhanced JOS discrimination in blocks 1-3 relative to the no-training group, using a 2(discrimination: hits vs. false alarms) \times 3(block: 1-3) by 2(group: training vs. no-training) mixed-factor ANOVA for each JOS measure. The complete ANOVA results are reported in Table 2.5. Of central interest was the three-way interaction. For AS JOSs, this interaction was significant, indicating that longer-duration anagrams in the training group improved AS JOS discrimination. However, the three-way interaction was not significant for S JOSs – the longer-duration anagrams in blocks 1-3 (as opposed to 2 s) did not result in more discriminating S JOSs in these blocks. Regardless, the Discrimination \times Group interaction suggests that S JOSs were more discriminating overall with longer-duration anagrams.

Table 2.5*Experiment 2: JOS Phase Discrimination ANOVAs in Blocks 1-3*

JOS(s)/Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs					
Discrimination × Group × Block*	1,86, 265.06	0.09	5.41	.01	.04
Discrimination	1, 140	23.95	166.85	< .001	.54
Discrimination × Group	1, 140	1.88	13.08	< .001	.09
Block*	1.91, 272.94	0.02	0.84	.43	.01
Block × Group*	1.91, 272.94	0.04	1.68	.19	.01
Block × Discrimination*	1.86, 265.06	0.03	1.51	.22	.01
Group	1, 140	2.58	13.33	< .001	.09
S JOSs					
Discrimination × Group × Block	2, 328	0.04	1.48	.23	.01
Discrimination	1, 164	4.64	63.06	< .001	.28
Discrimination × Group	1, 164	1.59	21.56	< .001	.12
Block*	1.87, 307.26	0.02	0.45	.64	.003
Block × Group*	1.87, 307.26	0.08	2.29	.10	.01
Block × Discrimination	2, 328	0.04	0.16	.85	.001
Group	1, 164	2.07	4.82	.03	.03

Note. * Huynh-Feldt correction was applied because assumption of sphericity was violated.

Did Training Improve JOS Discrimination in Block 4? We next focused on block 4 to determine whether training improved JOS discrimination where both groups received 2 s duration anagrams. For each JOS measure, we ran a 2(discrimination: hits, false alarms) \times 2(group: training vs. no-training) mixed-factor ANOVA (see Table 2.6). The effect of interest was the interaction, which was not significant either for AS or S JOSs. Thus, training with longer-duration anagrams, relative to 2 s anagrams, did not improve discrimination for either AS or S JOSs.

Solving Phase

Our solving phase analyses again averaged across the JOS phase blocks. Therefore, when we refer to the effect of training versus no-training on JOS predictiveness, we are referring to the general effect of experience with longer-duration anagrams on solving outcomes. The solving phase analyses followed Experiment 1, except group was added as a between-subjects factor. The means for the *proportion solved* measure appear in Figure 4b. The 3(JOS: AS, S, NS) by 2(group: training vs. no-training) ANOVA revealed a significant main effect of JOS, $F(2, 192) = 13.78$, $MSE = 0.45$, $p < .001$, $\eta^2_p = .13$. The proportion of anagrams solved was greater for anagrams that had received AS JOSs in the JOS phase rather than either S JOSs ($p = .001$) or NS JOSs ($p < .001$). In contrast, participants were equally likely to solve anagrams that received S JOSs or NS JOSs in the JOS phase ($p = 1.00$). Thus, replicating Experiment 1, S JOSs were not predictive of greater problem-solving success. The group main effect was not significant, $F(1, 96) = 2.00$, $MSE = 0.28$, $p = .16$, $\eta^2_p = .02$. Strikingly, training did not improve how well JOSs predicted later solving: JOS predictiveness was similar across groups, $F(2, 192) = 1.37$, $MSE = 0.04$, $p = .26$, $\eta^2_p = .01$ for the interaction.

The *solved versus not-solved outcome* measure means appear in Figure 2.5b, and Table 2.7 provides the 2(outcome: solved, not solved) by 2(group: training, no-training)

Table 2.6*Experiment 2 and 3: JOS Phase Discrimination ANOVAs in Block 4*

JOS(s)/Effect	Experiment 2					Experiment 3				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs										
Discrimination	1, 151	4.78	96.84	< .001	.39	1, 292	26.49	599.32	< .001	.67
Group	1, 151	0.03	0.49	.49	.003	1, 292	0.06	1.32	.25	.01
Discrimination \times Group	1, 151	0.07	1.38	.24	.01	1, 292	0.06	1.32	.25	.01
S JOSs										
Discrimination	1, 178	0.74	18.77	< .001	.10	1, 291	0.07	2.64	.13	.01
Group	1, 178	0.37	2.16	.14	.01	1, 291	0.002	0.02	.89	< .001
Discrimination \times Group	1, 178	0.04	1.11	.29	.01	1, 291	0.001	0.02	.90	< .001

ANOVAs. The main effects of group are not of interest because they average across outcomes. AS JOSs were more frequent among solved anagrams than among not-solved anagrams, and this effect was larger in the training group, resulting in a significant interaction (though the difference was significant in each group, $ps \leq .03$). S JOSs were not significantly more frequent among solved anagrams than among not-solved anagrams. Here, outcome interacted with group, but the outcome difference did not reach significance for either group ($ps \geq .05$). NS JOS were more frequent among not-solved anagrams than among solved anagrams, and outcome interacted with group; this effect was significant in the training group ($p < .001$) but not in the no-training group ($p = .42$). Thus, training with longer-duration anagrams in blocks 1-3 enhanced the predictiveness of AS and NS JOSs but not S JOSs.

Table 2.7*Experiments 2 and 3: Solved versus Not Solved Outcomes ANOVAs*

JOS(s)/Effect	Experiment 2					Experiment 3				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs										
Outcome	1, 157	1.28	34.72	< .001	.18	1, 274	27.81	579.93	< .001	.68
Group	1, 157	0.49	4.47	.04	.03	1, 274	1.70	24.77	< .001	.08
Outcome \times Group	1, 157	0.29	7.80	.01	.05	1, 274	0.24	5.04	.03	.02
S JOSs										
Outcome	1, 157	0.001	0.03	.87	< .001	1, 274	0.11	2.70	.10	.01
Group	1, 157	1.49	8.20	.01	.05	1, 274	0.93	11.74	< .001	.04
Outcome \times Group	1, 157	0.23	6.34	.01	.04	1, 274	0.003	0.07	.79	< .001
NS JOSs										
Outcome	1, 157	1.30	32.26	< .001	.17	1, 274	31.30	660.41	< .001	.71
Group	1, 157	0.26	3.20	.08	.02	1, 274	0.27	5.77	.02	.02
Outcome \times Group	1, 157	0.84	20.86	< .001	.12	1, 274	0.27	5.77	.02	.02

Discussion

Replicating Experiment 1, participants' AS JOSs and S JOSs both discriminated solvable from unsolvable anagrams. Experiment 2 extended this finding by establishing that both JOSs were discriminating even in a no-training group where anagram duration was always 2 s during the JOS phase. Experiment 2 also confirmed that presenting longer-duration anagrams in blocks 1-3 led to more discriminating AS JOSs. Importantly, however, this was not the case for S JOSs. Thus, inclusion of longer-duration anagrams increased the likelihood of spontaneous anagram solving, but it did not improve S JOS discrimination. In fact, training did not result in greater discrimination in block 4 (2 s anagrams for both groups) for either S or AS JOSs.

The solving phase for the training group replicated Experiment 1. Participants were more likely to solve anagrams that had been given AS JOSs than either S or NS JOSs. AS and NS JOSs predicted later anagram solving successes and failures, respectively. Importantly, replicating Experiment 1, S JOSs were not predictive of later problem-solving outcomes. Additionally, Experiment 2 showed that although training improved the predictiveness of AS and NS JOSs, it did not do so for S JOSs. Training also did not result in a greater proportion of anagrams solved, regardless of JOS.

In Experiment 1 and 2, the majority of anagrams were solved no matter the JOS (.71-.94; see Figure 2.4a and 2.4b). Participants were informed that each anagram was solvable, and were given 45 s to solve each one. These design elements may have increased solving efforts and, in turn, solving successes, which may have masked our ability to detect effects of training on JOS predictiveness. Experiment 3 revisited JOS predictiveness when efforts were made to reduce solving phase success, which also enabled us to examine how training during the JOS phase influences effort regulation in the solving phase.

Experiment 3

The JOS phase in Experiment 3 was identical to Experiment 2, allowing us to test the replicability of our findings with respect to JOS discrimination and the impact of training on JOS discrimination. However, the solving phase was modified to allow us to examine the generality of our findings regarding JOS predictiveness and to investigate the effects of training on how well JOSs predict effort regulation. In Experiment 3, the solving phase included 5 solvable and 5 unsolvable anagrams from each block of the JOS phase (rather than including only the 10 solvable anagrams from each block). In addition, we allowed participants to self-regulate their problem-solving effort: they could spend as much or as little time as they wished attempting to solve each anagram. On each trial, they either typed in the anagram solution, passed, or indicated that the anagram was not solvable (dubbed a *not-solvable response*). The inclusion of unsolvable anagrams, the ability to pass and make not-solvable responses, and to respond sooner than 45 s if no solution was found were expected to reduce the solving rate relative to Experiments 1 and 2. By lowering the solving rate, Experiment 3 was expected to provide a stronger test of JOS predictiveness, and of the potential effects of training on JOS predictiveness.

These modifications to the two-phase paradigm also provided new measures of the link between JOSs and later problem solving. One new measure was how long participants took to solve solvable anagrams, and another was how long they took to make not-solvable responses to unsolvable anagrams. The latter provides a novel measure of effort regulation that allowed us to examine, for example, whether participants spent longer solving anagrams when they had given an AS or S JOS versus an NS JOS, and whether training further impacted their effort regulation. We were also able to examine whether NS JOSs were associated with faster not-solvable responses for unsolvable anagrams, and whether training strengthened this

effect. A third new measure was the rate of not-solvable responses itself, which provided a parallel window onto these same questions.

Method

The experiment was pre-registered on OSF at <https://osf.io/bzuqc>.

Participants

A total of $N = 357$ additional MTurk workers were tested. Allocation to the training or no-training groups was randomized. We increased the sample size for each group by 50 given that the solving phase now included 5 rather than all 10 solvable anagrams from each block of the JOS phase. Here, 60 participants were excluded for failing more than one pre-registered exclusion criterion. The final sample consisted of 297 participants (221 female, 76 male, mean age = 42.25, $SD = 12.90$): 150 in the training group, and 147 in the no training group, in line with our pre-registration.

Stimuli

The Experiment 1 and 2 stimuli were used.

Procedure

The procedure followed Experiment 2, except the modifications to the solving phase to enable us to measure regulation of problem-solving effort. The solving phase now consisted of one of two sets of 20 solvable and 20 unsolvable anagrams from the JOS phase. To this end, a random half of the anagrams from each block were assigned to each set, and the set used in the solving phase was counterbalanced across participants. The solving phase instructions informed participants that half the anagrams were solvable, and half were not. They were told that they had as much time to try to solve each anagram as they wished, and they were instructed to either type in a solution, type the letter “P” to pass if they believed the anagram was solvable but were unable to solve it, or to type the letter “N” for “not solvable”

if they believed the anagram was unsolvable. Participants were given 5 JOS practice trials, and 5 solving practice trials using the same anagrams from the JOS practice trials.

Results

Experiment 3 was analysed as per Experiment 2, with additional analyses of unsolvable anagrams in the solving phase, and of self-regulated solving times.

JOS Phase

JOS discrimination means are provided in Figures 2.2d-2.2e for AS JOSs and in Figures 2.3d-2.3e for S JOSs (see Table 2.8 for ANOVA results and Table 2.4 for the linear contrasts).

Training Group. The discrimination pattern for AS JOSs and S JOSs replicated Experiments 1 and 2. In each case, AS and S JOSs were both discriminating, and discrimination decreased across blocks but was significant at each duration ($ps \leq .001$). For both AS JOSs and S JOSs, linear contrast analyses showed that the linear effect of block was significant, but only AS JOSs had a significant linear interaction between block and discrimination. For AS JOSs, although hits and false alarms both decreased linearly, the decrease in discrimination across blocks was greater for hits than for false alarms (but both were significant). For S JOSs, although the interaction was not significant, the analyses showed a similar pattern to Experiments 1 and 2: the change in discrimination across blocks was driven by a significant linear increase in false alarms ($p < .001$), whereas the decrease in hits was not significant ($p = .09$).

No-Training Group. The discrimination pattern in the no-training group also largely replicated Experiment 2. Discrimination was significant across all JOSs, did not significantly interact with block, and was significant in each block for each JOS, except for block 4 for S JOSs ($p = .42$).

Table 2.8*Experiment 3: JOS Phase ANOVAs Results*

Effect	Training					No-training				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs										
Discrimination	1, 145	90.63	723.57	< .001	.83	1, 143	47.64	333.00	< .001	.70
Block	2.90, 421.03*	0.90	50.50	< .001	.26	3, 429	0.03	1.61	.19	.01
Discrimination × Block	3, 435	0.53	33.73	< .001	0.20	3, 429	0.01	0.71	.55	.01
S JOSs										
Discrimination	1, 123	2.45	33.59	< .001	.22	1, 138	0.94	13.28	< .001	.09
Block	2.87, 352.49*	0.53	13.45	< .001	.10	2.86, 394.45*	0.18	4.64	.004	.03
Discrimination × Block	3, 369	0.14	4.30	.01	.03	3, 414	0.05	1.95	.12	.01

Note. * Huynh-Feldt correction was applied because assumption of sphericity was violated.

Did Longer-Duration Anagrams Improve JOS Discrimination in Blocks 1-3? The pattern of three-way interactions between discrimination, block, and group across JOSs replicated Experiment 2 (see Table 2.9). Although AS JOS discrimination decreased across blocks in the training group, longer-duration anagrams still led to significantly greater discrimination across blocks 1-3 for AS JOSs. In the no-training group, discrimination did not increase across blocks, and was significantly weaker than in the training group. The three-way interaction was not significant for S JOSs, and discrimination did not interact with group – longer duration anagrams did not sponsor greater S JOS discrimination.

Solving Phase

Solving phase analyses followed Experiment 2, with additional analyses for the self-regulated elements. Note that “pass” responses were too rare to analyse separately. As expected, the change to a self-regulated solving phase reduced the mean solving rate in Experiment 3 ($M = .65$, $SD = .20$) relative to Experiments 1 and 2 ($M = .77$, $SD = .23$), $t(775) = 7.09$, $p < .001$. This reduction of solving rate should make it easier to detect an impact of JOS on solving phase outcomes.

Proportion Solved. The *proportion solved* means for solvable anagrams appear in Figure 2.4c. The $3(\text{JOS: AS, S, NS}) \times 2(\text{group: training, no training})$ ANOVA revealed a significant main effect of JOS, $F(2, 400) = 168.01$, $MSE = 12.70$, $p < .001$, $\eta^2_p = .46$. The proportion of anagrams solved was greater for anagrams that received AS JOSs compared to either S JOSs ($p < .001$) or NS JOSs ($p < .001$). Unlike in Experiments 1 and 2, here the proportion of anagrams solved was greater for anagrams receiving S JOSs than NS JOSs ($p < .001$). As in Experiment 2, the main effect of group was not significant, $F(1, 207) = 0.51$, $MSE = 0.07$, $p = .48$, $\eta^2_p = .002$, nor was the interaction, $F(1, 400) = 0.01$, $MSE = 0.001$, $p = .99$, $\eta^2_p < .001$. Thus, training did not increase the overall proportion of anagrams solved in the solving phase, nor was the predictiveness of JOSs greater in the training group.

Table 2.9*Experiment 3: JOS Phase Discrimination ANOVAs in Blocks 1-3*

JOS(s)/Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs					
Discrimination × Group × Block*	2, 582	0.19	12.20	< .001	.04
Discrimination	1, 291	110.12	1048.40	< .001	.78
Discrimination × Group	1, 291	3.88	36.92	< .001	.11
Block*	1.96, 569.27	0.15	8.05	< .001	.03
Block × Group*	1.96, 569.27	0.43	24.02	< .001	.08
Block × Discrimination*	2, 582	0.08	4.75	.01	.02
Group	1, 291	4.60	34.02	< .001	.11
S JOSs					
Discrimination × Group × Block	1.96, 514.28	0.07	2.27	.10	.01
Discrimination	1, 262	3.47	50.61	< .001	.16
Discrimination × Group	1, 262	0.21	3.02	.09	.01
Block*	1.91, 500.08	0.08	2.01	.14	.01
Block × Group*	1.91, 500.08	0.30	7.94	< .001	.03
Block × Discrimination*	1.96, 514.28	0.04	1.37	.25	.01
Group	1, 262	4.16	18.30	< .001	.07

Note. * Huynh-Feldt correction was applied because assumption of sphericity was violated.

Did Training Improve JOS Discrimination in Block 4? Training with longer-duration anagrams did not significantly improve discrimination for either AS and S JOSs (see Table 2.6), replicating Experiment 2.

Solved vs. Not Solved Outcomes for Solvable Anagrams. The *solved versus not solved outcome* means appear in Figure 2.5c, and the $2(\text{outcome: solved vs. not solved}) \times 2(\text{group: training vs. no training})$ ANOVA results appear in Table 2.7. Incorrect solutions, “pass” responses, and not-solvable responses were all counted as not-solved outcomes. AS JOSs were more frequent among solved anagrams than among not-solved anagrams, and training interacted with outcome such that this effect was larger in the training group (though each was significant, $ps < .001$). In contrast, S JOSs did not significantly predict solving phase outcome, and the interaction with training was also not significant. NS JOSs were more frequent among not-solved anagrams than among solved anagrams, and here the interaction with outcome and training was (just) significant (and the effect was significant for each group, $ps < .001$). Thus, predictiveness of NS JOSs was again enhanced by training.

Not-Solvable Responses to Solvable vs. Unsolvable Anagrams. The *proportion of not-solvable responses in the solving phase* was calculated as the number of not-solvable responses in the solving phase that had received a given JOS during the JOS phase, divided by the total number of anagrams that had received that JOS during the JOS phase. For example, if a participant gave a not-solvable response in the solving phase for 5 out of 10 anagrams to which they had given an NS JOS in the JOS phase, their proportion of not-solvable responses in the solving phase for NS JOSs was .5. Their independence allowed for direct comparisons of the mean proportion of anagrams given a not-solvable response in the solving phase as a function of JOS. Due to the rarity of AS and S JOSs for unsolvable anagrams, these proportions were pooled across AS+S JOSs.

A $2(\text{JOS: AS+S, NS}) \times 2(\text{anagram type: solvable, unsolvable}) \times 2(\text{group: training, no-training})$ mixed-factor ANOVA was conducted on not-solvable responses (Table 2.10). The means appear in Figure 2.6. Here, we were interested in whether not-solvable responses were more likely for anagrams given NS JOSs than AS+S JOSs, whether anagram type

strengthened the likelihood of not-solvable responses for NS JOSs relative to AS JOSs, and whether training moderated the latter interaction. There was a significant main effect of JOS; not-solvable responses were more likely for anagrams given NS JOSs than AS+S JOSs. JOS interacted with anagram type; the difference in proportion of not-solvable responses between AS+S and NS JOSs was greater for solvable anagrams (though was significant for both solvable and unsolvable anagrams; $ps < .001$). JOS also interacted with group; although both groups showed more not-solvable responses at test for anagrams given NS JOSs than for AS+S JOSs ($ps < .001$), this pattern was more robust in the training group.

Figure 2.6

Experiment 3: Mean Proportions of Not-Solvable Outcomes to Solvable Anagrams (Bars show 95% CI of each mean)

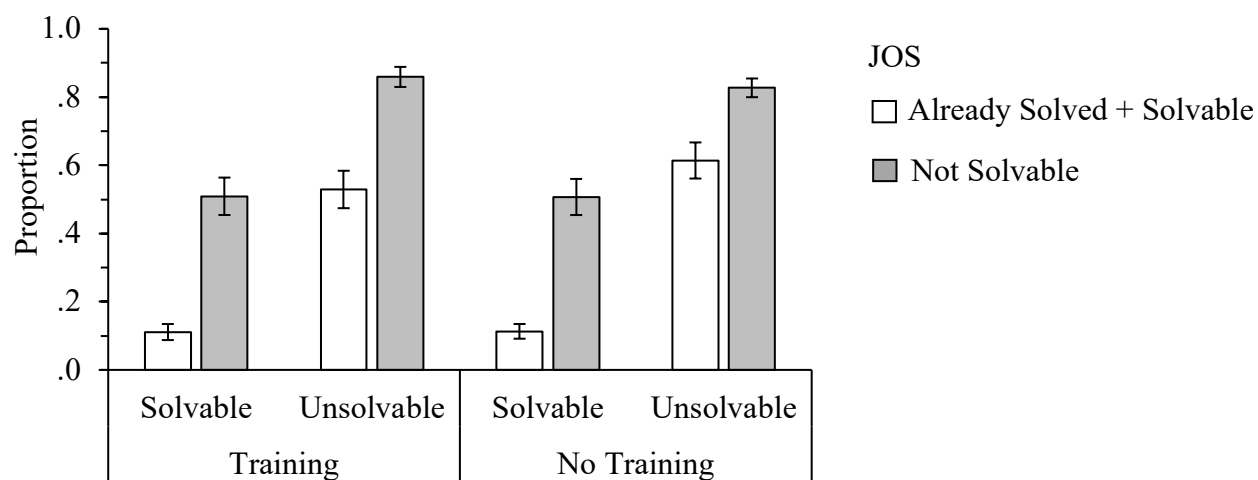


Table 2.10*Experiment 3: Proportion of Not-Solvable Responses ANOVA Results*

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	1, 266	29.82	739.65	< .001	.74
Anagram type	1, 266	42.31	625.07	< .001	.70
Group	1, 266	0.05	0.51	.48	.002
JOS × Anagram type	1, 266	1.03	21.38	< .001	.07
JOS × Group	1, 266	0.25	5.15	.02	.02
Group × Anagram type	1, 266	0.05	.82	.37	.003
JOS × Group × Anagram type	1, 266	0.21	4.37	.04	.02

Finally, the three-way interaction with anagram type was also significant. This interaction was followed up with separate interaction contrasts for the training and no-training groups (see Table 2.11). For each group, not-solvable responses were more likely to be provided for anagrams given NS JOSs than AS+S JOSs (i.e., a main effect of JOS). The interaction of JOS and anagram type was significant only in the no-training group, and reflected a smaller difference between JOSs for unsolvable than for solvable anagrams (though both were significant; $ps < .001$).

Table 2.11*Experiment 3: Interaction Contrasts for Proportion of Not-Solvable Responses ANOVA Results*

Effect	Training					No-Training				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	1, 128	17.10	396.07	< .001	.76	1, 138	12.80	246.70	< .001	.64
Anagram type	1, 128	19.06	394.05	< .001	.78	1, 138	23.47	358.78	< .001	.72
JOS \times Anagram type	1, 128	0.15	2.76	.10	.02	1, 138	1.13	26.34	< .001	.16

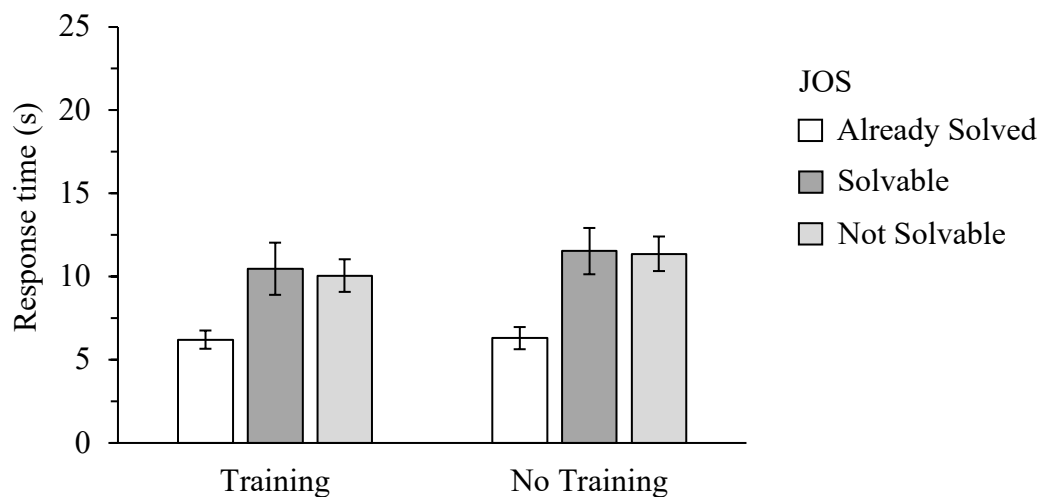
Do JOSs Predict Self-Regulated Response Times in the Solving phase?

Whether JOSs predict solving response times was analysed as per the not-solvable response analyses. Because solution times were negatively skewed, a base 10 logarithm transformation was applied to normally distribute the data. Thus, descriptive statistics are presented in seconds, but inferential statistics used the transformed means.

Solved Anagrams. The mean response time to correctly solve the solvable anagrams appear in Figure 2.7. The 3(JOS) \times 2(group) ANOVA revealed a main effect of JOS, $F(2, 280) = 81.15$, $MSE = 1.80$, $p < .001$, $\eta^2_p = .37$. AS JOSs were associated with shorter solution times than both S JOSs ($p < .001$) and NS JOSs ($p < .001$). In contrast, solution times were similar for anagrams that received S JOSs versus NS JOSs ($p = 1.00$). The main effect of group was not significant, $F(1, 140) = 2.84$, $MSE = 0.28$, $p = .09$, $\eta^2_p = .02$. The JOS \times group interaction was just shy of significance, $F(2, 280) = 2.96$, $MSE = 0.07$, $p = .053$, $\eta^2_p = .02$.

Figure 2.7

Experiment 3: Mean Solving Times for Anagrams (Bars show 95% CI of each mean)



Not-Solvable Responses to Solvable vs. Unsolvable Anagrams. The mean response time for making not-solvable responses appear in Figure 2.8. Table 2.12 shows the complete 2(JOS: AS+S, NS) \times 2(anagram type: solvable, unsolvable) \times 2(group: training, no-training)

mixed-factor ANOVA results. All three main effects were significant: not-solvable responses were faster for anagrams assigned NS JOSs than AS+S JOSs (JOS main effect), not-solvable responses were faster for solvable than unsolvable anagrams (anagram type main effect), and not-solvable response times were faster in the training than no-training group (group main effect). The interaction of JOS and anagram type was significant; the not-solvable response time difference between JOSs was larger for unsolvable than solvable anagrams (but both were significant; $ps \leq .003$). The interaction between JOS and group was significant; the not-solvable response time difference was larger for the no-training group than for the training group (both $ps < .001$). Thus, training reduced the difference in not-solvable response times for NS versus AS+S JOSs. The remaining effects were not significant.

Figure 2.8

Experiment 3: Mean Response Time for Not-Solvable Responses (Bars show 95% CI)

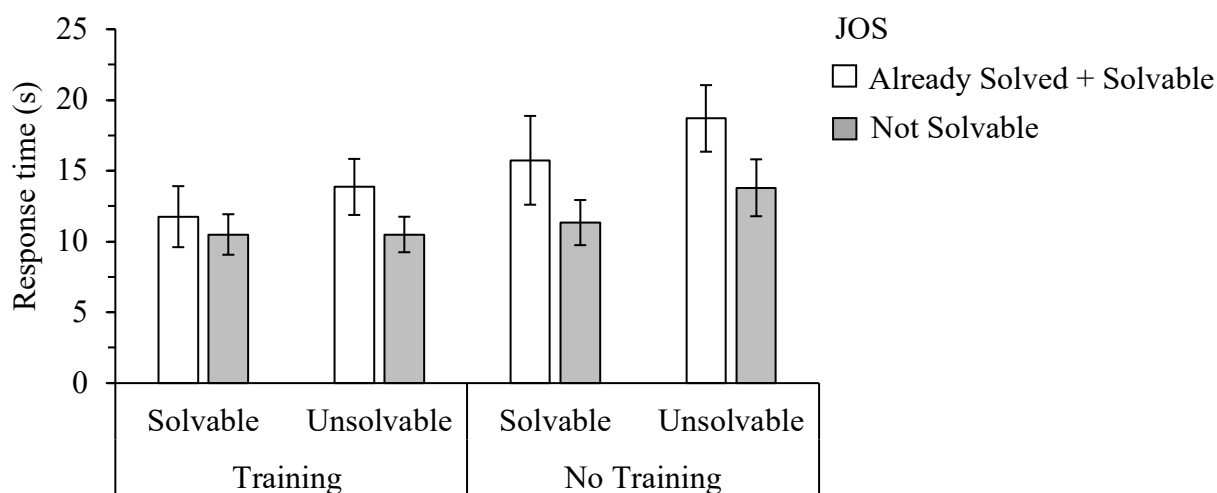


Table 2.12*Experiment 3: Mean Response Time for Not-Solvable Responses ANOVA Results*

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	1, 142	0.82	68.62	< .001	.33
Anagram type	1, 142	0.13	6.59	.01	.04
Group	1, 142	2.38	7.13	.01	.05
JOS × Anagram type	1, 142	0.12	8.31	.01	.06
JOS × Group	1, 142	0.05	4.41	.04	.03
Group × Anagram type	1, 142	0.04	1.90	.17	.01
JOS × Group × Anagram type	1, 142	0.02	1.66	.20	.01

Discussion

In terms of JOS discrimination, Experiment 3 replicated Experiment 2; AS and S JOSs were both discriminating, even in the no-training group where anagram duration was 2 s in all blocks. Longer-duration anagrams in blocks 1-3 increased the likelihood of AS JOSs rather than enhancing the discrimination of S JOSs, and training again did not increase discrimination in the final 2 s block for either S or AS JOSs.

Allowing participants to self-regulate their solving efforts (and the inclusion of unsolvable anagrams) reduced solving rates in Experiment 3, and thus provided a stronger test of whether JOSs (especially S JOSs) predict problem solving. AS JOSs and NS JOSs again predicted solving successes and failures, respectively. But even when self-regulation was permitted, we did not find any evidence that S JOSs predicted later problem-solving outcomes. As was found in Experiment 2, training improved the predictiveness of AS JOSs and NS JOSs, but not of S JOSs. Interestingly, in Experiment 3 solving rates for S JOSs were higher than for NS JOSs, unlike in Experiments 1 and 2. However, training still did not impact solving phase performance.

The design of Experiment 3 also allowed us to assess how often each JOS was associated with not-solvable responses in the solving phase. Solvable anagrams that received NS JOSs in the JOS phase received more not-solvable responses than anagrams that received AS+S JOSs, and this difference was greater in the training group. For unsolvable anagrams, the training group made more not-solvable responses for anagrams given NS JOSs than AS+S JOSs compared to the no-training group; the no-training group made more not-solvable responses for NS JOSs than AS+S JOSs, but the difference was smaller than the training group. Adding to Lauterman and Ackerman's (2019) findings that a 'not solvable' initial JOS predicts a 'not solvable' final JOS, longer-duration anagrams enhanced NS JOS predictiveness.

Experiment 3 also measured self-regulated response times during the solving phase. Unsurprisingly, anagrams that were reported to have been spontaneously solved during the JOS phase (AS JOSs) yielded the fastest solution times. However, solution times were similar for S JOSs and NS JOSs, and training with longer-duration anagrams did not impact this pattern. We also found faster not-solvable responses in the solving phase following NS JOSs than AS+S JOSs, particularly for unsolvable anagrams and for the no-training group. Lauterman and Ackerman (2019) reported that effort regulation following an S JOS was similar for solvable and unsolvable problems, suggesting that making an S JOS for an unsolvable problem may lead solvers to persevere on unsolvable problems. Our findings support theirs, and further establish that differences in not-solvable responses times between AS+S JOSs and NS JOSs are reduced via training with longer-duration anagrams.

In sum, Experiment 3 replicated Experiments 1 and 2 in terms of JOS discrimination and the ability of JOSs to predict solved versus not-solved outcomes. Adapting the two-phase paradigm to allow effort regulation and solving performance to vary extended our understanding of JOSs by revealing that S JOSs can be associated with a higher solving rate

relative to NS JOSs, whereas NS JOSs showed lower rates and faster ‘not solvable’ responses at test. Further, we found novel evidence that the ability of JOSs to predict the rate and speed of ‘not solvable’ responses was influenced by training, such that training lead to quicker and higher rates of not-solvable responses for anagrams given NS JOSs.

General Discussion

Three experiments provided an in-depth investigation of the first stage of meta-reasoning—judgements of whether problems are solvable or not. In our two-phase paradigm, participants first made JOSs to solvable and unsolvable anagram problems, and this JOS phase was followed by a solving phase. During the JOS phase, an ‘already solved’ (AS) JOS option was provided to allow participants to indicate having solved an anagram at this stage. A two-phase paradigm allows participants to focus on making intuitive JOSs in the JOS phase (at least at briefer anagram durations) and to focus on trying to solve the anagrams in the solving phase. Providing an AS JOS option allowed us to parse out solved anagrams from our discrimination measure in the JOS phase. Because JOSs are meant to be intuitive judgements (and intuition about problem solvability should precede solving; Ackerman & Thompson; 2017), it is important to separate intuitive JOSs from problems solved during the JOS process. This separation enabled us to more cleanly measure whether S JOSs predict solving outcomes and effort-regulation. We also examined the effects of training on JOS discrimination and predictiveness, by presenting anagrams for longer durations at first (16 s), which then halved across blocks (8 s, 4 s, 2 s). In Experiments 2 and 3 we compared JOSs in these training groups to no-training groups in which anagram duration was always 2 s. Below, we discuss in turn JOS discrimination and whether it was improved by training. We then discuss whether JOSs were predictive of later problem-solving outcomes and effort regulation, and whether these outcomes benefitted from training. Finally, we also discuss the

potential value of future research comparing two-phase and interleaved paradigms for capturing the initial stages of meta-reasoning.

JOS Discrimination

Our experiments provide evidence that participants' intuitions can discriminate solvable from unsolvable anagrams. AS and S JOSs were both found to be discriminating, even at our briefest anagram duration (2 s). Importantly, discrimination remained above chance when we excluded the anagrams that participants reported having solved during the task (i.e., those receiving AS JOSs).

Previous studies have reached different conclusions regarding the ability of solvable JOSs to discriminate solvable from unsolvable problems. Studies in which solvable JOSs were found to be discriminating did not allow participants to report having solved the problems during the JOS process (Balas et al., 2011; Bolte & Goschke, 2005; Novick & Sherman, 2003; Topolinski & Strack, 2009; Undorf & Zander, 2017). When already-solved items were reported and removed from analysis, Topolinski et al. (2016, Experiment 7) found that participants were only marginally sensitive to anagram solvability. Our experiments found that S JOS discrimination was significant, though it was notably weaker than AS JOS discrimination. Our η^2_p effect sizes for S JOS discrimination ranged from .09 to .11 in our no-training groups to .22 to .37 in our training groups, whereas Topolinski et al.'s was .06. Topolinski et al. also had a smaller sample and fewer JOS trials, thus their study may have lacked power. Regardless, our study is the first to provide clear evidence that S JOSs can be discriminating, even after excluding solutions arising during the JOS task.

An interesting question our study cannot address is the stimulus features participants use to successfully distinguish solvable from unsolvable anagrams. Perhaps participants' intuitions were biased by certain diagnostic letter combinations or differences in bigram frequencies. Another possible mechanism underlying S JOS discrimination is the unconscious

activation of semantic representations (Bowers et al., 1990) that would indicate an anagram is solvable. Future research should investigate the stimulus features that drive S JOS discrimination.

Even at our briefest anagram duration (2 s), AS JOSs were reported for 22-48% of the solvable anagrams during the JOS phase. Given that intuition about solvability precedes solving, we would expect a ‘solvable’ JOS to have preceded the solution on these trials (Ackerman & Thompson, 2017; Bowers et al., 1990). Thus, a potential disadvantage of removing AS JOS trials is that it may remove trials where accurate intuitions have occurred. In doing so, S JOS discrimination may be underestimated (and in turn may underestimate how well S JOSs predict solving outcomes, as discussed below). An alternative approach to capturing intuitive S JOSs is to establish a brief problem duration for each participant at which they no longer report AS JOSs. However, our concern with this approach is that the use of very brief problem durations may lead participants to rely on irrelevant cues to make their JOSs, or to simply engage in random responding. Thus, there are pros and cons to both approaches, and future research should compare them.

Turning to the impact of training, in Experiment 2, we found that AS JOS discrimination was more accurate in blocks 1-3 in our training group (who received longer-duration anagrams in these blocks) than in our no-training group. We expected that the training group would use their greater solving success as feedback to help calibrate their ‘solvable’ intuitions. However, training did not improve AS JOS or S JOS discrimination in the final block relative to the no-training group. Perhaps, then, the training group simply shifted their efforts to solving the anagrams when longer-duration anagrams were provided, rather than on trying to improve the accuracy of their intuitive judgements. If so, then our ‘training’ may not have helped participants learn to regulate their meta-reasoning during blocks 1-3. We recommend that future research consider alternative means of enhancing JOS

discrimination. For example, it could be worthwhile to examine the effects of providing explicit feedback about JOS accuracy (i.e., by indicating after each JOS whether the anagram was solvable or unsolvable), particularly given that prior studies have shown that trial-by-trial feedback improves discrimination in meta-memory tasks (e.g., Arnold et al., 2013; Higham, 2007; Sharp et al., 1988).

We consistently found that JOS discrimination in the training group weakened across blocks in a linear manner. Interestingly, this decrease in discrimination across blocks took a different form for AS and S JOSs. For AS JOSs, it reflected a linear decrease in hits across blocks (while false alarms remained similar), whereas for S JOSs, it reflected a linear increase in false alarms (while hits remained similar). AS JOSs and S JOSs appear to be impacted differentially by training. However, why this occurs remains to be determined. Regardless, this novel dissociation is indicative of a qualitative rather than quantitative difference between AS and S JOSs, as one would expect if AS JOSs reflect actual solving whereas S JOSs capture intuitions about solvability. This difference between AS and S JOSs, coupled with their different predictiveness of solving outcomes (as discussed next), help rule out the possibility that AS JOSs are simply stronger S JOSs.

JOS Predictiveness

Our study also clarified whether intuitions about problem solvability, as measured by JOSs, predict later reasoning performance and effort regulation. One of our key measures compared how often anagrams were solved as a function of how often they had received AS versus S versus NS JOSs in the JOS phase. In Experiments 1 and 2, anagrams that received AS JOSs were more likely to be solved than those that had received S or NS JOSs. But anagrams that received S JOSs were not more likely to be solved than those that received NS JOSs, consistent with prior evidence that JOSs are limited in their ability to predict reasoning success (e.g., Ackerman & Beller, 2017; Lauterman & Ackerman, 2019). However, because

participants in Experiments 1 and 2 knew that each anagram was solvable, they may have put equal effort into solve each anagram for the full 45 s regardless of their JOS. In Experiment 3 we included both solvable and unsolvable anagrams in the solving phase, and participants decided how much time to spend on their solving attempts. These conditions reduced the solving rate relative to Experiments 1 and 2, thus providing more room for solving rate to vary as a function of JOS. Here the solving rate was higher for anagrams that had received S than NS JOSs.

The extent to which solving phase effort and outcomes are influenced by memory for one's JOS is another important issue for future research to tackle. Remembering having indicated that an anagram is solvable is likely to increase one's efforts to solve it. In our two-phase paradigm, the delay between JOSs and solving attempts should reduce the likelihood that participants' solving efforts are solely determined by memory for the JOS—at least relative to an interleaved paradigm where each JOS is immediately followed by a solving attempt. However, the extent to which participants attempt to align their solving efforts with their intuitive judgements remains unknown. If intuitions about solvability are stable over time (Stagnaro et al., 2018) then the intuition that a problem is solvable might recur when the same problem is presented in the solving phase—even if the solver does not remember the JOS or intuition they experienced earlier for the same problem.

Our second measure of solving outcomes considered whether, for each JOS, solved outcomes were more likely than not-solved outcomes. We consistently found that solvable anagrams given AS JOSs were more frequent among solved anagrams than among not-solved anagrams, whereas solvable anagrams given NS JOSs were more frequent among not-solved anagrams than solved anagrams. But critically, even though S JOSs were discriminating, they were not more common among solved anagrams than among not-solved anagrams. Some prior research suggests that solvable JOSs predict solving outcomes (Markovits et al., 2015;

Siedlecka et al., 2016). We found that removing AS JOS trials eliminated S JOS predictiveness, suggesting that the effect in these studies may have arisen due to the inclusion of problems solved during the JOS task. Therefore, we recommend that where spontaneous solving is possible, an AS JOS option be provided to enable participants' intuitions to be separated from their solutions.

On the other hand, as discussed earlier, removal of AS JOSs may underestimate the predictiveness of S JOSs, given that intuitive feelings of solvability likely precede AS JOSs. Had we used problem durations short enough to eliminate AS JOSs, we may have obtained more solved than not-solved outcomes for S JOSs—so long as participants did not revert to random guessing or biases.

As was true in Topolinski et al. (2016), we did not measure whether participants' AS JOSs were accompanied by a solution at that time. Consequently, it remains unclear whether AS JOSs reflect high-confidence intuitions or actual solving. However, since solving rates for AS JOSs were very high, and solving times were fastest for AS JOSs, we suspect that AS JOSs typically reflect genuine solving. Nonetheless, future JOS research using a two-phase design could explore this question by asking participants to report solutions to anagrams they indicate as having already solved during the JOS phase.

Experiment 3 also assessed the rates of not-solvable responses during the solving phase. This rate was higher for anagrams given NS JOSs than for those given either of the other JOSs (i.e., AS+S JOSs). This finding is in line with Lauterman and Ackerman's (2019) evidence that initial JOSs predict final JOSs (i.e., a participant's final judgement about whether an unsolved problem was solvable). Similarly, not-solvable response times during the solving phase were shorter when anagrams were given NS JOSs than AS+S JOSs, in line with Lauterman and Ackerman's finding that participants spend more time on problems they judge as solvable, regardless of their actual solvability. Together, these findings highlight that

JOSs can predict later effort regulation and can help problem solvers optimize their effort regulation (i.e., so as not to waste efforts on unsolvable problems).

Next, we turn to a consideration of the impact of training on JOS predictiveness. Did exposure to blocks of longer-duration anagrams lead JOSs to be more predictive of solving phase outcomes? In general, we did not find this to be the case. However, in Experiment 3 only the training group produced more solved than not-solved outcomes for solvable anagrams given AS JOSs. Of course, this difference is not surprising given that longer-duration anagrams should result in more solving during the solving phase. We also found more not-solved outcomes than solved outcomes for solvable anagrams given NS JOSs in the training group, but not in the no-training group. Longer deliberation of solvability without a solution may lead participants to judge it as not solvable during the solving phase (Payne & Duggan, 2011). Given that the training group had longer to deliberate solvability in blocks 1-3, they may have exhausted all letter arrangements for some anagrams during the JOS phase and thus defaulted to not-solvable responses for them in the solving phase.

Importantly, training did not result in more solved than not-solved outcomes for anagrams given S JOSs. Earlier, we suggested that longer-duration anagrams may lead the training group to shift toward solving the anagrams rather than merely assessing solvability, thus robbing them of opportunities to learn how to regulate their JOSs. This might also explain why we did not detect an effect of training on S JOS predictiveness. Earlier, we also suggested that providing trial-level accuracy feedback after each JOS might improve S JOS discrimination. However, participants might use their memory for this feedback to regulate their efforts in the solving phase, rather than relying on their intuition. If so, then providing feedback might actually undermine solving performance. To assess this possibility, future research could examine whether providing trial-level feedback about JOS accuracy for one

set of problems affects discrimination for another set of problems presented without feedback, and whether JOSs predict solving outcomes selectively for the latter set.

Paradigms for Measuring JOSs

Our use of a two-phase paradigm, in conjunction with collecting AS JOSs, enabled us to separate the effects of intuitions from deliberate solving. However, it remains to be determined whether participants use memory for their JOSs to regulate their solving attempts—and whether memory for JOSs also impacts JOS predictiveness. In an interleaved paradigm, a solving attempt immediately follows each JOS, thus memory for the JOS likely influences one's problem-solving efforts. We are currently comparing these two paradigms.

Implications for Learners

Our results have some clear implications for learners. For instance, students taking timed tests need to learn how to strategically regulate their time and effort to maximize their performance. The ability to discriminate between questions they can versus cannot answer enables students to regulate their effort toward solvable problems. Our studies suggest that merely judging a problem to be solvable (S JOSs) was not predictive of later problem-solving success. Moreover, it can also mislead effort regulation; reasoners take longer to abandon problems they judged to be solvable, especially under greater time pressure. An important direction for future research is to investigate how to train and optimise JOSs to appropriately shift effort and increase successful solving.

Conclusion

Our study establishes that meta-reasoning judgments about solvability are sensitive to whether a problem is actually solvable and can sometimes influence subsequent regulation of problem-solving effort. We found that meta-reasoning judgements about solvability remained accurate when spontaneously solved items were excluded. Meta-reasoning research is still in the early stages, and our study highlights the need to measure solving during the JOS process

both for its effects on measures of intuition and on later problem-solving performance. Our findings highlight an interesting discrepancy regarding judgements of solvability, namely that they can be discriminating and yet not be predictive of later solving. More research is needed to examine the generality of our findings as a function of the type of problems being solved, and as a function of the paradigm used to measure JOS discrimination and predictiveness.

Supplementary Materials

Supplementary Table 2.1

Experiment 2: JOS Phase ANOVA Results for AS+S JOSs

Effect	Training Group					No-training Group				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2
Discrimination	1, 90	28.55	201.33	< .001	.69	1, 90	5.33	46.13	< .001	.34
Block	3, 270	0.003	0.11	.95	.001	3, 270	0.01	0.37	.77	.004
Discrimination \times Block	3, 270	0.40	17.54	< .001	.16	3, 270	0.01	0.80	.50	.01

Supplementary Table 2.2*Experiment 2: JOS Phase Discrimination ANOVAs in Blocks 1-3 for AS+S JOSs*

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2
Discrimination \times Group \times Block	2, 360	0.13	6.72	.001	.04
Discrimination	1, 180	24.83	239.97	< .001	.57
Discrimination \times Group	1, 180	4.94	47.78	< .001	.21
Block	2, 360	0.000	0.01	.99	< .001
Block \times Group	2, 360	0.01	0.36	.70	.002
Block \times Discrimination	2, 360	0.03	1.75	.18	.01
Group	1, 180	0.99	4.88	.03	.03

Supplementary Table 2.3*Experiment 2 and 3: JOS Phase Discrimination ANOVAs in Block 4 for AS+S JOSs*

Effect	Experiment 2					Experiment 3				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2
Discrimination	1, 180	4.74	102.88	< .001	.36	1, 295	14.27	328.02	< .001	.53
Condition	1, 180	0.20	2.02	.16	.11	1, 295	0.003	0.04	.83	< .001
Discrimination \times Group	1, 180	0.18	3.93	.046	.02	1, 295	0.03	0.63	.43	.002

Supplementary Table 2.4*Experiment 3: JOS Phase ANOVA Results for AS+S JOSs*

Effect	Training Group					No-training Group				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2
Discrimination	1, 149	67.83	801.09	< .001	.84	1, 146	27.70	215.22	< .001	.60
Block	3, 447	0.03	1.51	.21	.01	3, 438	0.05	1.76	.16	.01
Discrimination \times Block	2.89, 431.01*	1.30	61.30	< .001	.29	3, 438	0.02	1.06	.37	.01

Supplementary Table 2.5*Experiment 3: JOS Phase Discrimination ANOVAs in Blocks 1-3 for AS+S JOSs*

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η_p^2
Discrimination \times Group \times Block	1.98, 583.20	0.46	22.58	< .001	.07
Discrimination	1, 295	77.95	902.67	< .001	.75
Discrimination \times Group	1, 295	5.18	59.97	< .001	.17
Block	1.98, 582.93	0.10	4.19	.02	.01
Block \times Condition	1.98, 582.93	0.002	0.07	.93	< .001
Block \times Discrimination	1.98, 583.20	0.30	14.57	< .001	.05
Group	1, 295	0.04	0.34	.56	.001

Supplementary Table 2.6*Experiments 1-3: Mean Proportions and Standard Deviations for Solving Phase Solutions*

	<i>Training</i>		<i>No-training</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Experiment 1	.85	.16		
Experiment 2	.81	.21	.74	.25
Experiment 3	.71	.21	.68	.22

Supplementary Table 2.7*Experiment 3 Solving Phase: Mean Proportions and Standard Deviations for Final Not-Solvable Responses to Unsolvable Anagrams*

	<i>Training</i>		<i>No-training</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Experiment 3	.80	.18	.77	.20

Supplementary Table 2.8

Example of how Each Solving Phase Measure was Calculated for a Participant

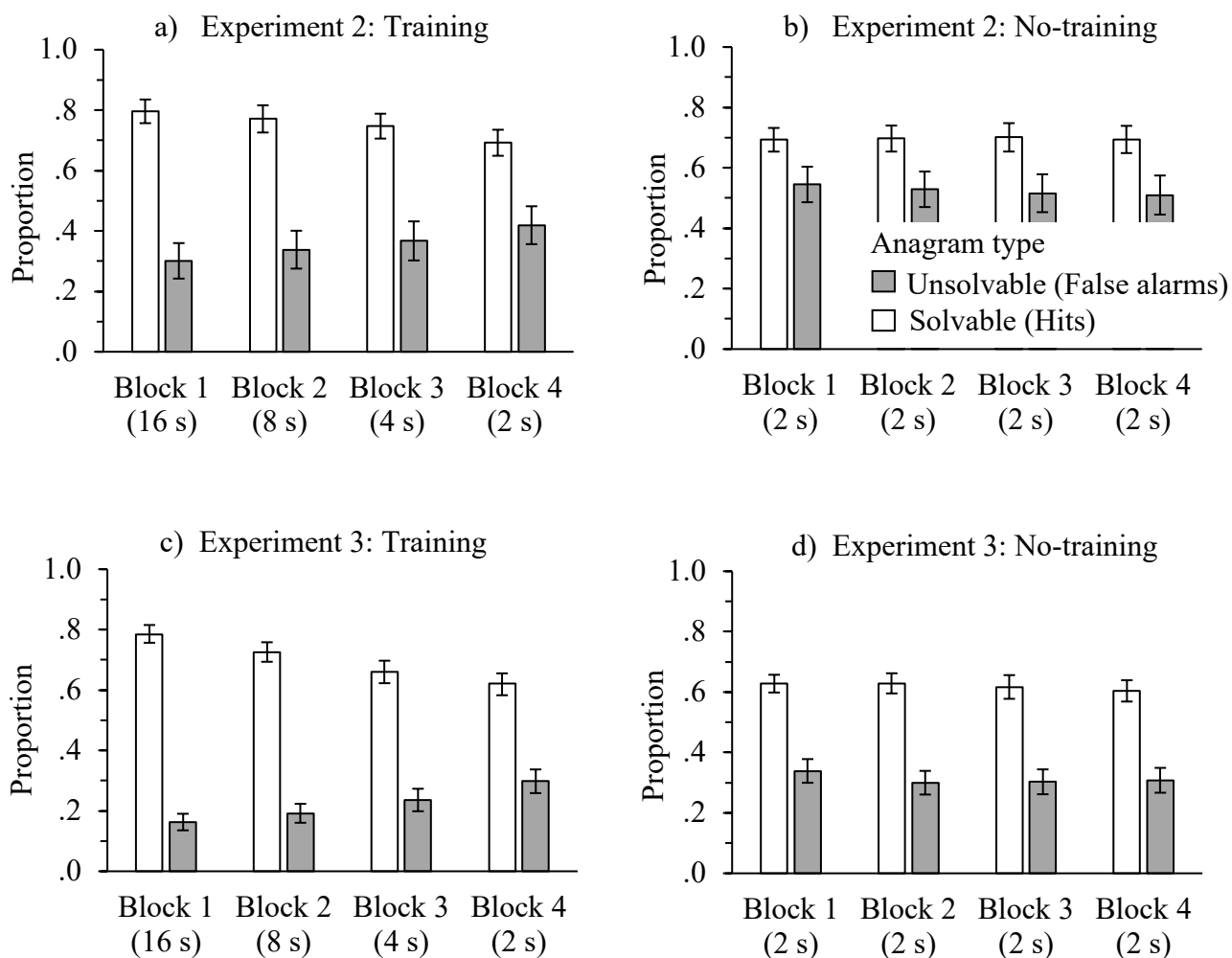
Anagram/Solution	JOS	Solved in Solving Phase?
HFTIG/FIGHT	Already Solved	Yes
BIACN/CABIN	Not Solvable	No
EZRIP/PRIZE	Solvable	Yes
ELCAB/CABLE	Solvable	Yes
TCAHM/MATCH	Solvable	Yes
RLCKE/CLERK	Not Solvable	No
DSTMI/MIDST	Already Solved	Yes
UKRND/DRUNK	Not Solvable	Yes
CIRTK/TRICK	Solvable	No
DUNWO/WOUND	Solvable	No
NOTUC/COUNT	Not Solvable	Yes
RFADU/FRAUD	Already Solved	Yes
ORFNW/FROWN	Solvable	Yes
HOTYU/YOUTH	Solvable	No
AESUB/ABUSE	Already Solved	Yes
HSFAL/FLASH	Already Solved	Yes
KURNT/TRUNK	Solvable	Yes
APTLN/PLANT	Already Solved	Yes
OLHTC/CLOTH	Already Solved	Yes
KLBCO/BLOCK	Not Solvable	No

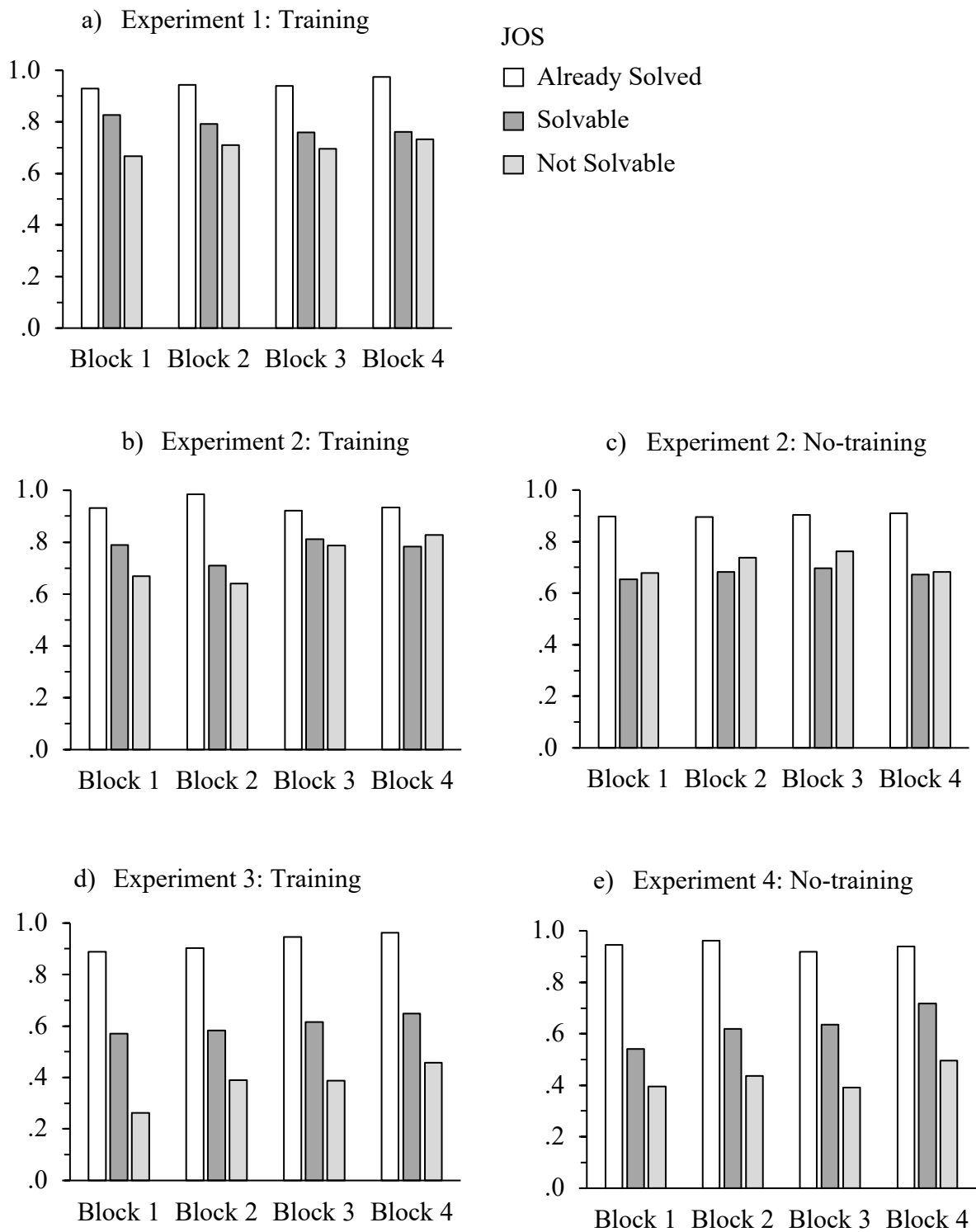
Notes. To calculate the *proportion solved* for each JOS, divide the number of solved anagrams that had received a given JOS during the JOS phase by the total number of anagrams that had received that JOS during the JOS phase. For example, the number of times an anagram given a S JOS was solved was 5. The total number of times participant gave a S JOS was 8. Therefore, the proportion solved for S JOSs was $5/8 = .63$. To calculate the *solved versus not-solved outcomes*, divide the number of solved anagrams that had received a given JOS during the JOS phase by the total number of anagrams that had received that JOS during the JOS phase. For example, the total number of anagrams *solved* was 15. Of the *solved* anagrams, the number that received a S JOS was 5. Therefore, the proportion of *solved outcomes* for S JOSs was $5/15 = .33$. The total number of anagrams *not solved* was 6. Of the *not-solved* anagrams, the number that received a S JOS was 3. Therefore, the proportion of *not-solved outcomes* for S JOSs was $3/6 = .50$.

Supplementary Figure 2.1

Experiments 2 and 3: Mean Proportions of Hits and False Alarms for AS+S JOSs in the JOS

Phase (Bars show 95% CI of each mean)



Supplementary Figure 2.2*Mean Proportion of Anagrams Solved in Each Block as a Function of JOS*

Chapter 3: Linking Judgements of Solvability, Solving Success, and Cognitive Reflection

Author contributions: GEB and I conceptualised the study design. I programmed the experiment, collected and cleaned the data, and performed the data analyses. PW advised me on which data analyses to carry out and assisted in interpretation of the data. I drafted the manuscript and GEB and PW provided critical revisions. GEB approved the final version of the manuscript for submission.

Abstract

As people reason they also engage in meta-reasoning wherein they assess the quality of their reasoning processes. Meta-reasoning begins with a Judgement of Solvability (JOS)—an intuitive assessment of a problem’s solvability. Burton et al. (2022) found that people’s ‘solvable’ JOSs discriminated solvable from unsolvable anagrams, but JOSs only predicted later solving for anagrams that they indicated having solved while making their JOS. Here we investigated whether individual differences in cognitive reflection (i.e., one’s inclination to reflect on one’s cognitive processes) are related to these outcomes. To this end, we reanalysed two of Burton et al.’s experiments with consideration of participants’ Cognitive Reflection Test scores. Greater cognitive reflection led to more ‘already solved’ JOSs, particularly when longer-duration anagrams were presented in initial blocks within the JOS task. Although greater cognitive reflection was related to a higher proportion of anagram solving during the solving trials, it was only related to JOS predictiveness for the anagrams that participants reported having solved during the JOS task. In sum, cognitive reflection was associated with anagram-solving ability, but it did not predict better meta-reasoning discrimination or predictiveness.

Introduction

Meta-reasoning involves self-evaluating one's reasoning processes (Ackerman & Thompson, 2017). People meta-reason to determine whether to engage in reasoning, the likelihood that their reasoning will be successful, and how confident they are in their reasoning outcomes (Ackerman et al., 2020). According to Ackerman and Thompson's (2017) meta-reasoning framework, problem-solving begins with a *Judgement of Solvability* (JOS), which is an intuitive assessment of problem solvability that occurs before a problem-solving attempt. Some research has shown that these initial meta-reasoning judgements can accurately distinguish between solvable and unsolvable problems (i.e., JOS discrimination; Balas et al., 2011; Bolte & Goschke, 2005; Burton et al., 2022; Novick & Sherman, 2003; Topolinski et al., 2016; Topolinski & Strack, 2009a; Undorf & Zander, 2017), and can sometimes predict problem-solving success and effort-regulation (i.e., *JOS predictiveness*; Markovits et al., 2015; Siedlecka et al., 2016).

Recently, Burton et al. (2022) reported that intuitive JOSs were able to discriminate solvable from unsolvable anagrams, even when anagrams solved during the JOS task were excluded from the analyses. JOSs were more discriminating when anagrams were presented for longer durations to enable reasoners to develop better intuition about each anagram's solvability. However, although solvable (*S*) JOSs were discriminating, *S* JOSs were not more frequent for anagrams that were successfully solved, versus anagrams that were not solved. Moreover, *S* JOSs did not generally lead to greater solving rates than not solvable (*NS*) JOSs. Thus, JOSs indicated correct beliefs about anagram solvability but were generally not predictive of later problem-solving outcomes.

Here we investigated whether individual differences in meta-reasoning, specifically regarding *cognitive reflection*, are related to one's ability to make accurate and predictive JOSs. Cognitive reflection refers to a disposition to reflect on one's thought processes.

Greater cognitive reflection has been shown to facilitate better *meta-reasoning*. Specifically, those with greater cognitive reflection have final confidence judgements that are better calibrated with objective reasoning performance (Coutinho et al., 2021; Duttler, 2016; Noori, 2016; Pennycook et al., 2017). To examine the potential effect of cognitive reflection on the initial monitoring during meta-reasoning, we reanalysed two of Burton et al.'s (2022) experiments in which Cognitive Reflection Test (CRT) scores were collected. We measured whether CRT scores affect JOS discrimination and predictiveness. In doing so, our study tested whether the effect of cognitive reflection on reasoning extends to initial meta-reasoning judgements about problem solvability.

Cognitive Reflection and Meta-Reasoning

Dual-process models of reasoning suggest that there are two distinct reasoning processes: a fast System 1, and a slow System 2 (Evans, 2007; Stanovich, 1999). System 1 is automatic and relies on heuristics, but if these heuristics are unreliable, System 1 can generate erroneous reasoning (De Neys et al., 2010; Evans & Curtis-Holmes, 2005; Hoppe & Kusterer, 2011). For example, Frederick's (2005) now famous bat and ball problem, "*A bat and a ball together cost \$1.10. The bat costs \$1 more than the ball. How much does the ball cost?*" can generate a strong intuitive System 1 response of 10 cents. Reasoners who reflect on their System 1 response may realise that 10 cents is incorrect, and may need to override their erroneous intuition using their more analytical System 2 to achieve the correct response of 5 cents (Evans, 2019; Kahneman, 2011; Pennycook et al., 2017). However, not all reasoners will detect errors or biases in their System 1 processes; many are "cognitive misers" who lack the willingness and ability to detect and correct their initial System 1 response (Pennycook et al., 2015a; Toplak et al., 2014). Those who detect errors in their reasoning might be disposed to reflect on their cognitive processes to a greater extent (Pennycook et al., 2015b; Šrol & De Neys, 2021). Thus, some researchers contend that the

ability to reflect on one's cognitive processes and detect errors in reasoning relates to meta-reasoning (Mata et al., 2013; Pennycook et al., 2017).

Is Cognitive Reflection Related to JOS Discrimination and Predictiveness?

The ability to “decouple” from erroneous intuition, and instead engage in more analytic thinking, might improve one's ability to regulate one's JOSs. Burton et al. (2022) allowed participants to report when they had spontaneously solved an anagram during the JOS task. They argued that solving the anagram during a JOS trial might help participants learn to regulate their JOSs, given that an intuition that a problem is solvable should precede a solution (Ackerman & Beller, 2017). After several JOS trials, a more reflective thinker might note that their intuition has been biased by certain misleading anagram letter combinations or differences in bigram frequencies. They might observe this after noticing a miscalibration between their intuition and a spontaneous solution during a JOS trial. For example, a reasoner may feel an anagram is unsolvable if the anagram has a small vowel-to-consonant ratio (e.g., ESLTY is a solvable anagram of STYLE, but only has one consonant). However, if the solution is spontaneously retrieved during the JOS, that provides feedback to confirm that they were using unreliable cues to inform their intuition (Burton et al., 2022). Reflective thinkers might then readjust their intuition and find more reliable cues with which to make their JOSs, thus decoupling from their incorrect intuition and self-regulating more accurate intuition. Hence, the ability to make more accurate JOSs may depend on a reasoner's disposition to detect errors or biases in their JOS intuition, and instead regulate more accurate meta-reasoning (Mata et al., 2013), in turn improving JOS discrimination and predictiveness.

Furthermore, anagrams are insight reasoning problems, and solving anagrams sometimes requires “restructuring” of the problem after encountering a solving impasse (Ash & Wiley, 2006; Ohlsson, 2011). For example, a reasoner might decide to rearrange the letters

of an anagram differently after multiple failed solving attempts, leading to a sudden instantiation of the solution (Gilhooly & Fioratou, 2009). Of course, not all insight reasoners will restructure the problem; some may maintain an unsuccessful problem-solving strategy despite its ineffectiveness (DeYoung et al., 2008). The tendency to restructure an insight problem might relate to a disposition to disconnect from erroneous System 1 reasoning. An intuitive feeling of solvability should promote greater problem-solving effort regulation, and in turn, greater problem-solving success (Pennycook et al., 2015a; Thompson et al., 2011). However, a reasoner lacking in cognitive reflection may not identify that their current reasoning strategy is unsuccessful, and in turn, may not separate from their incorrect intuition about their reasoning strategy by restructuring the problem – instead, they may simply skip the problem. Thus, the ability of a reasoner’s ‘solvable’ JOSs to predict problem-solving success may depend on whether they are motivated to reflect on their insight reasoning processes; in this respect, cognitive reflection would be related to JOS predictiveness.

The Effect of Task Factors on the Relationship Between JOSs and Cognitive Reflection

Burton et al. (2022) presented initial blocks of anagrams for either long durations (i.e., training group) or short durations (i.e., no-training group) and found that anagram presentation duration impacted JOS discrimination. Unsurprisingly, when given longer durations to develop intuition about solvability, participants made ‘solvable’ JOSs that were more discriminating than at shorter durations. More surprisingly, however, training did not lead to more predictive JOSs relative to no-training.

Whether training influences JOS discrimination and predictiveness might depend in part on cognitive reflection ability. At briefer presentation durations, reasoners are more likely to rely on unreliable heuristic cues to make their decisions (Ackerman, 2019; Benjamin, 2005; Kahneman et al., 1982). However, reflective thinkers may be better able to reflect and identify when they are using these unreliable cues, thus leading to more accurate

regulation of JOSs at shorter durations. Hence, the inclusion of longer-duration anagrams in the training group might be of greater benefit to those with *poorer* cognitive reflection – because longer-duration anagrams are likely to lead to more anagram-solving successes that might help less reflective participants generate a better sense of anagram solvability. Overall, we expected that greater cognitive reflection would be related to better JOS discrimination and predictiveness for both training and no-training groups. However, we expected that training might work to mitigate individual differences in cognitive reflection on JOS discrimination and predictiveness.

Overview

Further analyses were conducted on the data from Experiments 2 and 3 of Burton et al. (2022). The JOS phase was identical in both experiments: participants made JOSs to 4 blocks of anagrams, each consisting of a random mixture of 10 solvable and 10 unsolvable anagrams. Participants judged each anagram as solvable, not solvable, or already solved (*S*, *NS*, *AS*). In a training group, anagrams were presented for 16 s in block 1, and anagram duration thereafter halved across blocks (8 s, 4 s, 2 s). In a no-training group, anagrams were presented for 2 s in each block. After the anagram was presented, participants had 3 s to make their JOS. Participants then completed a solving phase, which differed across the two experiments. In Experiment 2, participants were given up to 45 s to try to solve each of the 40 solvable anagrams from the JOS phase (presented in random order). In Experiment 3, half of the solvable and unsolvable anagrams from each block (40 anagrams in total) were presented for a solving attempt, and no time limit was placed on solving (to enable self-regulation of effort). Here, for each trial, the participant either typed in a solution, passed, or classified it as not solvable.

After the solving phase, all participants completed our behavioural measure of cognitive reflection: the 7-item Cognitive Reflection Test (CRT; Frederick, 2005; Toplak et

al., 2014). The 7-item CRT is designed to assess one's propensity to engage in reflective thinking by posing "trick" questions which are designed to elicit a rapid but incorrect response. Some research argues that the CRT is a valid measure of individual differences in metacognition because of its ability to separate those who reflect on their thinking and detect errors or biases in their reasoning, or have unbiased reasoning, from those who do not (Pennycook et al., 2014; Pennycook et al., 2015a).

Method

The experiment was pre-registered on OSF at <https://osf.io/cq2kb>. The data for this study are available in Open Science Framework at <https://doi.org/10.17605/OSF.IO/JD5S9>.

Participants

A total of 595 participants were recruited through Amazon's Mechanical Turk (MTurk) via TurkPrime (Litman et al., 2017). The MTurk inclusion and data exclusion criteria are specified in Burton et al. (2022). Here we analysed data from 479 participants from Experiments 2 and 3 of Burton et al. (322 female, 157 male, mean age = 42.9, $SD = 13.2$), comprising 91 participants in each of the training and no-training groups from Experiment 2, and 150 in the training group and 147 in the no-training group from Experiment 3.

Stimuli

The stimuli were 40 solvable and 40 unsolvable anagrams (see Burton et al., 2022).

The 7-item version of the CRT was used to measure cognitive reflection (Frederick, 2005). Each CRT item was designed to cue an intuitive but incorrect response. Correct answers were summed to calculate a CRT score. Higher scores indicate higher cognitive reflection ability.

Procedure

The experiments were conducted online in Qualtrics (2019). The JOS phase was identical in both experiments. Participants were instructed that on each JOS phase trial, they would be presented with an anagram for a set duration, some of which could form a solution word (e.g., DSTMI - MIDST) and some of which could not (e.g., ZEREB). Their task was to make one of three solvability judgements for each anagram: “YES it is solvable”, “NO it is not solvable”, or “I have already solved it”. Participants had 3 s to provide their judgement after the anagram was presented. Participants completed 80 JOS phase trials over 4 blocks (20 JOS phase trials per block). In the training group, participants were told that anagram duration would decrease across blocks (16 s, 8 s, 4 s, 2 s). In the no-training group, participants were told that anagram duration would be 2 s in each block.

Following the JOS phase, participants completed the solving phase where 40 anagrams from the JOS phase were presented in random order. In Experiment 2, participants were given up to 45 s to try to solve each of the 40 solvable anagrams from the JOS phase. In Experiment 3, half of the solvable and unsolvable anagrams from each block (40 anagrams in total) were presented, there was no time limit, and participants could either type in a solution, “P” to pass (if they believed the anagram was solvable but were unable to solve it) or “N” to indicate that they thought it was ‘not solvable’.

Participants completed the 7-item CRT immediately after the solving phase. They were informed that they could spend as much time solving each problem as they needed.

Results

The results of these experiments are reported in Burton et al. (2022). Here we examined whether cognitive reflection is related to JOS discrimination and/or predictiveness and whether any effects of cognitive reflection on these measures differed when participants had experience with longer-duration anagrams compared to when anagram duration was

always 2 s (i.e., the potential impact of training). Given these foci, our analyses of both JOS discrimination and predictiveness were collapsed across JOS phase blocks.

Separate ANCOVAs were conducted for each dependent variable (JOS discrimination, proportion solved, and solving outcomes). In each ANCOVA, Group (training vs. no-training) and Experiment (2 vs. 3) were between-subjects factors and mean-centred CRT was a covariate. Categorical main effects or interactions were explored using pairwise comparisons, and significant main effects or interactions involving the CRT were further investigated using regression analyses. Analyses used .05 as the alpha level and used η^2_p as the measure of effect size for the ANCOVAs, and R^2 for the regressions.

Effects of CRT on JOS Phase Discrimination

We used Burton et al.'s (2022) JOS discrimination measures to create a *JOS discrimination difference score* by subtracting participants' mean false alarm rate from their mean hit rate, separately for AS JOSs and S JOSs. Thus, discrimination scores for each JOS could range from +1 (hit rate = 1, false alarm rate = 0) to -1 (hit rate = 0, false alarm rate = 1). Table 3.1 presents the complete ANCOVA results for each JOS. Our focus here was on whether cognitive reflection, as assessed using the CRT, was related to discrimination, and whether the relationship between CRT and discrimination was moderated by training or experiment factors.

In the AS JOS ANCOVA, the main effect of CRT was significant: Higher CRT scores were associated with greater AS JOS discrimination, $R^2 = .03$, $B = 0.02$. The CRT \times group interaction was significant, and this was qualified by a significant CRT \times group \times experiment interaction. Figure 3.1 provides a graphical depiction of the three-way interaction. The three-way interaction was followed up using separate interaction contrasts for each experiment (see Table 3.2), for which we focus on the CRT \times group interaction. The CRT \times group interaction was significant in Experiment 2, but not in Experiment 3. In Experiment 2, higher CRT

scores were associated with greater AS JOS discrimination in the training group (i.e., when longer duration anagrams were present), $F(1, 85) = 14.44$, $MSE = 1.38$, $p < .001$, $R^2 = .15$, $B = .05$, whereas CRT did not affect AS JOS discrimination in the no-training group (i.e., when longer duration anagrams were not presented), $F(1, 85) = 0.35$, $MSE = 0.02$, $p = .56$, $R^2 = .004$, $B = -.01$.

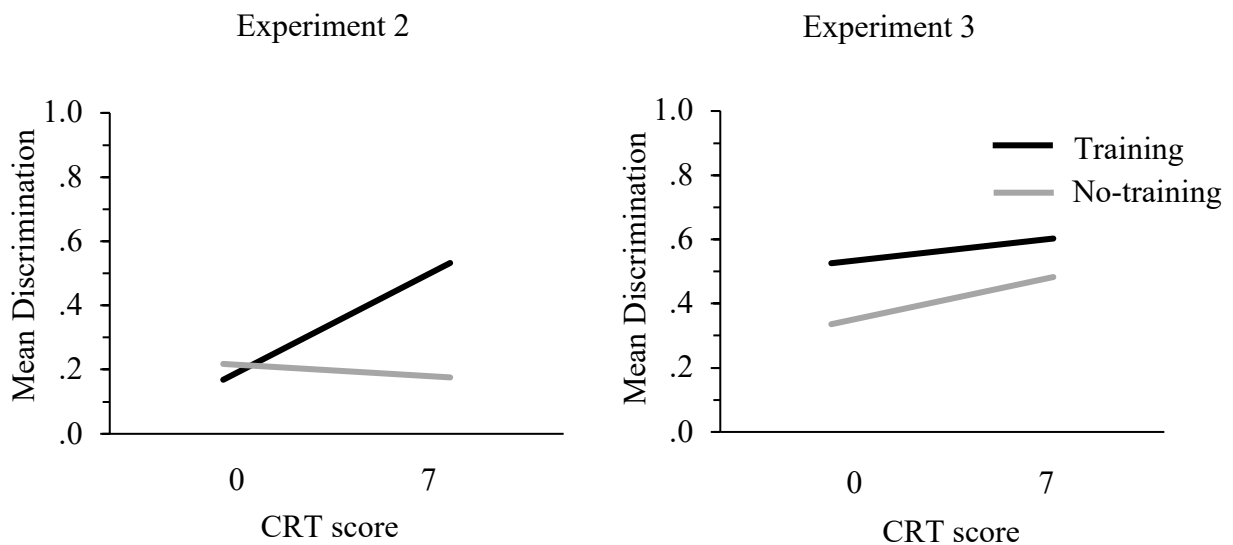
Table 3.1

JOS Discrimination: CRT ANCOVA Results by JOS

<i>JOS(s)/Effect</i>	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs					
CRT	1, 463	0.94	13.31	< .001	.03
CRT × Group	1, 463	0.37	5.28	.02	.01
CRT × Experiment	1, 463	0.03	0.43	.51	.00
CRT × Group × Experiment	1, 463	0.73	10.32	.001	.02
Experiment	1, 463	4.99	71.02	< .001	.13
Group	1, 463	2.44	34.66	< .001	.07
Experiment × Group	1, 463	0.01	0.09	.77	.00
S JOSs					
CRT	1, 469	0.00	0.04	.84	.00
CRT × Group	1, 469	0.03	0.79	.38	.00
CRT × Experiment	1, 469	0.01	0.21	.65	.00
CRT × Group × Experiment	1, 469	0.05	1.38	.24	.00
Experiment	1, 469	0.23	6.02	.02	.01
Group	1, 469	0.65	17.07	< .001	.04
Experiment × Group	1, 469	0.15	4.02	.05	.01

Figure 3.1

AS JOSs: Mean Discrimination as a Function of Experiment, Group, and CRT Score

**Table 3.2**

AS JOS Discrimination: CRT \times Group Interaction Contrasts

<i>Experiment/Effect</i>	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
Experiment 2					
CRT	1, 170	0.53	6.84	.01	.04
CRT \times Group	1, 170	0.86	11.21	.10	.06
Group	1, 170	0.87	11.29	.001	.06
Experiment 3					
CRT	1, 293	0.41	6.22	.01	.02
CRT \times Group	1, 293	0.04	0.58	.45	.00
Group	1, 293	1.83	27.56	< .001	.09

It is unclear why the CRT \times group interaction occurred only in Experiment 2, given that the JOS phase was identical in Experiments 2 and 3. Mean CRT scores did not differ significantly between Experiment 2 ($M = 3.02$, $SD = 2.35$) and Experiment 3 ($M = 3.33$, $SD = 2.34$), $t(476) = 1.22$, $p = .22$. However, AS JOSs were more discriminating in Experiment 3 than Experiment 2 (note the robust main effect of Experiment in Table 3.1). Relatedly, AS JOS variance was lower in Experiment 3 than in Experiment 2, resulting in a violation of the homogeneity of variance assumption, Levene's $F(3, 467) = 4.61$, $p = .003$. Our Discussion notes some additional potential reasons for this unexpected cross-experiment difference.

In the S JOS ANCOVA, neither the main effect of CRT nor any interactions involving CRT were significant. Thus, unlike for AS JOSs, higher CRT scores were not linked with more accurate S JOS discrimination, and the CRT did not interact with group and/or experiment.

Solving Phase

We next examined whether CRT scores moderated the relationship between JOSs and the two solving-phase measures reported in Burton et al. (2022): *proportion solved* and *proportion of solved versus not-solved outcomes*. Here our focus was on whether there was a main effect of the CRT, and whether the CRT interacted with our other factors.

Proportion Solved

The *proportion solved* measure was calculated as the number of anagrams solved in the solving phase that were given a particular JOS during the JOS phase. For example, if a participant gave S JOSs to 8 anagrams in the JOS phase, and then went on to solve 4 of those anagrams in the solving phase, their proportion solved for S JOSs would be .5. The proportions for each JOS were independent and could range from 0 to 1. The supplementary materials from Burton et al. (2022) provide a full illustrative example.

The mean proportion of anagrams solved as a function of JOS type (AS vs. S vs. NS) served as the repeated-measures factor in a mixed-factor ANCOVA, in which experiment and group were between-group factors and mean-centered CRT was a covariate (see Table 3.3). Although JOS type was significant (see Burton et al., 2022), the CRT did not moderate how well JOSs predicted the proportions of anagrams solved (i.e., the CRT \times JOS interaction was not significant). The ANCOVA revealed a main effect of CRT: higher CRT scores were associated with a higher overall proportion of anagrams solved, $R^2 = .05$, $B = .02$. The only significant interaction involving the CRT was the three-way interaction with group and experiment, which was followed up via separate CRT \times Group interaction contrasts for each experiment. This interaction was marginal in Experiment 2 and was not significant in Experiment 3 (see Table 3.4). Figure 3.2 shows that, in Experiment 2, the association between CRT score and solving was significant in the training group (i.e., when longer duration anagrams were present), $F(1, 88) = 16.38$, $MSE = 0.94$, $p < .001$, $R^2 = .16$, $B = .04$, whereas it was just at the threshold of significance in the no-training group (i.e., when longer duration anagrams were not present), $F(1, 89) = 4.13$, $MSE = 0.17$, $p = .05$, $R^2 = .04$, $B = .02$. Interestingly, then, the CRT predicted solving outcomes only for longer-duration anagrams, and only when participants were required to provide a solution to each anagram (i.e., when the experiment did not enable them to self-regulate their solving efforts).

Table 3.3*Solving Phase: Proportion Solved ANCOVA Results*

<i>Effect</i>	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
CRT	1, 297	1.75	13.96	< .001	.05
JOS*	1.98, 587.18	6.61	109.05	< .001	.27
CRT × JOS*	1.98, 587.18	0.07	1.07	.34	.00
CRT × Group	1, 297	0.16	1.28	.26	.00
CRT × Experiment	1, 297	0.47	3.73	.054	.01
CRT × JOS × Group*	1.98, 587.18	0.04	0.66	.52	.00
CRT × JOS × Experiment*	1.98, 587.18	0.01	0.10	.90	.00
CRT × Group × Experiment	1, 297	0.59	4.71	.03	.02
CRT × Experiment × Group × JOS*	1.98, 587.18	0.02	0.31	.73	.00
Group	1, 297	0.08	0.60	.44	.00
Experiment	1, 297	4.10	32.65	< .001	.10
JOS × Group*	1.98, 587.18	0.02	0.40	.67	.00
JOS × Experiment*	1.98, 587.18	2.26	36.85	< .001	.11
JOS × Group × Experiment*	1.98, 587.18	0.04	0.72	.49	.00

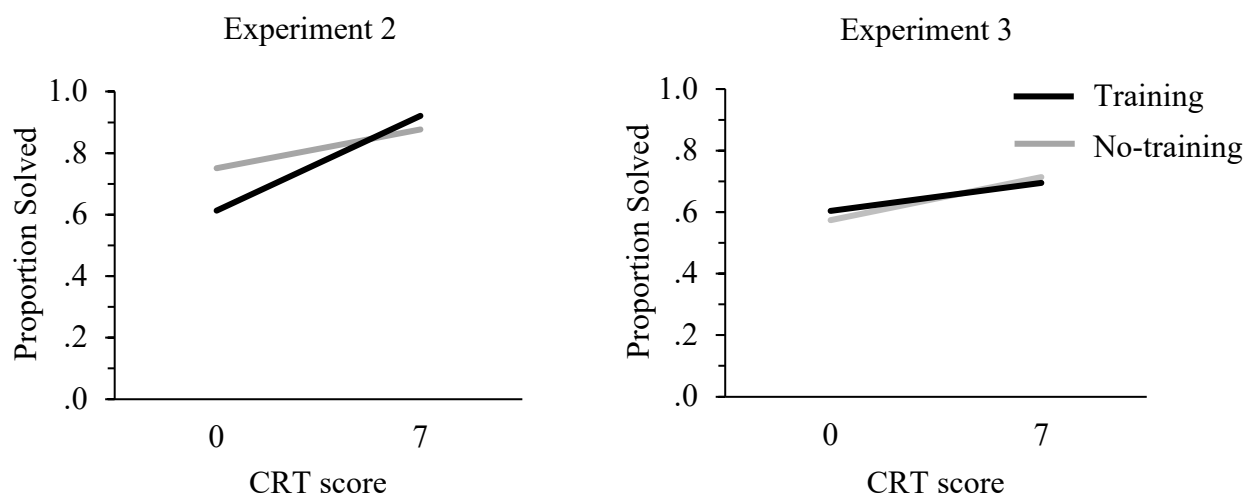
Note. * Huynh-Feldt correction was applied because assumption of sphericity was violated.

Table 3.4*Mean Proportion Solved: CRT × Group Interaction Contrasts*

<i>Experiment/Effect</i>	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
Experiment 2					
CRT	1, 177	0.96	19.69	< .001	.10
CRT × Group	1, 177	0.17	3.49	.06	.02
Group	1, 177	0.12	2.44	.12	.04
Experiment 3					
CRT	1, 293	0.43	10.67	.001	.04
CRT × Group	1, 293	0.02	0.48	.50	.00
Group	1, 293	0.00	0.09	.76	.00

Figure 3.2

Solving Phase: Mean Proportion Solved Collapsed Across JOS as a Function of Experiment, Group, and CRT Score



Proportion of Solved Outcomes – Proportion of Not-Solved Outcomes.

Our second measure of JOS predictiveness was *the proportion of solved versus not-solved outcomes* as reported in Burton et al. (2022). For this measure, we separately calculated the proportion of *solved outcomes* and the proportion of *not-solved outcomes* for each JOS. For example, if a participant solved 8 anagrams in total, and 6 of those anagrams received an S JOS, the proportion of *solved outcomes* for their S JOSs would be .75. If a participant failed to solve 10 anagrams in total, and 3 of those anagrams received an S JOS, the proportion of *not-solved outcomes* would be .3. We then calculated the *solving outcome difference* by subtracting the proportion of not-solved outcomes from the proportion of solved outcomes for each JOS. Thus, for each JOS, the solving outcome difference could range from +1 to -1. A positive score indicated that solved outcomes were more frequent than not-solved outcomes among that JOS, and a negative score indicated that not-solved outcomes were more frequent than solved outcomes among that JOS. Solving outcome differences for each JOS were then analysed using a 2(Experiment: 2, 3) by 2(Group: training, no-training) ANCOVA, again with mean-centred CRT score as a covariate. Table 3.5 shows the complete results for each ANCOVA.

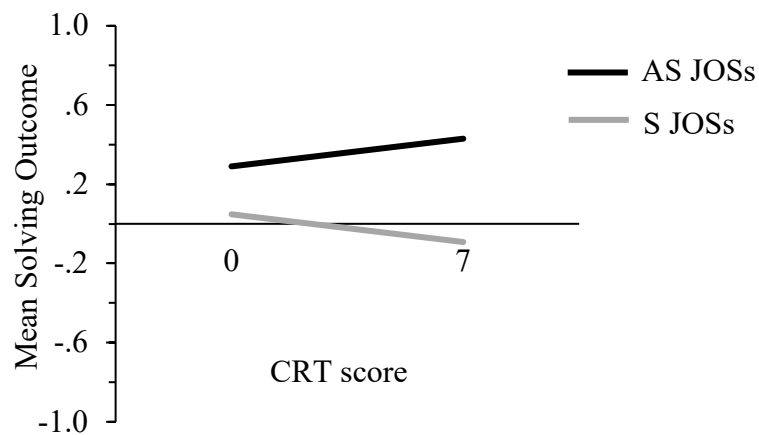
In the AS JOSs ANCOVA, the main effect of CRT was significant: Figure 3.3 shows that higher CRT scores predicted more positive solving outcome differences, $F(1, 435) = 7.67$, $MSE = 0.94$, $p = .01$, $R^2 = .02$, $B = .02$. The S JOS ANCOVA also revealed a main effect of the CRT but in the opposite direction: higher CRT scores predicted more *negative* solving outcome differences, $F(1, 435) = 8.94$, $MSE = 0.73$, $p = .003$, $R^2 = .02$, $B = -.02$. The main effect of CRT was not significant in the NS JOS ANCOVA thus the CRT did not predict solving outcome differences for these JOSs.

Table 3.5*Solving Phase: Solving Outcome ANOVA Results by JOS*

<i>JOS(s)/Effect</i>	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
AS JOSs					
CRT	1, 429	0.51	5.52	.02	.01
CRT \times Group	1, 429	0.01	0.10	.75	.00
CRT \times Experiment	1, 429	0.22	2.44	.12	.01
CRT \times Group \times Experiment	1, 429	0.01	0.08	.78	.00
Group	1, 429	1.03	11.25	.001	.03
Experiment	1, 429	11.70	127.59	< .001	.23
Group \times Experiment	1, 429	0.00	0.02	.88	.00
S JOSs					
CRT	1, 429	0.67	8.18	.004	.02
CRT \times Group	1, 429	0.04	0.44	.51	.00
CRT \times Experiment	1, 429	0.00	0.24	.88	.00
CRT \times Group \times Experiment	1, 429	0.01	0.06	.81	.00
Group	1, 429	0.24	2.96	.09	.01
Experiment	1, 429	0.01	0.15	.70	.00
Group \times Experiment	1, 429	0.48	5.90	.02	.01
NS JOSs					
CRT	1, 429	0.03	0.36	.55	.00
CRT \times Group	1, 429	0.04	0.45	.50	.00
CRT \times Experiment	1, 429	0.11	1.16	.28	.00
CRT \times Experiment \times Group	1, 429	0.10	1.02	.31	.00
Group	1, 429	1.72	18.43	< .001	.04
Experiment	1, 429	10.04	107.81	< .001	.20
Group \times Experiment	1, 429	0.39	4.16	.04	.01

Figure 3.3

Solving Phase: Solving Outcomes Among AS and S JOSs as a Function of CRT Score



Note. Solving outcomes for AS and S JOSs were analysed separately (see Table 3.5).

These findings suggest that participants higher in cognitive reflection may be better at rapidly solving the anagrams in the JOS phase, and hence their JOSs were more influenced by the (AS) solutions found during the JOS task. Burton et al. (2022) noted that removing AS JOS trials may have underestimated S JOS predictiveness, and they further suggested that anagram durations that are short enough to reduce AS JOSs may lead to more predictive S JOSs. The latter effect may be stronger for those higher in cognitive reflection, leading to more AS JOSs, and leaving a smaller set of S JOSs that were predictive in the wrong direction (i.e., the negative relationship suggests that S JOSs were more frequent among not-solved anagrams than among solved anagrams). Thus, the intuition regarding solvability for individuals with higher cognitive reflection may have been compromised by a greater reliance on anagram solutions that occurred during the JOS process.

S JOS discrimination versus S JOS predictiveness

Burton et al. (2022) separately measured S JOS discrimination and S JOS predictiveness, but they did not examine whether one's ability to discriminate solvable from unsolvable anagrams related to one's ability to later solve the solvable anagrams. Thus, here we assessed whether better S JOS discrimination was related to a higher proportion of

anagrams solved for anagrams given S JOSs and whether the CRT moderated this relationship. We analysed the proportions solved for S JOSs in a 2(Experiment: 2, 3) by 2(group: training, no-training) between-groups ANCOVA, with mean-centred CRT and mean-centred S JOS discrimination included as covariates. AS JOS discrimination and predictiveness were not examined because anagrams receiving AS JOSs were almost always solved (see Figure 4 in Burton et al., 2022). We analysed the proportion solved measure (rather than the solved vs. not-solved outcome measure) here because our focus was on whether better S JOS discrimination predicted more anagrams solved, rather than on whether S JOSs were more frequent among solved than among not-solved outcomes.

The ANCOVA (Table 3.6) revealed only one key significant effect: the main effect of S JOS discrimination. We used a standard linear regression to further investigate the relationship between discrimination and the proportions solved for S JOSs. S JOS discrimination positively predicted the rates of solving for anagrams given S JOSs, $F(1, 406) = 37.19$, $MSE = 3.71$, $p < .001$, $R^2 = .08$, $B = 0.50$. Thus, individuals who were more discriminating with their S JOSs were more likely to later solve anagrams they had given an S JOS. Cognitive reflection did not moderate this relationship, and the relationship did not vary as a function of whether the solving phase was self-regulated or not.

Table 3.6

Proportion Solved for S JOSs: ANCOVA with Mean-Centred S JOS Discrimination and Mean-Centred CRT

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
Discrimination	1, 392	2.69	28.08	< .001	.07
CRT	1, 392	0.26	2.75	.10	.01
Discrimination × CRT	1, 392	0.24	2.65	.10	.01
CRT × Group	1, 392	0.04	0.43	.51	.00
CRT × Experiment	1, 392	0.14	1.45	.23	.00
CRT × Group × Experiment	1, 392	0.18	1.86	.17	.01
CRT × Discrimination × Group	1, 392	0.00	0.04	.85	.00
CRT × Discrimination × Experiment	1, 392	0.06	0.57	.45	.00
CRT × Discrimination × Group × Experiment	1, 392	0.07	0.77	.38	.00
Group	1, 392	0.04	0.42	.52	.00
Experiment	1, 392	1.81	18.86	< .001	.05
Group × Discrimination	1, 392	0.07	0.74	.39	.00
Experiment × Discrimination	1, 392	0.00	0.04	.85	.00
Experiment × Group	1, 392	0.03	0.32	.58	.00
Experiment × Group × Discrimination	1, 392	0.05	0.47	.49	.00

Discussion

Previous studies have shown that greater cognitive reflection results in better-calibrated confidence judgements when meta-reasoning (Coutinho et al., 2021; Duttler, 2016; Noori, 2016; Pennycook et al., 2017). Our reanalysis of Burton et al.'s (2022) Experiments 2 and 3 sought to clarify whether greater cognitive reflection, as assessed using participants'

CRT scores, also relates to an ability to make better initial meta-reasoning judgements (i.e., more discriminating and predictive JOS). Our analyses revealed that greater cognitive reflection was rather selective in its effect, in that it was chiefly associated with how often participants reported anagrams as ‘already solved’ during the JOS, and signalled anagram-solving ability more generally.

Cognitive Reflection Predicted Anagram Solving Ability, but Not Solvability Intuition

We anticipated that if cognitive reflection is related to meta-reasoning, then reflective thinkers might generate more accurate intuitions about insight problem solvability. Reflective thinkers might self-regulate more accurate intuition after detecting biases in their intuition. However, we found that cognitive reflection was only related to more rapid insight reasoning ability, rather than more discriminating meta-reasoning decisions about problem solvability. Although studies have reported that cognitive reflection predicts reasoning ability (Patel et al., 2019; Toplak et al., 2011, 2014), we found that greater cognitive reflection did not result in more accurate and predictive JOSs.

Our findings are all the more surprising given recent evidence that cognitive reflection is related to meta-reasoning (e.g., Coutinho et al., 2021; Duttler, 2016; Noori, 2016; Pennycook et al., 2015a; Pennycook et al., 2017; Šrol & De Neys, 2021). Importantly, studies measuring the relationship between cognitive reflection and meta-reasoning have almost exclusively used heuristics and biases tasks (e.g., base rate problems, conjunction fallacy task, belief bias syllogisms). These tasks are specifically designed to mislead meta-reasoning processes by cueing incorrect, intuitive responses. In contrast, our study measured whether cognitive reflection is related to meta-reasoning in a task that was not designed to lead to biased reasoning (i.e., a neutral task, e.g., de Chantal et al., 2020; Goel & Dolan, 2003). Perhaps the anagram task is *too neutral* to detect any effects of cognitive reflection on

intuitive JOS discrimination and predictiveness and does not enable participants to detect self-generated conflict in their intuitions.

To examine this possibility, future research should examine whether cognitive reflection is related to JOS discrimination and predictiveness using stimuli that are designed to elicit biased JOSs. For example, Topolinski et al. (2016) found that processing fluency was an unreliable heuristic cue indicating solvability – pronounceable anagrams were judged as solvable more often than unpronounceable anagrams, contrary to actual solvability. Thus, future studies should assess whether greater cognitive reflection acts as a protective factor against unreliable heuristic cues when making intuitive JOSs.

Cognitive reflection was related to the ability to rapidly solve anagrams during the JOS process, but it was not related to how well S JOSs predicted problem-solving outcomes. In fact, greater cognitive reflection predicted more S JOSs among not-solved anagrams than among solved anagrams. Burton et al. (2022) noted that the inclusion of longer-duration anagrams may have led to lower S JOS predictiveness. The present study suggests that this pattern may have been exacerbated in those with greater cognitive reflection, such that making S JOSs predicted later solving failures of those anagrams. Given that intuitive impressions of solvability should precede AS JOSs, perhaps S JOS intuition would have been more predictive in those with greater cognitive reflection had we used shorter problem durations (i.e., less than 2 s) that minimized spontaneous solutions during the JOS task. Future studies should investigate this possibility.

Furthermore, recent hybrid dual-process models of reasoning challenge the assumption that better cognitive reflection requires detecting and correcting erroneous intuitions (e.g., Bago & De Neys, 2019; De Neys & Pennycook, 2019). For instance, when solving the bat and ball problem, some reasoners can immediately intuit the correct response of 5 cents without needing intervention from System 2 (Burič & Konrádová, 2020; Burič &

Šrol, 2020). Thus, perhaps reasoning and meta-reasoning depend on two separate processes: a reasoner's disposition to dissociate from System 1 and engage System 2 reasoning after detecting errors in their intuition (termed *smart deliberation*), or a reasoner's disposition to use correct intuition (termed *smart intuition*; Raoelison et al., 2020).

Hence, smart intuition may more strongly relate to rapid anagram solving rather than meta-reasoning intuitions about problem solvability. Solutions to insight problems can occur rapidly with little awareness of how the answer was generated, which can happen even before a reasoner feels they need to restructure the problem (Bowden & Jung-Beeman, 2003; Metcalfe, 1986; Metcalfe & Wiebe, 1987; Novick & Sherman, 2008; Weisberg, 1992). Typically, those who are more skilled at the given insight problem tend to experience these rapid, sudden solutions (Novick & Sherman, 2003), which can arise when the reasoner unconsciously activates semantic representations of possible solutions (Bowers et al., 1990). Given that cognitive reflection ability is related to better insight reasoning (Patel et al., 2019), smart intuitive responses on the CRT might have been more strongly related to rapid solving during the JOS, rather than intuitions about solvability.

Thus, we acknowledge that our ability to detect a relationship between cognitive reflection and S JOS discrimination, and with JOS predictiveness overall, may have been influenced by our use of the single-response CRT. Researchers have recently proposed a two-response CRT paradigm, in which reasoners are given two opportunities to provide solutions to each CRT item: an initial response under time pressure, and a second response without time pressure (Burič & Šrol, 2020; Strudwicke et al., 2022). The two-response CRT is intended to distinguish individuals who respond intuitively and correctly, those who decouple from their erroneous intuitions, and those who do neither. Interestingly, it appears that the majority of correct answers on the CRT are intuited immediately, rather than after error detection and correction (Bago & De Neys, 2019; Burič & Konrádová, 2020). If smart

intuition is related to rapid insight solving, then perhaps our sample was over-represented with smart-intuitive thinkers who solved the anagrams during the JOS process, thus limiting the predictive power of the CRT. Future studies measuring whether cognitive reflection is related to JOS discrimination and predictiveness may benefit from separating smart deliberative responses from smart intuitive responses using the two-response CRT, and testing whether they differ in their association with initial meta-reasoning. Another future direction could assess whether smart deliberative versus smart intuitive cognitive styles are associated with JOS discrimination and predictiveness for non-insight problems (where rapid solving is less likely). Such a study would also allow us to delve into the differential cognitive processes used for making JOSs, and how people regulate their solving efforts based on their JOSs.

We also found that cognitive reflection was predictive of the proportion of anagrams solved, regardless of JOS. For reflective thinkers who use smart deliberation, the disposition to “decouple” from erroneous System 1 reasoning might relate to the tendency to know when to “restructure” an insight reasoning problem, leading to more successful problem-solving, whereas reflective thinkers who use smart intuition might have already intuited the response at the JOS stage. Future studies could implement the two-response CRT to assess whether smart intuition is indeed related to anagram solving during the JOS and whether smart deliberation is perhaps related to insight problem restructuring.

Inclusion of Longer-Duration Anagrams Increased Solving for Reflective Thinkers

Our study also explored whether the inclusion of longer-duration anagrams moderated the relationship between cognitive reflection and JOS discrimination and predictiveness. We expected that their inclusion would reduce how strongly cognitive reflection was associated with JOS discrimination and predictiveness, by allowing less-reflective reasoners an opportunity to generate a better sense of anagram solvability. However, this was not the case.

Training with longer-duration anagrams chiefly bolstered the relationship between cognitive reflection and AS JOS discrimination. Greater cognitive reflection was associated with more anagrams being reported as ‘already solved’ during the JOS task, particularly when longer-duration anagrams were included in the task. Training also strengthened the relationship between cognitive reflection and the overall proportion of anagrams solved in Experiment 2, yet training had no bearing on JOS predictiveness.

Related to an earlier point, it remains possible that our ability to detect whether training was associated with the relationship between cognitive reflection and JOSs was undermined by our use of a single-response CRT. Perhaps a stronger association between cognitive reflection and training would be detected if we could parse out reflective thinkers who use smart deliberation from those who use smart intuition. Individuals with a greater disposition to reflect on their intuition might be more sensitive to unreliable cues regarding a problem’s solvability when durations are shorter, and participants without this disposition might have their intuition accuracy bootstrapped via trials with longer-duration anagrams. We recommend that future research investigating meta-reasoning processes implement a two-response CRT to enable comparisons of smart intuitive responses versus slow deliberative responses.

Surprisingly, training moderated the relationship between cognitive reflection and AS JOS discrimination in Experiment 2 but not in Experiment 3, even though the JOS phase was identical in both. Although we cannot pinpoint the reason for this discrepancy, some cross-experiment differences between our samples may be relevant. The variance in AS JOS discrimination was lower in Experiment 3, which may have minimised how much the CRT moderated the effect of training on JOS discrimination. In addition, the Experiment 2 data were collected from a US sample between April and September 2020, when social distancing requirements led many to the online labour market. Although our MTurk criteria required an

approval rating of 95-100% and completion of at least 100 Human Intelligence Tasks (HITs), the influx of new workers may have led to more variance in performance that enabled these relationships to be detected (Arechar & Rand, 2021; Lee & Hoffman, 2020).

Finally, we found that cognitive reflection was a strong predictor of the proportion solved in the training group, but only in Experiment 2 when participants attempted to provide a solution to each anagram (cf. in Experiment 3, where unsolvable anagrams were also tested). Interestingly, Burton et al. (2022) found that solving rates in the solving phase were not greater for longer-duration anagrams. In the present study, this effect emerged only when cognitive reflection was a factor, and only when the solving phase was not self-regulated. Burton et al. (2022) noted that the high overall solving rates in Experiment 2 likely occurred because participants knew each anagram was solvable—therefore they may have exerted solving effort on each trial regardless of anagram duration in the JOS phase. Because more-reflective participants were better anagram solvers overall, this effort may have seeded greater solving success relative to less-reflective participants. Moreover, given evidence that cognitive reflection is related to greater working memory capacity (Toplak et al., 2011, 2014), perhaps providing more time to evaluate solvability during the JOS led to greater maintenance of possible solution representations, which transferred to the solving phase (Barrouillet et al., 2007). Whether cognitive reflection is related to working memory for anagram solving is thus another potential area for research exploration.

Conclusion

Our study found that cognitive reflection was associated with meta-reasoning judgements about solvability, but only by increasing the proportion of problems solved during the JOS task. As such, greater cognitive reflection was not found to be related to *intuition* about problem solvability, per se. Our study provides a provocative demonstration that cognitive reflection can be related to actual problem solving, but not to initial meta-

reasoning judgements about problem solvability. More research is needed to establish whether cognitive reflection might be shown to predict JOS discrimination and predictiveness when problems designed to mislead JOSs are used, and when anagram durations are short enough to minimise solving during the JOS task. Finally, we recommend that future studies use a two-response version of the CRT to measure whether smart intuitive versus slow deliberative responses differ in their JOS discrimination and subsequent problem-solving success.

Chapter 4: Unpacking the Relationship Between Initial Judgements of Solvability and Problem Solving: Interleaving Impacts Meta-Reasoning

Author contributions: GEB and I conceptualised the study design. I programmed the experiment and collected the data, cleaned the data for analysis, and performed the data analyses and interpretation. I drafted the manuscript and GEB and PW provided critical revisions. GEB approved the final version of the manuscript for submission.

Abstract

Judging whether a problem is solvable is a key metacognitive step in problem solving. Some studies have shown that Judgements of Solvability (JOSs) can discriminate solvable from unsolvable problems and can predict problem-solving success. Here, we examined the influence of interleaved vs. blocked designs on JOS discrimination and predictiveness. Participants made JOSs for briefly presented anagrams that were either solvable or unsolvable, by judging them as ‘solvable’, ‘not solvable’, or ‘already solved’. Solving attempts either followed each JOS (interleaved design) or occurred after all JOSs were made (blocked design). JOSs were more accurate and predictive of solving outcomes when interleaved with solving attempts than when blocked. Whether anagrams were presented for 2 s or 4 s during the JOS task influenced the rate of ‘already solved’ JOSs, but did not moderate the effects of design on JOS accuracy or predictiveness. Thus, interleaving JOSs and solving attempts can bootstrap metacognitive accuracy, effort regulation, and problem-solving success.

Introduction

The cognitive processes that underlie reasoning have been extensively researched, but the monitoring and control processes that facilitate reasoning—*meta-reasoning*—have only recently become a research focus. Awareness of our reasoning processes can inform us about whether our reasoning strategy is likely to be successful (known as *monitoring*), which in turn can inform decisions to change tact, seek help, or continue using the reasoning strategy (known as *control*; Ackerman & Thompson, 2017; Evans & Fisher, 2011; Koriat & Goldsmith, 1996; Koriat et al., 2006). Meta-reasoning processes can bolster successful reasoning outcomes and reduce incorrect reasoning outcomes (Gangemi et al., 2015). Thus, in some situations one's ability to meta-reason may be as important as one's ability to reason.

Meta-reasoning occurs in stages, beginning with a brief assessment of problem solvability, dubbed a *Judgement of Solvability* (JOS; Ackerman & Thompson, 2017). Reasoners make an initial JOS to decide if a problem is solvable, and if so, whether to make a problem-solving attempt (Payne & Duggan, 2011; Toplak et al., 2014). Because JOS research is in the early stages, studies have used different designs for measuring whether JOSs can accurately distinguish solvable from unsolvable problems (which we term *JOS discrimination*), and whether these JOSs are predictive of later solving success (which we term *JOS predictiveness*). Some studies have interleaved JOSs and solving attempts, such that on each trial, participants make a JOS and then are immediately prompted to solve the problem (e.g., Balas et al., 2011; Topolinski & Strack, 2009a; Valerjev & Dujmović, 2020). Others have used a blocked design in which JOSs and solving are separated into two phases (Ackerman & Beller, 2017; Burton et al., 2022; Lauterman & Ackerman, 2019). For example, in Burton et al., participants first completed a *JOS phase*, in which they made JOSs for a set of reasoning problems, to capture participants' intuitive judgements. Participants

then completed a *solving phase*, where they then attempted to solve the problems from the JOS phase, to measure how well solving outcomes were predicted by JOSs.

Studies that interleaved JOSs and solving have tended to report that JOSs are discriminating and predictive (e.g., Balas et al., 2011; Bolte & Goschke, 2005; Markovits et al., 2015; Novick & Sherman, 2003; Siedlecka et al., 2016; Topolinski & Strack, 2009a; Undorf & Zander, 2017). On the other hand, studies using a blocked design have found that the ability of JOSs to discriminate and predict solving outcomes and effort regulation is limited, such that JOSs are sometimes discriminating but not predictive of solving outcomes (Burton et al., 2022), are discriminating and predictive but only under certain conditions (e.g., Burton et al., 2022; Lauterman & Ackerman, 2019), or are neither discriminating nor predictive (Ackerman & Beller, 2017).

Thus, design can influence JOS discrimination and effort regulation, sponsoring different conclusions regarding how well JOSs predict solving performance. Our study sought to clarify the extent to which interleaved versus blocked designs impact meta-reasoning, specifically JOS discrimination and JOS predictiveness. We also explored whether JOS discrimination and predictiveness in each study design is affected by how much time participants are given to develop their intuitions about a problem's solvability (i.e., problem presentation duration). To these ends, the current study used elements of Burton et al.'s (2022) in-depth exploration of JOS discrimination and predictiveness, as discussed next.

JOS Measurement in a Blocked Design

Using anagrams as problem-solving stimuli, and a blocked design, Burton et al. (2022) established that JOSs were discriminating. However, JOSs were far less discriminating after excluding problems that were reportedly solved during the JOS task (i.e., spontaneous solutions; Novick & Sherman, 2003; Topolinski et al., 2016). For JOS predictiveness, anagrams given an 'already solved' (*AS*) JOS were more likely to lead to

problem solving successes, and anagrams given ‘not solvable’ (*NS*) JOSs were more likely to lead to problem solving failures. Interestingly, even though ‘solvable’ (*S*) JOSs discriminated solvable from unsolvable anagrams, they were not predictive of later problem solving. That is, in the solving phase, anagrams that received *S* JOSs were not more common among solved anagrams than among not-solved anagrams, and they also did not lead to more solving than for anagrams that received *NS* JOSs. These outcomes suggest that solutions arising during the JOS task may have exaggerated discrimination and predictiveness of ‘solvable’ JOSs in studies that did not measure and parse out spontaneous solving.

Burton et al. (2022) also established that JOS discrimination and predictiveness were influenced by how much time participants were given to develop their intuitions about a problem’s solvability. In their blocked design, *S* JOSs were more discriminating in blocks with longer-duration anagrams in the JOS phase. With longer-duration anagrams, *AS* JOSs were more frequent, because reasoners likely moved away from simply making a JOS to attempting to solve the anagram. However, at the longer durations, *S* JOSs would still have captured intuition when solving efforts did not produce solutions, and at the shorter durations, they would capture intuition when solving efforts had not yet occurred. Thus, as long as the researcher can identify and separate already-solved problems, use of longer duration problems can accurately capture solvers’ intuitions, while also gauging when initial JOSs are likely to transition into solving attempts.

The Influence of Design on JOS Discrimination and Predictiveness

We expected that use of an interleaved versus blocked design would impact meta-reasoners’ intuitions. In an interleaved design, judging a problem to be solvable on a JOS trial, and then solving it during the following solving trial, provides the meta-reasoner with informative feedback (i.e., that their intuitive JOS was correct). Given that feedback can improve the accuracy of intuitive decisions (Glöckner & Witteman, 2010), participants might

use this feedback to inform their intuitions on future JOS trials, thereby bootstrapping their JOS discrimination and predictiveness. In an interleaved design, participants also incrementally gain experience making JOSs and solving problems. This accumulation of solving experience can also improve problem-solving ability (Novick & Sherman, 2008; Shynkaruk & Thompson, 2006), which in turn can improve the ability to accurately judge problem solvability (Novick & Coté, 1992), thereby generating more discriminating and predictive JOSs.

Regardless of design, researchers typically pick a problem duration that limits spontaneous solving but provides enough time for participants to potentially generate accurate intuitions (e.g., Ackerman & Beller, 2017; Balas et al., 2011; Bolte & Goschke, 2005; Lauterman & Ackerman, 2019; Siedlecka et al., 2016; Topolinski et al., 2016; Valerjev & Dujmović, 2020). Studies measuring JOSs usually present insight problems, which are short, verbal problems that are solved suddenly with little understanding of how the answer was produced (Bowden & Jung-Beeman, 2003; Metcalfe, 1986; Metcalfe & Wiebe, 1987; Novick & Sherman, 2008; Weisberg, 1992). Although insight problems are useful for measuring JOSs (see Burton et al., 2022), there is always a risk that participants might solve some of them during the JOS task (Novick & Sherman, 2003). Thus, spontaneous solving during the JOS trials can impact the measurement of intuition accuracy even in a blocked design (see Burton et al., 2022).

Even though both interleaved and blocked designs may bootstrap JOS intuitions with how they influence solving, a blocked design has the potential virtue of capturing more naïve intuitive judgements compared to an interleaved design. In a blocked design, shorter problem durations limit the occurrence of solutions found during the JOS, which limits how much feedback reasoners receive about their JOSs (compared to an interleaved design, in which solving attempts always occur right after the JOS). However, the blocked design may

therefore lead to lower estimates of the discrimination and predictiveness of JOSs by not providing participants with as much opportunity to regulate and improve their JOS discrimination and solving success.

Another way that design may influence JOS predictiveness is through the availability of memory for the JOS. Memory for JOSs will be more available in the interleaved design than in the blocked design, which may prompt reasoners to put more effort into their solving attempt after making an S JOS than after making a NS JOS. Greater effort expenditure is likely to seed greater problem-solving success (Pennycook et al., 2015a), thus S JOSs will be more likely to lead to more solving success than in a blocked design. In kind, problems given a NS JOS should lead to less effort for that problem in the interleaved design, given that people are reluctant to invest effort in problem solving when likelihood of success is low (Ackerman, 2019; De Neys et al., 2013; Payne & Duggan, 2011).

Participants can also regulate their efforts based on the JOS in a blocked design, of course, but memory for the JOS they gave to each problem will be less accessible than in the interleaved design due to greater interference from successive JOSs (i.e., interference) and longer time between making the JOS in the JOS trials and solving the problem during the solving trials (i.e., delay; Berman et al., 2009). Thus, JOSs may have less impact on solving effort regulation in a blocked design than in an interleaved design. On the other hand, there is some evidence that intuitions about problem solvability can be stable across time (Stagnaro et al., 2018), thus it is also possible that intuitions from the JOS trials in a blocked design may match intuitions that arise later during the solving phase. If so, JOSs in a blocked design should still predict later solving outcomes and effort regulation, just to a lesser extent than in an interleaved design.

Effects of Problem Duration in Blocked versus Interleaved Designs

Problem duration can also impact JOSs. In a blocked design, Burton et al. (2022) showed that S JOSs are far more discriminating when participants had more time to make their JOSs (though surprisingly JOSs were not more predictive of problem-solving outcomes). They also showed that solutions arising during the JOS task (i.e., AS JOSs) were more likely when participants had more time to make JOSs. Here, we sought to examine the influence of study design on JOS discrimination and predictiveness when 2 s vs. 4 s anagram durations were used (between groups).

As discussed earlier, the availability of the JOS in memory should be greater in an interleaved design. In turn, this will lead solvers to align their solving efforts with their JOS (i.e., more effort for a problem given an S JOS, less effort for a problem given an NS JOS), and will ultimately lead to solving successes/failures that correspond with the JOS. However, during the solving trials, participants in a blocked design should be more likely to remember their JOS in the 4 s than 2 s duration (Tversky & Sherman, 1975), thus sponsoring higher JOS predictiveness in the 4 s group. Therefore, although we expected JOSs to be more predictive in the interleaved design overall, we expected that differences in JOS predictiveness across the 2 s and 4 s durations would be greater in the blocked design.

Method

The experiment was pre-registered on Open Science Framework (OSF) at <https://osf.io/e4zc3> (embargoed until March 29th, 2023). The data for this study are available in Open Science Framework at <https://doi.org/10.17605/OSF.IO/JD5S9>.

Participants

Participants ($N = 255$) were recruited through Amazon's Mechanical Turk (MTurk) via TurkPrime (Litman et al., 2017) and each received USD \$2. Our MTurk criteria required a Human Intelligence Task (HIT) approval rating of 95-100% and completion of at least 100

HITs. We excluded 60 participants who met one or more pre-registered exclusion criterion (correctly solved less than 10% of anagrams, did not complete the study, failed an attention check, did not provide a JOS on 25% or more of JOS trials, more than 2 SD outside the mean study completion time). The final sample sizes were 51 and 49 for the blocked and interleaved conditions for the 2 s group (66 female, 31 male, 3 non-binary; mean age = 38.7, $SD = 11.4$), respectively, and 44 and 51 for the 4 s group (63 female, 30 male, 2 non-binary; mean age = 38.3, $SD = 9.6$).

Stimuli

We selected 20 solvable and 20 unsolvable 5-letter anagrams from the 40 of each type used in Burton et al. (2022). We excluded solvable anagrams with the highest and lowest solving rates from a pilot study (see Burton et al.); the 20 we selected had a mean solving rate of 75.0% (range 56.5-87.0%). We created 2 blocks of 10 solvable and 10 unsolvable anagrams, roughly equated for solving rate, and we also created 2 orders of the 20 anagrams within each block. Solvable and unsolvable anagrams were then randomly mixed with the constraint that there were no more than 3 of a given anagram type in a row. Thus, 4 different lists of the 40 anagrams were created, and their assignment was counterbalanced across participants.

Design

The experiment used a 2(duration: 2 s, 4 s) by 2(design: blocked, interleaved) between-subjects factorial design. Data for the 2 s and 4 s anagram duration groups were collected in turn (back-to-back), and within each duration group the assignment to the blocked versus interleaved design was randomized.

Procedure

The experiment was conducted online using Qualtrics software (Qualtrics, 2019). Participants were instructed that on each trial they would see a sequence of letters (i.e., an

anagram), some of which could be rearranged to spell a word (e.g., DSTMI - MIDST) and hence were 'solvable', and others of which did not have a solution word (e.g., ZEREB) and hence were "unsolvable". They were also told that the anagrams would be presented for only 2 s or 4 s (as per their group assignment).

In the blocked groups, participants made a JOS to each of the 40 anagrams (JOS phase) then attempted to solve each of the 40 anagrams in the same order (solving phase). In the interleaved groups, participants made a JOS to an anagram then were immediately presented with it again with a text box underneath and a prompt for them to attempt to solve it. For the JOS task, participants were told that their task was to make one of three solvability judgements for each anagram in the allotted time: "YES it is solvable" (S JOS), "NO it is not solvable" (NS JOS), or "I have already solved it" (AS JOS). On JOS trials the anagram was presented for the assigned duration (2 s or 4 s). Once the anagram disappeared, the 3 JOS options appeared as response boxes, and participants had 3 s to click on a response. If they failed to make their JOS within 3 s, a message appeared asking them to respond within 3 s. This message remained on the screen for 4 s to discourage participants from continuing to try to solve anagrams after they disappeared.

For the solving task, participants were told that they had as much time to try to solve each anagram as they wished. They were instructed to either type in a solution, type the letter "P" to pass if they believed the anagram was solvable but were unable to solve it, or to type the letter "N" for 'not solvable' if they believed the anagram was unsolvable. All participants first completed 5 practice JOS trials (3 solvable, 2 unsolvable), followed by 5 solving trials using the same anagrams.

Results

The JOS trials assessed whether JOSs accurately distinguished solvable from unsolvable problems (i.e., JOS discrimination), and whether anagram duration and design

influence JOS discrimination. The solving trials assessed how well JOSs predict successful solving and ‘not solvable’ responses, and response times for these outcomes. The measures we used and the analyses we report match our previous study (Burton et al., 2022). However, here we also examined the relationship between JOS discrimination and JOS predictiveness to see whether better JOS discrimination was associated with better prediction of solving outcomes. We used .05 as our alpha level and η^2_p as our measure of effect size. We also analysed the extent to which JOSs predict self-regulated solving time in each design, but for brevity these are reported in our Supplementary Materials.

JOS Trials

JOS discrimination was assessed by measuring whether hit rates (i.e., correctly judging solvable anagrams as solvable) exceeded false alarm rates (i.e., incorrectly judging unsolvable anagrams as solvable). Hits and false alarms were converted to proportions by dividing them by the total number of JOS trials (excluding trials where participants failed to enter a response within the 3 s time limit after anagram presentation). These proportions were calculated separately for AS and S JOSs (see Figure 4.1).

AS and S JOS discrimination were each analysed using a 2(discrimination: hits, false alarms) \times 2(duration: 2 s, 4 s) \times 2(design: blocked, interleaved) mixed-factor ANOVA. Table 4.1 provides the complete ANOVA results for each JOS. The four key effects reviewed below for each JOS are: (1) whether JOSs were discriminating (main effect of discrimination; i.e., hits > false alarms), (2) whether JOS discrimination was greater for 4 s than 2 s anagrams (discrimination \times duration interaction), (3) whether JOS discrimination was greater in the interleaved than blocked design (discrimination \times design interaction), and (4) whether anagram duration moderated the latter interaction (discrimination \times design \times duration).

Figure 4.1

JOS Trials: Mean Proportions of Hits and False Alarms (Bars show 95% CI of each mean)

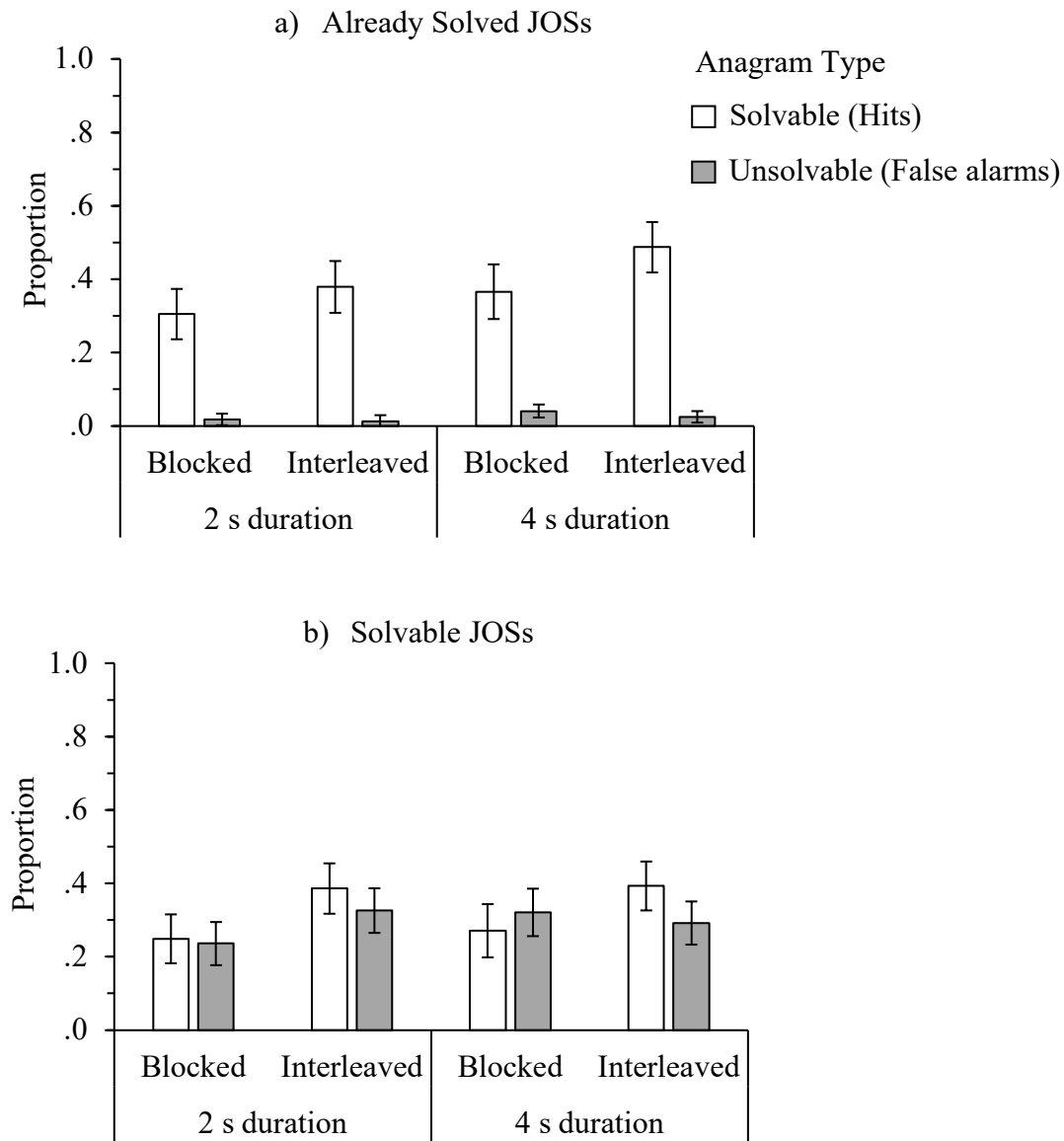


Table 4.1*JOS Trial Discrimination: ANOVA Results by JOS*

<i>JOS/Effect</i>	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
Already Solved JOSs					
Discrimination	1, 190	12.51	413.29	< .001	.69
Duration	1, 190	0.25	7.05	.01	.04
Design	1, 190	0.18	5.03	.03	.03
Discrimination \times Duration	1, 190	0.11	3.53	.06	.02
Discrimination \times Design	1, 190	0.29	9.44	.002	.05
Duration \times Design	1, 190	0.01	0.23	.64	.00
Discrimination \times Duration \times Design	1, 190	0.02	0.66	.42	.00
Solvable JOSs					
Discrimination	1, 191	0.09	4.05	.046	.02
Duration	1, 191	0.04	0.47	.50	.00
Design	1, 191	0.62	7.49	.01	.04
Discrimination \times Duration	1, 191	0.00	0.11	.74	.00
Discrimination \times Design	1, 191	0.24	10.35	.002	.05
Duration \times Design	1, 191	0.11	1.29	.26	.01
Discrimination \times Duration \times Design	1, 191	0.07	2.86	.09	.02

For AS JOSs, hits ($M = .38$, $SD = .26$) significantly exceeded false alarms ($M = .02$, $SD = .06$), thus AS JOSs accurately discriminated solvable from unsolvable anagrams.

Discrimination was more accurate with 4 s than 2 s anagrams, but was significant for both durations ($ps < .001$). As anticipated, we found that AS JOS discrimination was more accurate in the interleaved than blocked design (but was significant for both designs; $ps <$

.001). The three-way interaction was not significant, thus use of longer-duration anagrams did not increase the discrimination advantage in the interleaved design.

S JOSs also accurately discriminated solvable from unsolvable anagrams, although the difference between hits ($M = .33$, $SD = .25$) and false alarms ($M = .29$, $SD = .22$) was very modest and only just reached significance. Unlike for AS JOSs, S JOS discrimination was not more accurate for 4 s than 2 s anagram durations. As for AS JOSs, S JOS discrimination was more accurate in the interleaved than blocked design. In fact, S JOS discrimination was significant in the interleaved design ($p < .001$) but not in the blocked design ($p = .40$). The three-way interaction was again not significant.

In sum, interleaving JOSs with solving attempts increased the accuracy of both AS and S JOSs—and here S JOSs were only discriminating in the interleaved design. In contrast, Burton et al. (2022) found significant S JOS discrimination in a blocked design at both 4 s and 2 s anagram durations. To evaluate the strength of our evidence for a null effect, we calculated a Bayes factor (as specified by Van Doorn et al., 2007) using JASP with default priors (JASP Team, 2022, Version 0.16.2). Unlike conventional null hypothesis significance testing, the Bayes factor indicates the odds ratio of the alternative hypothesis relative to the null hypothesis; factors greater than 1 favour the alternative hypothesis, whereas factors less than 1 favour the null hypothesis. The Bayes factor from a repeated-measures ANOVA examining discrimination in the blocked design (across duration) favoured the null hypothesis, $BF_{10} = 0.21$. Our Discussion notes between-study differences that may explain why S JOSs were not discriminating in the blocked design.

Solving Trials

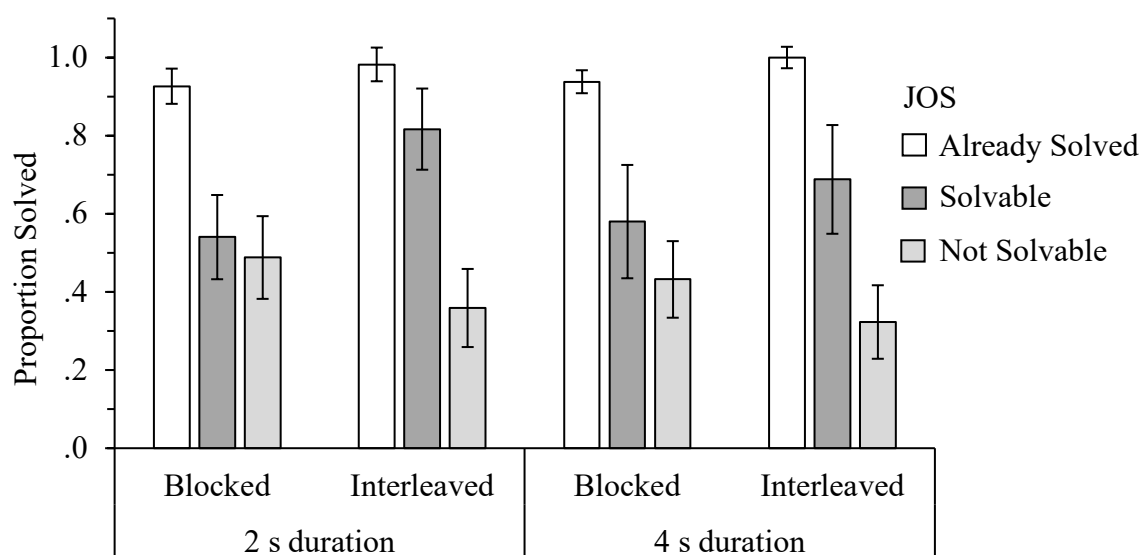
To assess how well JOSs predict problem-solving success in each design, we used two measures reported in Burton et al. (2022): *proportion solved* and *solved versus not solved outcomes*.

Proportion Solved

For each JOS, we calculated the *proportion solved* as the number of solved anagrams that had received that JOS divided by the total number of anagrams that had received that JOS. For example, if a participant assigned S JOSs to 10 anagrams and solved 6 of them, then their proportion solved would be .6 for S JOSs. The proportion solved for each JOS was independent and could range from 0 to 1 (see Figure 4.2 for the means).

Figure 4.2

Solving Trials: Mean Proportion Solved (Bars show 95% CI of each mean)



The mean proportion of anagrams solved as a function of JOS (AS vs. S vs. NS) served as the repeated-measures factor in a mixed-factor ANOVA in which design and duration were the between-group factors (see Table 4.2). Only two effects were significant: The main effect of JOS and its interaction with design. Focusing on the latter, although the solving rate for anagrams given AS JOSs was near ceiling, the proportion solved was nonetheless greater in the interleaved than blocked design ($M = .99$, $SD = .05$ vs. $M = .93$, $SD = .13$; $p = .002$). Critically, the same pattern occurred for S JOSs ($M = .75$, $SD = .31$ vs. $M = .56$, $SD = .41$; $p < .001$). Conversely, the solving rate for anagrams given NS JOSs was lower in the interleaved than blocked design ($M = .31$, $SD = .33$ vs. $M = .46$, $SD = .28$; $p = .01$).

These results suggest that the interleaved group was more tenacious in their attempts to solve anagrams receiving AS and S JOSs, but less tenacious for anagrams receiving NS JOSs.

Thus, the provided JOS appears to influence subsequent solving effort regulation to a greater extent in an interleaved design than in a blocked design.

Table 4.2

Solving Trials: Proportion Solved ANOVA Results

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	1.94, 256.03	11.23	178.81	< .001	.58
Duration	1, 131	0.06	0.56	.46	.00
Design	1, 131	0.22	2.18	.14	.02
JOS × Duration	1.94, 256.03	0.04	0.58	.56	.00
JOS × Design	1.94, 256.03	0.80	12.90	< .001	.09
Duration × Design	1, 131	0.07	0.68	.41	.01
JOS × Duration × Design	1.94, 256.03	0.09	1.40	.25	.01

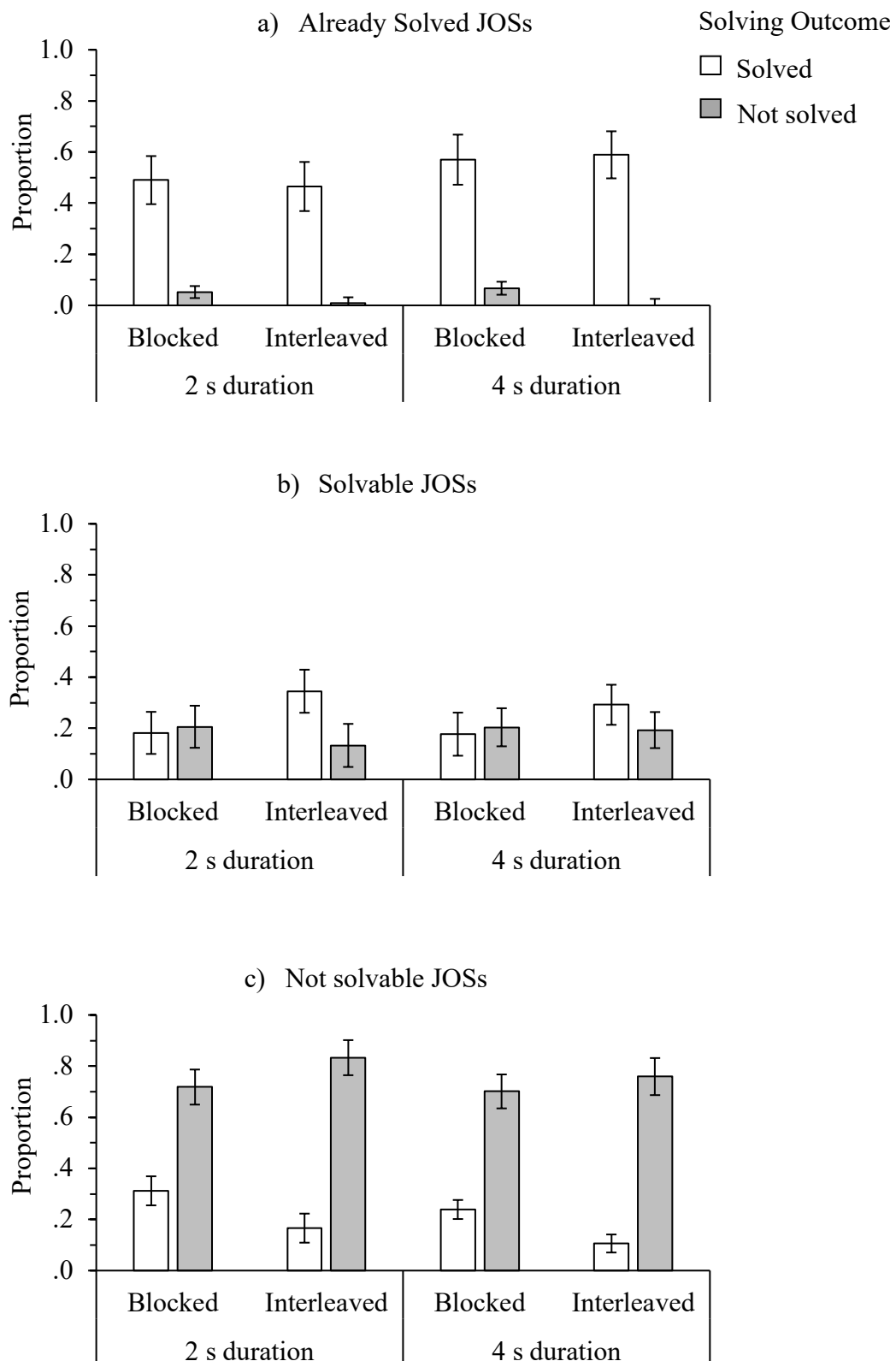
Solved vs. Not Solved Outcomes

Our second measure of how well JOSs predict problem solving compared, for each JOS, the proportion of solved anagrams that had received that JOS to the proportion of not-solved anagrams that had received that JOS (see Burton et al., 2022). For example, if 8 anagrams were solved, and of those anagrams 6 had received S JOSs, the proportion solved in the solving phase for S JOSs would be .75. If 12 anagrams were not solved, and 3 of those not-solved anagrams had received S JOSs, the proportion not-solved on solving trials for S JOSs would be .25. For each JOS, these two proportions are independent and could range from 0 to 1 (see Figure 4.3 for the means).

Figure 4.3

Solving Trials: Proportion of Solved Versus Not Solved Outcomes for Solvable Anagrams

(Bars show 95% CI of each mean)



For each JOS, the mean proportion of solved anagrams that received that JOS versus the proportion of not-solved anagrams that received that JOS (i.e., solved versus not solved outcomes) served as the repeated-measures factor in a mixed-factor ANOVA in which design and duration were the between-group factors (see Table 4.3). Here we focus on whether there were a higher proportion of solved outcomes than not-solved outcomes for a given JOS (i.e., the main effect of outcome), and whether this pattern interacted with duration and/or design.

AS JOSs were more frequent among solved anagrams than among not-solved anagrams ($M = .50, SD = .33$ vs. $M = .03, SD = .09$). This difference interacted only with duration, being larger for the 4 s duration ($M = .55, SD = .32$, vs. $M = .03, SD = .09$) than the 2 s duration ($M = .46, SD = .33$, vs. $M = .03, SD = .08$), though both were significant, $ps < .001$. S JOSs were also more frequent among solved anagrams than among not-solved anagrams ($M = .27, SD = .29$ vs. $M = .18, SD = .23$). This difference interacted only with design. Importantly, it was robust in the interleaved group ($M = .34, SD = .32$ vs. $M = .16, SD = .24$; $p < .001$) but completely absent in the blocked group ($M = .20, SD = .23$, vs. $M = .20, SD = .22$; $p = .44$). Finally, NS JOSs were less frequent among solved anagrams than among not-solved anagrams ($M = .21, SD = .18$ vs. $M = .75, SD = .25$). This difference interacted only with design, being larger in the interleaved group ($M = .14, SD = .14$, vs. $M = .80, SD = .25$) than in the blocked group ($M = .28, SD = .19$, vs. $M = .71, SD = .24$), though both effects were significant ($ps < .001$).

This measure suggests that interleaving JOSs with solving attempts modulated effort regulation by increasing solving effort following S JOSs while decreasing solving effort following NS JOSs. AS JOSs were far more frequent among solved than among not-solved anagrams in both designs, confirming that participants used AS JOSs appropriately (see also Burton et al., 2022).

Table 4.3*Solving Trials: Solved vs. Not Solved Outcomes ANOVA Results by JOS*

<i>JOS/Effect</i>	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
Already Solved JOSs					
Outcome	1, 173	21.74	425.63	< .001	.71
Duration	1, 173	0.25	4.14	.04	.02
Design	1, 173	0.08	1.27	.26	.01
Outcome × Duration	1, 173	0.22	4.23	.04	.02
Outcome × Design	1, 173	0.06	1.21	.27	.01
Design × Duration	1, 173	0.00	0.04	.84	.00
Outcome × Design × Duration	1, 173	0.03	0.48	.49	.00
Solvable JOSs					
Outcome	1, 173	0.37	7.64	.01	.04
Duration	1, 173	0.00	0.00	1.00	.00
Design	1, 173	0.21	2.70	.10	.02
Outcome × Duration	1, 173	0.07	1.51	.22	.01
Outcome × Design	1, 173	0.73	14.89	< .001	.08
Design × Duration	1, 173	0.00	0.01	.92	.00
Outcome × Design × Duration	1, 173	0.07	1.35	.25	.01
Not Solvable JOSs					
Outcome	1, 173	26.39	621.23	< .001	.78
Duration	1, 173	0.29	6.40	.01	.04
Design	1, 173	0.06	1.40	.24	.01
Outcome × Duration	1, 173	0.10	0.22	.64	.00
Outcome × Design	1, 173	1.12	26.36	< .001	.13
Design × Duration	1, 173	0.01	0.23	.63	.00
Outcome × Design × Duration	1, 173	0.03	0.63	.43	.00

‘Not solvable’ Responses to Solvable vs. Unsolvable Anagrams.

The *proportion of not-solvable responses* was calculated as the number not-solvable responses on solving trials that had received a given JOS divided by the total number of anagrams that had received that JOS (Burton et al., 2022). For example, if a participant gave

a not-solvable response in the solving phase for 5 out of 10 anagrams to which they had given an NS JOS in the JOS phase, their proportion of not-solvable responses in the solving phase for NS JOSs would be .5. Due to the rarity of AS and S JOSs for unsolvable anagrams, these proportions were calculated across both AS and S JOSs (AS+S; see Figure 4.4 for the means).

We analysed the mean proportion of not-solvable responses as a function of JOS and anagram type, which were repeated-measures factors in a mixed-factor ANOVA with design and duration as between-group factors (see Table 4.4). Four effects were significant: a JOS main effect (not-solvable responses were more likely following NS JOSs than AS+S JOSs), an anagram type main effect (not-solvable responses were more likely for unsolvable than solvable anagrams), the JOS \times anagram type interaction (the JOS effect was larger for solvable anagrams), and, of note, the JOS \times design interaction. This interaction was due to the JOS effect being larger in the interleaved design, although it was significant in both designs ($ps < .001$). Indeed, not-solvable responses were more likely following NS JOSs in the interleaved design ($M = .80, SD = .18$) than in the blocked design ($M = .66, SD = .20, p < .001$), but were less likely following AS+S JOSs in the interleaved design ($M = .20, SD = .21$) than in the blocked design ($M = .31, SD = .19, p < .001$). Thus, JOSs had a stronger impact on one's decision to enter a 'not solvable' response on solving trials in the interleaved than blocked design.

Figure 4.4

Solving Trials: Mean Proportion of Not-Solvable Responses (Bars show 95% CI of each mean)

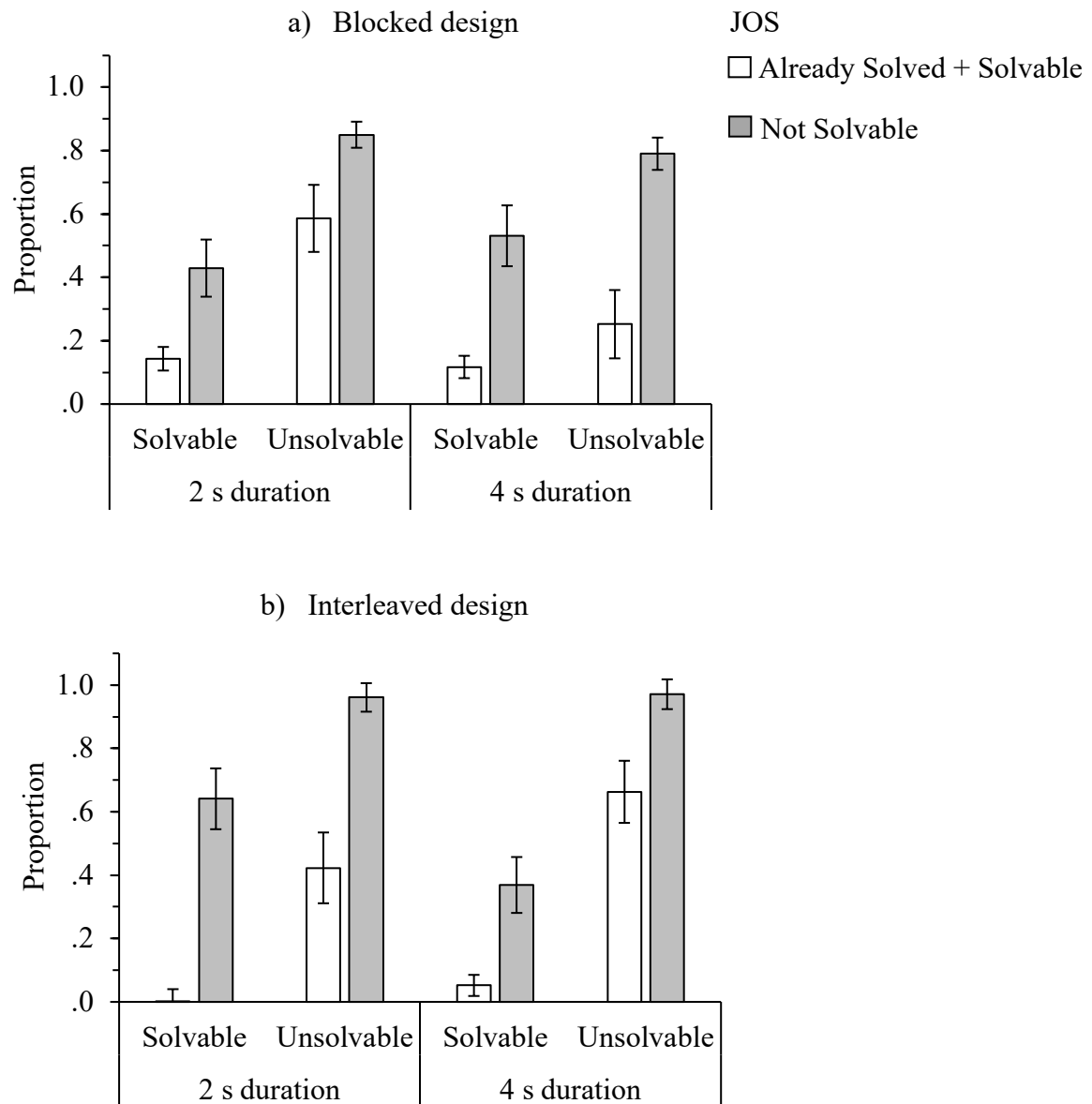


Table 4.4*Solving Trials: Proportion of Not-Solvable Responses ANOVA Results*

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	1, 163	33.66	535.41	< .001	.77
Anagram type	1, 163	21.90	304.51	< .001	.65
Duration	1, 163	0.00	0.01	.92	.00
Design	1, 163	0.03	0.41	.52	.00
JOS × Duration	1, 163	0.06	0.94	.33	.01
JOS × Design	1, 163	3.56	56.65	< .001	.26
JOS × Duration × Design	1, 163	0.02	0.27	.60	.00
JOS × Anagram type	1, 163	0.21	5.03	.03	.03
JOS × Duration × Anagram type	1, 163	0.00	0.08	.78	.00
JOS × Design × Anagram type	1, 163	0.01	0.22	.60	.00
Anagram type × Duration	1, 163	0.24	3.29	.07	.02
Anagram type × Design	1, 163	0.08	1.13	.29	.01
Duration × Design	1, 163	0.02	0.20	.66	.00
Anagram type × Duration × Design	1, 163	0.01	0.17	.69	.00
JOS × Anagram type × Duration × Design	1, 163	0.13	3.02	.08	.02

S JOS discrimination versus S JOS predictiveness

Finally, for the first time, we assessed whether greater S JOS discrimination resulted in a greater proportion solved for anagrams given S JOSs (using the *proportion solved* measure). We also assessed whether anagram duration and study design moderated this relationship. The relationship between AS JOS discrimination and predictiveness was not investigated because anagrams given AS JOSs were almost always solved (see Figure 4.2).

We chose the proportion solved measure rather than the solved vs. not-solved outcome measure because we were interested in examining whether better S JOS discrimination predicted more anagrams solved, rather than whether S JOSs were more frequent among solved outcomes than among not-solved outcomes.

We created a *S JOS discrimination* difference score by subtracting participants' mean false alarms from their mean hits for S JOSs. Thus, participants' S JOS discrimination could range from +1 (100% hits, 0% false alarms) to -1 (0% hits, 100% false alarms). The proportion solved for S JOSs was our outcome variable. We analysed S JOS discrimination scores in a 2(duration: 2 s, 4 s) by 2 (design: blocked, interleaved) between-groups ANCOVA with (mean-centred) S JOS discrimination included as a predictor covariate. Our analyses of the proportion solved measure already established a significant main effect of design, and non-significant main effect of duration, thus we do not repeat or consider the main effects of duration and design.

The ANCOVA (Table 4.5) revealed two key significant effects: a main effect of S JOS discrimination, and an interaction of design and duration. A linear regression was used to show the relationship between S JOS discrimination and S JOS proportion solved. S JOS discrimination was a significant positive predictor of the rates of solving for anagrams given S JOSs, $F(1, 163) = 16.21$, $MSE = 0.75$, $p < .001$, $R^2 = .09$, $B = 0.19$. Thus, participants who made more discriminating S JOSs were more likely to go on to solve the anagrams to which they had assigned S JOSs.

To follow-up the design \times duration interaction we used pairwise comparisons for each design. In the interleaved design, the proportion of anagrams given S JOSs that were solved was higher for the 2 s group than the 4 s group ($M = .83$, $SD = .24$ vs. $M = .69$, $SD = .35$; $p = .03$), whereas this difference was not significant in the blocked design ($M = .60$, $SD = .35$ vs. $M = .67$, $SD = .46$; $p = .32$). We comment on this unexpected result in our Discussion.

Table 4.5*Proportion Solved for S JOSs: ANCOVA with Mean-Centred S JOS Discrimination*

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
Discrimination	1, 158	2.58	26.42	< .001	.14
Duration	1, 158	0.06	0.62	.59	.00
Design	1, 158	0.68	7.01	.01	.04
Duration × Design	1, 158	0.51	5.18	.02	.03
Duration × Discrimination	1, 158	0.00	0.01	.91	.00
Design × Discrimination	1, 158	0.16	1.62	.21	.01
Duration × Design × Discrimination	1, 158	0.10	1.04	.31	.01

Discussion

When JOSs are interleaved with solving attempts, studies have typically found that JOSs discriminate between solvable and unsolvable problems and also predict later problem-solving success (e.g., Balas et al., 2011; Bolte & Goschke, 2005; Markovits et al., 2015; Novick & Sherman, 2003; Siedlecka et al., 2016; Topolinski & Strack, 2009a; Undorf & Zander, 2017). However, studies that block JOSs and solving attempts into two separate phases have not consistently found JOSs to be discriminating and/or predictive of problem-solving success (e.g., Ackerman & Beller, 2017; Burton et al., 2022; Lauterman & Ackerman, 2019). Our study established that design can influence the extent to which JOSs are discriminating and predictive. We found that intuitions about solvability were more accurate and predictive when JOSs were interleaved with solving attempts, and this pattern was stable across two anagram durations that differed in how often they yielded spontaneous solutions during the JOS trials. Below, we discuss the differences in interleaving versus

blocking on JOS discrimination and predictiveness, our duration manipulation, and design selection recommendations for future JOS research.

Interleaving Influences JOS Discrimination and Predictiveness

As anticipated, providing incremental experience solving anagrams in an interleaved design made participants more sensitive to anagram solvability. Interleaving provides participants with feedback about their intuitions when they solved (or did not solve) the anagram—feedback that they could use to bootstrap the accuracy of their intuitions on future JOS trials. Indeed, our study showed that JOSs were more discriminating and predictive in an interleaved design. Interleaving also resulted in a higher rate of AS JOSs during the JOS trials. The solving practice provided after each JOS worked to improve participants' ability to solve the anagrams during the JOS trials. Thus, in addition to bootstrapping S JOSs, interleaved designs may also foster problem-solving success.

Although Burton et al. (2022) reported above-chance discrimination for S JOSs in a blocked design, our study did not replicate this result. Here, S JOSs were discriminating in the interleaved design, but not in the blocked design, at either 2 s or 4 s anagram durations. That S JOSs were discriminating in the interleaved design is somewhat reassuring, but it leaves the question of why this did not occur in the blocked design. We cannot pinpoint the reason here, but we do point out some cross-study differences. In the Burton et al. study, participants made 80 JOSs across 4 blocks (20 solvable and 20 unsolvable in each block), whereas ours made half as many. Thus, more practice making JOSs may be necessary to sponsor discriminating S JOSs. There were also some differences between the anagrams selected across the two studies. In Burton et al., the 40 anagrams selected from a pilot study had a wider range of solving rates (50-100%) compared to the present study (57-87%). Consequently, there were fewer “easy” anagrams and “hard” anagrams in the present study. Perhaps exposure to some very easy (and/or very hard) anagrams is necessary for participants

to learn how to effectively regulate their JOSs. Given that comparative difficulty of test items can influence metacognitive judgements (Arnold et al., 2017; Arnold & Prike, 2015), we recommend that future research consider the impact of intermixing easier problems with more difficult problems on JOS discrimination.

For the first time, we found evidence that S JOSs can be more frequent among solved anagrams than among not-solved anagrams. This result arose in our interleaved design, whereas it was not significant in the blocked design here or in Burton et al. (2022). We expected that the greater accessibility of the JOS in an interleaved design might influence effort regulation and subsequent problem-solving success. Although participants in the blocked design may also regulate their efforts based on their JOS, delay and interference from successive JOS trials likely render the original JOS less available in memory. Our JOS predictiveness results suggest that following a JOS trial with a solving trial may lead to more persistent solving attempts for anagrams given S JOSs, given that S JOSs yielded higher rates of solving and were also more frequent among solved anagrams than among not-solved anagrams. Our study suggests that the availability of the JOS may be an important determinant of how well JOSs predict problem-solving success. However, our study design did not allow us to confirm this possibility. Future research could directly measure participants' memory for their JOSs in both designs after each solving trial. This could be done by re-presenting each anagram after the solving trial and asking participants to report their original JOS.

Longer-Duration Anagrams Increased Solving During JOS Trials but did not Moderate the Effects of Blocking vs. Interleaving

The current study also explored whether anagram duration moderated the effects of study design on JOS discrimination and predictiveness. Overall, the 4 s duration increased the frequency of AS JOSs among solved (versus among not-solved) anagrams, but otherwise,

anagram duration moderated few of our outcomes. However, one unexpected effect of anagram duration warrants comment. In the interleaved design, we found that S JOSs were more predictive of solving when anagram durations were *shorter* (2 s rather than 4 s). Making JOSs should be more difficult for 2 s anagrams than for 4 s anagrams, thus perhaps the 2 s group exerted more effort during the JOS task. Because JOSs would be more salient during solving trials in the interleaved design, greater effort may therefore also have been applied to the solving attempt, resulting in greater success. In other words, it is possible that the 2 s duration created a “desirable difficulty” for learning in terms of JOS discrimination (Bjork, 1994; Bjork et al., 2013; Yue et al., 2013). The briefer duration may have made judging solvability more difficult, but it may have also engaged processes which generated greater effort regulation, thus leading to more solving successes for S JOSs. Of course, further research is needed to confirm this finding and establish its cause.

In general, though, anagram duration had much less influence on JOS discrimination and predictiveness in the present study than in Burton et al. (2022). This could reflect the between-group manipulation of duration here versus within-subjects in Burton et al. (where duration in the ‘training’ condition decreased across 4 blocks from 16 s, to 8 s, to 4 s, to 2 s). Burton et al.’s parametric manipulation may have increased JOS discrimination in the later blocks through practice effects (Keren, 2014), which in turn may have sponsored a larger difference in JOS discrimination between the 2 s and 4 s blocks. In addition, the within-subject manipulation of anagram duration is also more sensitive. Whether an impact of a within-subject manipulation of anagram duration moderates the effect of study design on JOS discrimination and predictiveness could also be a target for future research.

Conclusion

Our study establishes that interleaved designs impact meta-reasoning such that they produce JOSs that are more sensitive to problem solvability, and more predictive of later

problem-solving success. Importantly, whether these effects of interleaving are desirable or not will depend on the research question. Our findings do not challenge the use of blocked designs. Indeed, blocked designs likely capture more naïve intuitions about solvability (Burton et al., 2022), and thus may be better suited to the study of how problem-solving characteristics such as problem length, difficulty, or fluency bias JOSs (such as problem length, difficulty, or fluency; e.g., Balas et al., 2011; Lauterman & Ackerman, 2019; Topolinski et al., 2016; Valerjev & Dujmović, 2020). Additionally, interleaved and blocked designs may represent different metacognitive test-taking strategies for learners. Imagine an exam in which students must answer 3 of 4 problems. Here, students may make a JOS for each question, and then attempt to solve the 3 questions they feel are most solvable (akin to a blocked design). Now imagine an exam in which students must answer 50 problems. Here, students are unlikely to make a JOS to each question at the outset—rather—they will likely make a JOS and then complete the solving attempt if they feel it is solvable (akin to an interleaved design). In the same vein, there is no “gold standard” design for measuring JOSs. Rather, researchers should consider the design choice carefully in light of their research questions about meta-reasoning.

Supplementary Materials

Does JOS Predictiveness of Solving Response Times Vary Depending on Duration and Design?

To analyse whether study design and duration influenced how well JOSs predicted response times on solving trials, we calculated the mean solution response times for AS, S, and NS JOSs, as well as mean not-solvable response times for AS+S and NS JOSs. Because mean solution times were negatively skewed, a base 10 logarithm transformation was applied to the mean solution times to normally distribute the data. Thus, descriptive statistics are presented in seconds, but inferential statistics were conducted using the transformed means.

Solved Anagrams

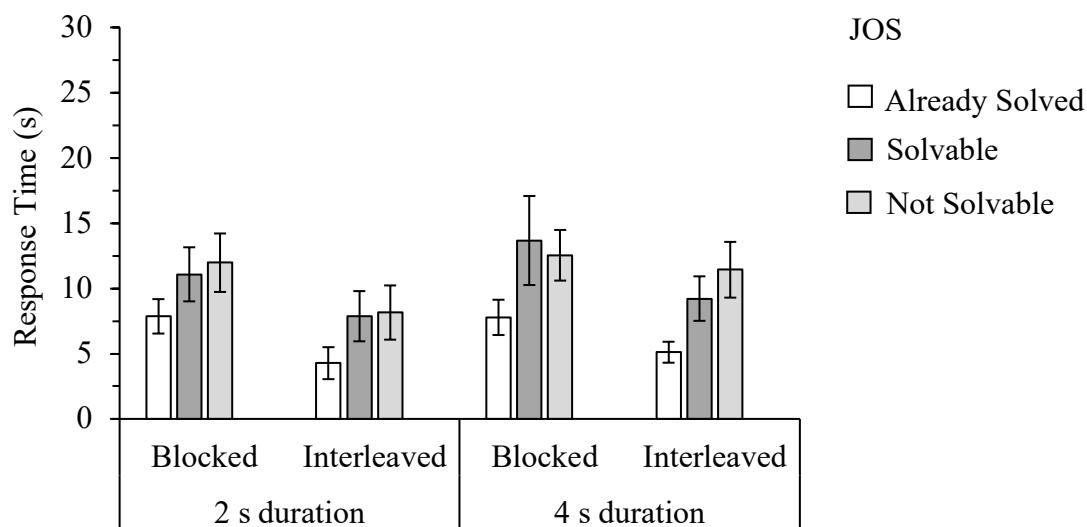
We analysed mean solving response times for each JOS in a 3(JOS) \times 2(duration) \times 2(design) mixed-factor ANOVA (see Supplementary Table 4.1). There were four significant effects: main effects of JOS, duration, and design, and an interaction between JOS and duration. Our focus was on the main effect of design, as well as the significant interaction between JOS and duration (given that it qualifies each main effect of JOS and duration). The means are provided in Supplementary Figure 4.1.

Supplementary Table 4.1*Mean Solving Time on Solving Trials: ANOVA Results*

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	2, 202	1.70	76.69	< .001	.43
Duration	1, 101	0.50	4.57	.04	.04
Design	1, 101	1.19	17.33	< .001	.15
JOS × Duration	2, 202	0.08	3.42	.04	.03
JOS × Design	2, 202	1.11	0.03	.33	.01
Duration × Design	1, 101	0.05	0.41	.52	.00
JOS × Duration × Design	2, 202	0.01	0.21	.81	.00

Supplementary Figure 4.1

Solving Trials: Mean Response Times for Solved Anagrams (Bars show 95% CI of each mean)



The main effect of design indicates that interleaving led to faster solution response times ($M = 7.56$, $SD = 3.94$) than did blocking ($M = 10.30$, $SD = 5.01$). Faster solutions for interleaved anagrams likely occurred because participants were already in solving process when the solving trial appeared. In the blocked design, the delay between JOSs and solving meant that participants probably had to begin solving again when the anagram reappeared on a solving trial, thus leading to longer solving response times.

Solutions following an AS JOSs were provided at similar speeds when anagram duration was 2 s ($M = 5.95$, $SD = 3.37$), and 4 s ($M = 6.41$, $SD = 3.80$, $p = .68$). The differences in solving response times between the 2 s and 4 s durations occurred for S JOSs and NS JOSs. Anagrams receiving S JOSs had faster solution response times at the 2 s duration ($M = 9.57$, $SD = 5.57$) than 4 s duration ($M = 11.12$, $SD = 7.75$, $p = .03$). The same pattern occurred for NS JOSs; solutions for NS JOSs were provided faster for 2 s anagram durations ($M = 10.21$, $SD = 5.82$) than 4 s anagram durations ($M = 12.03$, $SD = 6.26$, $p = .01$). Interestingly, shorter anagram durations appear to decrease solution response times for S JOSs and NS JOSs. Making JOSs for 2 s anagrams (which did not produce a clear-cut solution immediately) were likely more difficult than making JOSs for 4 s anagrams, therefore participants who had 2s anagram durations might have been expending more effort to evaluate solvability. More effort expenditure during the JOS process might have generalised to the solving trials, in which greater effort led to more rapid solutions.

Not-Solvable Response Times for Solvable vs. Unsolvable Anagrams

Not-solvable response times were analysed in a 2(JOS: AS+S, NS) \times 2(anagram type: solvable, unsolvable) \times 2(duration: 4 s, 2 s) \times 2(design: interleaved, blocked) mixed-factor ANOVA (see Supplementary Table 4.2 for the ANOVA results, and Supplementary Figure 4.2 for the means). There were three significant main effects of JOS, anagram type, and duration. A significant JOS \times duration emerged, as did a JOS \times duration \times design interaction.

We discuss the main effect of anagram type, the JOS \times duration interaction, and the JOS \times duration \times design interaction in turn.

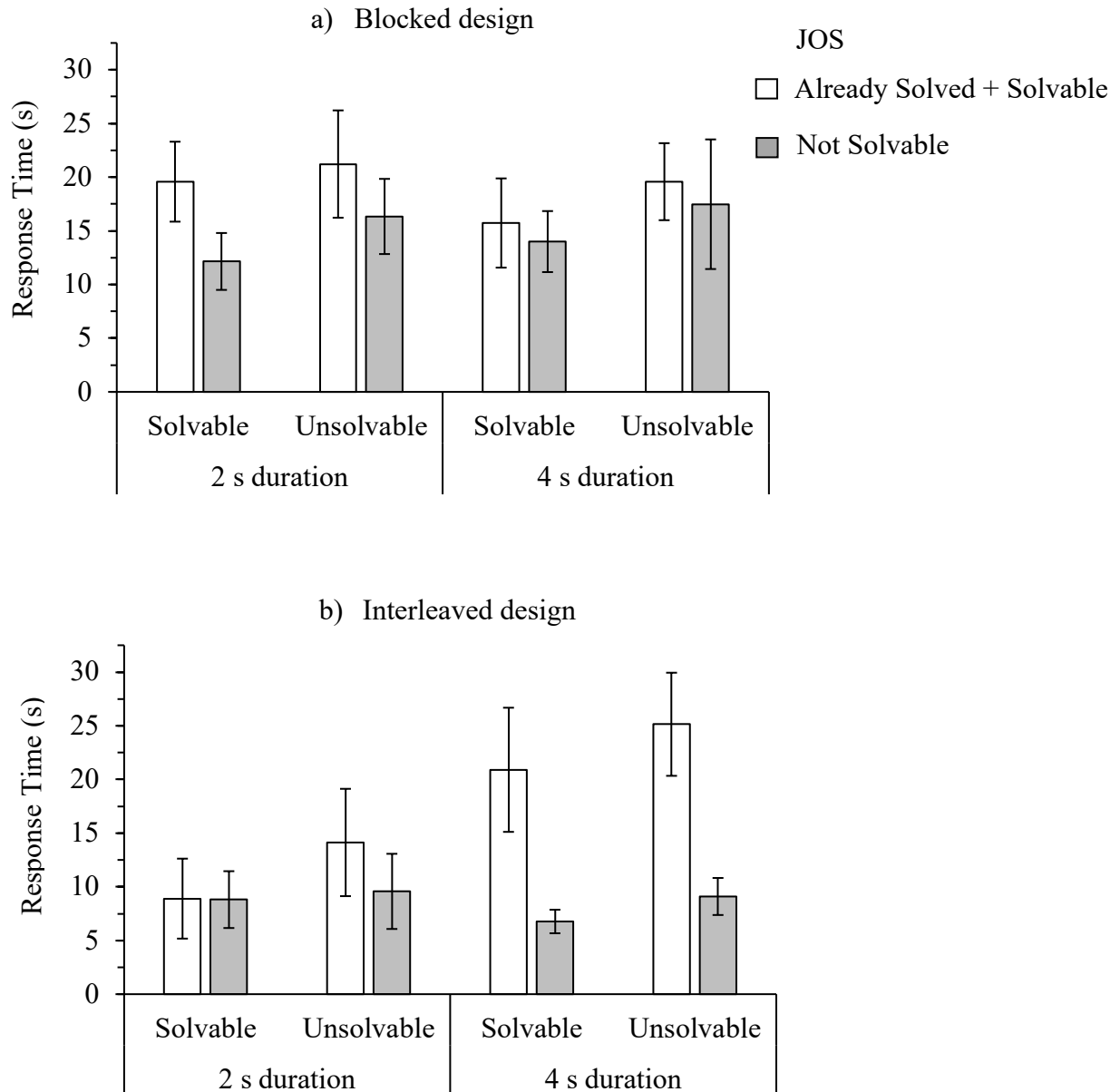
Supplementary Table 4.2

Mean 'Not Solvable' Response Times on Solving Trials: ANOVA Results

Effect	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	1, 82	52.12	52.12	< .001	.39
Anagram type	1, 82	1.00	38.03	< .001	.32
Duration	1, 82	1.77	4.99	.03	.06
Design	1, 82	1.33	3.74	.06	.04
JOS \times Duration	1, 82	0.49	16.72	< .001	.17
JOS \times Design	1, 82	0.22	7.44	.01	.08
JOS \times Duration \times Design	1, 82	0.56	19.18	< .001	.19
JOS \times Anagram type	1, 82	0.09	3.57	.06	.04
JOS \times Duration \times Anagram type	1, 82	0.01	0.53	.47	.01
JOS \times Design \times Anagram type	1, 82	0.02	0.95	.33	.01
Anagram type \times Duration	1, 82	0.00	0.10	.75	.00
Anagram type \times Design	1, 82	0.01	0.27	.61	.00
Duration \times Design	1, 82	1.28	3.59	.06	.04
Anagram type \times Duration \times Design	1, 82	0.00	0.03	.86	.00
JOS \times Anagram type \times Duration \times Design	1, 82	0.07	2.92	.09	.03

Supplementary Figure 4.2

Solving Trials: Mean Response Times for Not-Solvable Responses (Bars show 95% CI of each mean)



For the main effect of anagram type, our results directly replicated those of Burton et al. (2022); not-solvable responses were provided faster for solvable ($M = 10.76$, $SD = 7.92$) than unsolvable anagrams ($M = 14.41$, $SD = 10.80$). The interaction of JOS and duration shows that not-solvable responses for AS+S JOSs were provided faster when anagrams were 2 s ($M = 14.03$, $SD = 10.98$) than 4 s ($M = 20.28$, $SD = 10.99$, $p = .01$), but there was no difference in not-solvable response times for NS JOSs between the 2 s ($M = 11.07$, $SD = 7.74$) and 4 s ($M = 11.75$, $SD = 9.98$, $p = .36$) anagram durations. Importantly, although differences in not-solvable response times between AS+S and NS JOSs were significant in both durations (both $ps \leq .01$), the difference was much greater for 4 s anagram durations. The results indicate that 4 s anagram durations lead to longer not-solvable response times for anagrams that were initially appraised as solvable.

The JOS \times duration \times design interaction was also significant and was followed up with separate interaction contrasts for the 2 s and 4 s durations, collapsed across anagram type (the complete ANOVA results appear in Supplementary Table 4.3). The JOS \times design interaction was significant only in the 4 s duration and reflected a greater difference in not-solvable response times between AS+S and NS JOSs in the interleaved design ($M = 23.88$, $SD = 11.59$ vs. $M = 8.23$, $SD = 5.39$, $p < .001$) than the blocked design ($M = 17.66$, $SD = 9.88$ vs. $M = 15.84$, $SD = 12.34$, $p = .06$). Thus, 4 s anagram durations led to a greater ability of JOSs to predict not-solvable response times, especially when JOSs were interleaved with solving attempts.

Supplementary Table 4.3*Mean 'Not Solvable' Response Times by Duration on Solving Trials: ANOVA Results*

Effect	2 s Duration					4 s Duration				
	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p	<i>df</i>	<i>MSE</i>	<i>F</i>	<i>p</i>	η^2_p
JOS	1, 90	0.20	13.87	< .001	.13	1, 67	1.70	77.17	< .001	.54
Design	1, 90	1.43	6.74	.01	.07	1, 67	0.01	0.05	.82	.00
JOS × Design	1, 90	0.01	0.73	.39	.01	1, 67	0.88	40.04	< .001	.37

The results support our argument that participants may have a desire to ensure their JOSs are consistent with their solving attempts, especially if their JOS immediately preceded their solving attempt. Here, when JOSs were interleaved with solving, participants clearly spent longer deliberating attempting to solve the anagram given a AS+S JOS before choosing to provide a not-solvable response. However, this result was robust only for 4 s anagrams. When given longer to evaluate solvability, participants might be more convinced about their JOSs, thus taking longer to decide to provide a not-solvable response for anagrams appraised as solvable.

Supplementary Table 4.4*Mean Proportions and Standard Deviations for Solving Rates*

Design	2 s Duration		4 s Duration	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Interleaved	.70	.23	.69	.21
Blocked	.68	.22	.68	.22

Supplementary Table 4.5*Mean Proportions and Standard Deviations for Final Not-Solvable Responses to Unsolvable Anagrams*

Design	2 s Duration		4 s Duration	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Interleaved	.76	.18	.80	.18
Blocked	.79	.19	.72	.22

Chapter 5: General Discussion

The overarching aim of my thesis was to elucidate the first stage of meta-reasoning—the initial Judgement of Solvability (JOS). To this end, I examined the influence of task and measurement factors on whether (and which) JOSs can distinguish between solvable and unsolvable problems (i.e., JOS discrimination), and the ability of JOSs to predict problem-solving success, failures, and effort regulation (i.e., JOS predictiveness). I also explored whether individual differences in cognitive reflection underpin the ability of JOSs to discriminate and predict problem-solving outcomes. This final chapter serves to draw together my findings, acknowledge limitations, and suggest additional future directions.

Summary of Experiment Findings

My experiments used longer problem presentations on JOS trials compared to previous studies (e.g., Novick & Sherman, 2003, Topolinski et al., 2016), and participants were also given three JOS options: ‘already solved’ (AS), ‘solvable’ (S), and ‘not solvable’ (NS). Providing an AS JOS option allowed me to separate intuitive JOSs from problems that were solved during the JOS task. This judgement addition enabled me to determine whether intuitive S JOSs (cf. AS JOSs) about anagrams are discriminating. After making a JOS, participants then attempted to solve the anagram. In Experiments 1-3, I manipulated anagram presentation duration across 4 JOS phase blocks of a *training group* (Block 1: 16 s, Block 2: 8 s, Block 3: 4 s, Block 4: 2 s). Experiments 2 and 3 measured the influence of this form of training in comparison to a *no-training group* (where anagram duration remained 2 s in each of the 4 blocks). I expected that giving participants more time to judge solvability might help them generate a better intuitive sense of solvability (Schuster et al., 2020). Participants also received implicit feedback about the accuracy of their intuitions when they solved the anagrams at the longer durations, which I anticipated would help them regulate their JOSs at the shorter durations (Leopold & Leutner, 2015; Leutner et al., 2007). With longer duration

anagrams, participants are likely to move away from simply making a JOS to actively attempting to solve the anagram, but S JOSs and NS JOSs would still capture their intuitions about solvability when solving efforts during the JOS task did not generate solutions. JOSs and solution attempts were set up in blocks such that participants made a JOS for the entire set of reasoning problems and then attempted to solve each one in a separate solving block.

After separating solutions found during the JOS (AS JOSs) from JOS intuitions (S JOSs), Experiments 1-3 typically revealed above-chance discrimination for S JOSs, even at the briefest anagram duration (2 s). These findings align with several other studies suggesting that JOSs can be sensitive to problem solvability (e.g., Balas et al., 2011; Bolte & Goschke, 2005; Markovits et al., 2015; Novick & Sherman, 2003; Siedlecka et al., 2016; Topolinski & Strack, 2009a; Undorf & Zander, 2017). However, S JOSs discrimination was much weaker than AS JOSs discrimination, indicating that studies that did not parse out solved problems may have inflated JOS accuracy. Moreover, I found that S JOSs were not generally predictive of anagram solving success. I also found that training did not lead to more discriminating S JOSs in block 4 of the training group or more predictive S JOSs.

Experiment 3 used a self-regulated solving phase in which participants were informed that they would be presented with a subset of both solvable and unsolvable anagrams from the JOS trials. They could either attempt a solution, pass if they believed the anagram was solvable but could not solve it, or make a not-solvable response if they believed the anagram was unsolvable, akin to a final JOS (Ackerman & Beller, 2017; Lauterman & Ackerman, 2019). The use of self-regulated solving trials added an element of uncertainty regarding anagram solvability that was absent in Experiments 1 and 2. This change may have led to greater problem-solving disengagement for solvable anagrams. Indeed, overall rates of solving were lower in the self-regulated experiments. Here, participants showed a lower rate of anagram solving for anagrams that had received NS JOSs than S JOSs (cf. Experiments 1

and 2). However, S JOSs were again not more frequent among solved anagrams than among not-solved anagrams when the solving phase was self-regulated.

Chapter 3 reanalysed data from Experiments 1-3 with consideration of participants' Cognitive Reflection Test (CRT) scores, to see whether individual differences in *cognitive reflection* were related to JOS discrimination and predictiveness. Here I also investigated whether allowing participants more time to develop their intuitions with longer-duration anagrams bolstered JOS discrimination and predictiveness for those lower in cognitive reflection. I found that cognitive reflection served to increase the likelihood of reporting an anagram as already-solved and general anagram-solving performance overall, but not the ability of S JOSs to discriminate or predict problem-solving outcomes. Chapter 3 notes some possible reasons why a more reflective thinking style did not predict more accurate and predictive S JOS intuitions, including potential individual differences in how correct responses are generated on the CRT (i.e., *smart intuitive* versus *smart deliberative* response types). That is, smart intuition might more strongly predict rapid solving during the JOS, but smart deliberation might better relate to an ability to make accurate and predictive S JOSs. I suggested that future research should separate these processes using a two-response version of the CRT (e.g., Burič & Šrol, 2020; Strudwicke et al., 2022).

In Experiment 4, JOSs and solving attempts were either *blocked* into two separate phases (akin to Experiments 1-3) or were *interleaved* such that a solving attempt followed each JOS. Solving trials in each design were self-regulated, as in Experiment 3. Here, S JOSs were more discriminating and more predictive in the interleaved than the blocked design. In the interleaved design, I suggested that memory for the JOS would be more available given that solving immediately precedes the JOS. Greater memory availability may have led to more persistent solving attempts assuming that problem solvers align their solving efforts with their JOSs, which may therefore increase solving rates (Pennycook et al., 2015a). In line

with this possibility, the interleaved design also led to a higher likelihood of an anagram being reported as already solved during the JOS. Thus, relative to the blocked design, interleaving provides participants with continuous experience attempting to solve anagrams, which may stimulate greater sensitivity to solvability.

Prior research has mainly focussed on whether JOSs are influenced by heuristic stimulus cues indicating solvability (e.g., Ackerman & Beller, 2017; Balas et al., 2011; Markovits et al., 2015; Topolinski et al., 2016; Topolinski & Strack, 2009a; Valerjev & Dujmović, 2020). My thesis shows that JOS discrimination and predictiveness can be influenced by experiment factors (such as anagram duration length, study design, and whether the solving trials are self-regulated), and measurement factors (such as whether solutions found during the JOS are separated from intuitions about solvability). Given that meta-reasoning is still an emerging research area, these findings inform the decisions researchers must make when designing studies that rely on JOSs. For example, a researcher interested in measuring intuition without interference from deliberate solving attempts should consider using a blocked design for measuring JOSs.

My experiments revealed several other interesting findings. For example, although S JOSs were often discriminating and predicted solving effort-regulation, they did not always predict successful problem-solving. Another interesting finding was that for self-regulated solving trials, participants rarely used a “pass” response – they either solved the anagram or indicated it was not solvable. Additionally, although cognitive reflection was not related to meta-reasoning intuition, it *was* related to the rate of anagrams solved, as well as to how rapidly participants solved the anagrams. I next discuss each of these novel findings.

JOSs Can be Discriminating, but not Predictive of Problem-Solving Success

Ackerman and Thompson (2017) theorise that intuitive JOSs are related to one's decision to attempt solving, in addition to how much effort one invests in solving. In my experiments with self-regulated solving trials (Experiments 3 and 4), NS JOSs predicted greater proportions of decisions to disengage from solving (i.e., final not solvable responses on a solving trial), and S JOSs predicted longer anagram-solving attempts. Furthermore, each of these relationships was stronger when the solving attempt occurred immediately after the JOS was provided (i.e., interleaved design) than when JOSs and solving attempts were separated into blocks (i.e., blocked design). However, my experiments revealed an interesting discrepancy: although reasoners could distinguish between solvable and unsolvable anagrams, their S JOSs did not typically predict their actual solving outcomes.

An important question is why S JOSs were not as predictive as they were discriminating. I found some potential answers to this question. For example, S JOSs were more frequent among successfully solved anagrams when the study design was interleaved than blocked (see Chapter 4) and predicted higher rates of solving compared to 'not solvable' JOSs when the solving trials were self-regulated (see Chapter 2) due to greater memory availability for the JOS. I anticipated that intuition about anagram solvability would operate similarly at both the JOS stage and the solving stage, given evidence that intuition remains stable over time (Stagnaro et al., 2018). This should have resulted in greater effort regulation for anagrams judged as 'solvable', which should have seeded greater problem-solving success (Pennycook et al., 2015a). Consistent with this possibility, Experiments 2, 3, and 4 revealed that participants who made more discriminating S JOSs *were* more likely to go on to solve anagrams they intuitively felt were solvable. This result lends some credence to the possibility that solvability intuition about an anagram remains stable at the JOS stage and the solving stage.

The finding that more accurate S JOS discrimination was related to more predictive S JOSs supports the idea that the ability to have accurate meta-reasoning intuitions about solvability, and the ability to problem-solve, are linked to each other. Indeed, Novick and Sherman (2003) found that those with greater anagram solving expertise were able to make more accurate solvability judgements, providing further support for the association between solvability intuitions and problem-solving. However, given that the relationship between S JOS discrimination and S JOS predictiveness was correlational, I cannot infer whether JOS intuition indeed causes greater effort regulation, and in turn, a greater likelihood of problem-solving. It remains possible that the ability to make intuitive solvability judgements and the ability to problem solve rely on (at least somewhat) dissociable cognitive skills.

Participants in my experiments made Type 1 signal-detection judgements on JOS trials, whereby they attempted to distinguish between signals (solvable anagrams) and noise (unsolvable anagrams) (Maniscalco & Lau, 2012). Type 1 judgements require sensitivity to stimulus-level features such as bigram frequency or diagnostic letter combinations to distinguish between signals and noise. These sensitivities to stimulus features help discriminate solvability, but they might not translate to better anagram-solving ability if reasoners do not have the knowledge structures for carrying out anagram solving, such as an understanding of solving procedures and strategies like anagram restructuring (Burič & Konrádová, 2020; Stanovich, 2018). If these two skills are at least partially dissociable this would explain why S JOSs were generally discriminating but not predictive.

It is worth noting that Type 1 JOSs do not capture participants' metacognitive beliefs about whether they believed they could solve each anagram. Participants' S JOSs may have been more predictive had I asked participants to judge whether *they* thought that they could solve the anagrams, rather than judge whether the anagrams were actually solvable. Self-efficacy has been found to relate to academic performance outcomes (Ackerman et al., 2002;

Honicke & Broadbent, 2016; Talsma et al., 2018). Higher self-efficacy may promote a self-fulfilling prophecy regarding performance because those with high self-efficacy tend to show greater perseverance on tasks (Pajares, 2006). In this way, metacognitive (i.e., Type 2) JOSs may be more strongly related to solving effort regulation, and in turn, to problem-solving success for problems that reasoners believe they can solve. Although I had anticipated that Type 1 JOSs would also promote greater perseverance on anagrams judged as solvable, including unsolvable anagrams on JOS and solving trials may have decreased motivation to solve some anagrams or provide rational JOSs, given that whether they could solve the anagram was out of the participants' control (Skinner, 1979). Ackerman and Beller (2017) had participants rate their own personal ability to solve the problem. However, their study also presented participants with a mixture of solvable and unsolvable problems, so participants' personal JOS ratings may have been influenced by the perceived likelihood that the problem was solvable. Thus, future studies should measure how well participants make JOSs that relate to their beliefs about whether they could solve each anagram for solvable anagrams only, to test whether S JOSs more strongly predict problem-solving successes. These metacognitive JOSs could also be examined across both blocked and interleaved designs, and at different durations, to see whether study design and/or duration moderate how well metacognitive JOSs predict problem-solving success.

As a further consideration here, Novick and Sherman (2003) argued that there may be two distinct processes for solving anagrams: pop-out solutions and search solutions. These two processes might differentially influence how well JOSs about anagram solvability can discriminate and predict later problem-solving success. Pop-out solutions are sudden and instantaneous, whereas search solutions involve serially testing possible solutions. Novick and Sherman found that participants who used either strategy were able to discriminate between solvable and unsolvable anagrams. However, interestingly, participants who used

pop-out solution strategies had more accurate JOS discrimination. Moreover, this result was not compromised by solution retrieval during the solvability judgement. In my experiments, those participants that used pop-out solution strategies may have made more discriminating S JOSs and generated more rapid solutions on solving trials (Aziz-Zadeh et al., 2009; Novick & Sherman, 2003). Meanwhile, participants that relied on search solutions may have still been able to make accurate S JOSs, if the ability to make an accurate JOS for an anagram depends on a serial search for information. However, the use of this strategy may have led them to run out of time on time-limited solving trials, or perhaps to set a low threshold for deciding when to disengage from problem-solving on self-regulated solving trials (i.e., they were quicker to decide that an anagram was ‘not solvable’ on a solving trial if their search fell short), rendering their S JOSs not as predictive. In this way, perhaps the dissociation between S JOS discrimination and predictiveness was driven by differences in solving strategies.

To further explore this possibility, future studies could ask participants to report how their anagram solution arose (i.e., whether it came to fruition suddenly, or after a deliberative solving process). Trials or participants could then be separated to examine whether the use of a pop-out and/or search solution process affects S JOS predictiveness. In addition, it would be worthwhile to test whether anagram duration and study design moderate how well S JOS predict problem-solving success when a pop-out versus search strategy was used. For example, search solvers might make more predictive S JOSs in the interleaved design because they have already commenced the solution search process during the JOS.

Do JOSs Predict Feelings of Rightness?

Although I found that S JOSs generally did not predict successful problem-solving outcomes, my experiments established that JOSs do predict how much effort people regulate toward problem-solving. For example, participants were faster to disengage from solving an unsolvable anagram when they had assigned it an NS JOS and were slower to disengage

when they had assigned it an S JOS. My experiments also established that NS JOSs were more frequent among solving failures than among solving successes. These findings provide useful information about how people regulate effort following a JOS. However, a question that my thesis did not address is how JOSs influence subsequent meta-reasoning monitoring stages, specifically the *feeling of rightness* (Ackerman & Thompson, 2017).

The feeling of rightness is simply an assessment of whether an initially generated solution feels right. If the feeling of rightness is strong, then a reasoner may choose to stop reasoning and provide the answer, whereas if it is weak, then they may opt to reconsider their answer or continue reasoning until they generate an answer they are confident in. In Ackerman and Thompson's (2017) meta-reasoning theory, a feeling of rightness judgement follows the initial JOS, but it is not clear whether these judgements are linked or independent. That is, does a strong feeling that a problem is solvable relate to whether reasoners feel an initially generated response is right? If JOSs predict reasoning outcomes, then we would expect that they would also relate to feelings of rightness. Given that my experiments found evidence that reasoners' S JOSs do not generally predict problem-solving successes, it may be that S JOSs are dissociable from feelings of rightness. Conversely, S JOSs may lead to feelings of rightness if reasoners are motivated to mitigate any dissonance between their initial JOSs and feelings of rightness. In this case, JOSs might relate to feelings of rightness, yet they might not relate to problem-solving outcomes.

The paradigms I developed for measuring JOS discrimination and predictiveness would need to be modified for measuring whether JOSs predict feelings of rightness. Anagrams (at least unique solution anagrams) are not ideal stimuli for this purpose because they do not offer an alternative but incorrect solution. With these stimuli, a strong feeling of rightness would follow a correct solution and a weak feeling of rightness would follow an incorrect solution. In this way, measuring feelings of rightness on an anagram task would

provide no more useful information than simply measuring solving outcomes. Instead, heuristics-and-biases problems might be used to measure whether JOSs and feelings of rightness relate to each other in a dual-process framework. However, such problems might be problematic for measuring JOSs given that the intuitive but incorrect response on these tasks may be generated rapidly after reasoners have processed the problem (Bago & De Neys, 2017; Strudwicke et al., 2022). Therefore, researchers could include an ‘already solved’ JOS with these problems to exclude those solved during the JOS. However, in many cases, these problems are not actually solved correctly so some of these AS JOSs would be mistaken. Therefore, a better direction would be to use analytic problems that do not cue an attractive, prepotent response, and that are more likely to be solved using search solution strategies (e.g., Raven’s Matrices).

Pass Responses During Self-Regulated Solving

When problem-solving was self-regulated (Experiments 3 and 4), pass responses were very rare during the solving trials (indeed, they were too rare to analyse). When a solution was not found for a solvable anagram, participants were more inclined to provide a final not-solvable response rather than to pass. Also, solution response times for solvable anagrams in Experiments 3 and 4 were faster than the 45 s period provided in Experiments 1 and 2, with mean solution response times for each JOS ranging from 4.3 to 13.7 s. Most participants appear to have allocated less than 45 s to their solving attempts before responding, and thus may have anchored their decision about final solvability based on whether their solving time exceeded their set-solving threshold (see Payne & Duggan, 2011).

Final decisions about solvability may have also been based on the perceived likelihood of anagram solving success. If participants felt they were likely to solve the problem within their response-time threshold, then their failure to do so may have sponsored a not-solvable response. Consistent with that possibility, research has shown that participants

who are more decisive about the items they recall truncate their memory search earlier than those who are less decisive (Dougherty & Harbison, 2007). In my experiments, participants who believed they would be able to solve each solvable anagram may have more likely to indicate an anagram was not-solvable on a solving trial if they could not generate a solution for it. Future research could explore whether self-perceived anagram-solving ability predicts how often participants make ‘pass’ versus ‘not-solvable’ responses on solving trials.

Relatedly, my experiment was advertised on MTurk as a “solving word scrambles” task. This label may have attracted participants who believed they were good at solving anagrams.

Thus, future studies might explore the impact of anagram-solving experience (or experience playing the recent craze of Wordle) on solving-phase performance, as well as on JOS discrimination and predictiveness.

Individual Differences in Meta-Reasoning and Insight Reasoning

In Chapter 3, I found that individuals higher in cognitive reflection merely solved more of the longer-duration anagrams (i.e., they reported a higher rate of AS JOSs) rather than making more accurate and predictive intuitive JOSs (i.e., S JOSs). These findings further emphasize the importance of separating intuitions from solutions when measuring JOSs. If I had not parsed out problems that were solved from the JOS measure, greater cognitive reflection would have appeared to improve S JOS discrimination and predictiveness. This study also highlighted some possible barriers to measuring the relationship between cognitive reflection and JOS intuition, namely, the inability to separate the two types of correct responses (smart deliberation versus smart intuition) on the Cognitive Reflection Test (CRT; Frederick, 2005; Toplak et al., 2014), and whether these responses differentially predict meta-reasoning ability. Although cognitive reflection was not related to solvability intuition in this study, Chapter 3 highlighted that cognitive reflection nonetheless influences anagram solving ability.

Cognitive Reflection Predicts Anagram Solving

The findings from Chapter 3 are among the first to show that cognitive reflection is related to another measure of insight reasoning, namely anagram solving ability. This result provides new evidence for the CRT's predictive power beyond numerical reasoning ability (Welsh et al., 2013). However, some researchers have rejected the notion that the CRT is related to insight-reasoning ability. For example, some contend that the conflict detection processes which underly correct responses on the CRT are qualitatively distinct from “restructuring” in insight reasoning (Toplak et al., 2011, 2014; West et al., 2012). Nevertheless, Patel et al. (2019) highlighted an interesting similarity between the lily-pad problem on the CRT (see Chapter 1) and the following insight reasoning problem from Gilhooly and Murphy (2005):

“There is a container of Murples. The Murples double in number every day. The container will be full in 60 days. In how many days will it be half full?” (p. 285)

The Murples problem is conceptually identical to the lily-pad problem. Indeed, the Murples problem clusters with other insight-reasoning problems, suggesting that these problems capture cognitive restructuring, or insight (Gilhooly & Murphy, 2005, as cited in Patel et al., 2019). Furthermore, the CRT has been found to predict the number of correct responses on the Remote Associates Task (Barr et al., 2015). Hence, errors on the CRT might arise due to difficulties in restructuring the problem or to a lack of insight reasoning ability (Patel et al., 2019). However, other evidence using the Remote Associates Task indicates that cognitive reflection predicts insight reasoning without restructuring after an impasse (Ash et al., 2018). This result suggests that more reflective thinkers may generate more potential solution representations than less reflective thinkers and hence do not need to restructure the problem.

Chapter 3 provided new evidence of a relationship between the CRT and verbal insight reasoning ability (anagram solving). As discussed in Chapters 1 and 3, anagram solving may occur when the solution is spontaneously and rapidly retrieved with little awareness of how it was generated (Bowden & Jung-Beeman, 2003; Chu & MacGregor, 2011; Gilhooly & Fioratou, 2009; Metcalfe, 1986; Metcalfe & Wiebe, 1987; Novick & Sherman, 2008; Salvi et al., 2016; Weisberg, 1992). Alternatively, anagram solving may occur via unconscious cognitive restructuring (Ash & Wiley, 2006; Ohlsson, 2011), or conscious cognitive restructuring (Gilhooly & Fioratou, 2009) such as rearranging the letters of the anagram following a solving impasse. These anagram-solving strategies may be similar to reasoning on the CRT, where reasoners either intuit the response immediately (Burič & Konrádová, 2020; Burič & Šrol, 2020) or engage in more deliberative reasoning after error-detection in their initial, intuitive response (Pennycook et al., 2015a; Toplak et al., 2014). Future research should establish the relationship between insight reasoning ability and cognitive reflection ability. For example, such research could examine whether the relationship between cognitive reflection and insight reasoning generalises to other insight reasoning problems, such as Rebus puzzles (MacGregor & Cunningham, 2008) or The Eight Coin Problem (Öllinger et al., 2013).

Effects of Open-Minded Thinking on JOS Discrimination and Predictiveness

I did not find that greater cognitive reflection promoted more accurate and predictive JOS intuition. It may be worthwhile to consider other potential individual differences that might offer advantages in meta-reasoning. One such individual difference is the ability to engage in actively open-minded thinking (Stanovich & West, 2007). To reason well, reasoners must consider and try to sidestep potential biases in their reasoning processes (Pennycook et al., 2015a). Thus, one's willingness to reconsider one's beliefs when presented with evidence to disconfirm their beliefs, that is, *open-minded thinking*, may

promote better meta-reasoning. Research has found that open-minded thinking predicts reasoning accuracy on heuristics-and-biases tasks, where unbiased reasoning is necessary (Stanovich & West, 2008; Toplak et al., 2011; West et al., 2008). Greater open-minded thinking has also been shown to predict more accurate metacognitive decisions (Haran et al., 2013; Mellers et al., 2015; Strudwicke et al., 2022), and reduced overconfidence on heuristics-and-biases tasks (Kleitman et al., 2019).

As discussed in Chapter 3, JOS accuracy and predictiveness may depend on one's ability to decouple from incorrect, or biased intuitions about solvability and self-regulate more accurate JOSs. In Chapter 3, I argued that a reflective thinking style might sponsor better JOS discrimination and predictiveness given that reflective thinkers have a disposition to "decouple" from biased reasoning (Pennycook et al., 2015a). Alternatively, however, decoupling from biased intuitions about solvability might relate to actively open-minded thinking. If actively open-minded thinking requires reasoners to readjust their beliefs based on disconfirming evidence, then perhaps open-minded thinkers would be more amenable to feedback about their intuitions (i.e., when they solve anagrams during JOS trials) that confirms whether their JOSs are correct or incorrect. For example, if an open-minded thinker observes a miscalibration between their intuitive feelings about a problem's solvability and the occasions they uncover a solution on a JOS trial, they might then readjust their beliefs about problem solvability. However, less open-minded thinkers may be less inclined to readjust, given that they tend to be less flexible and more dogmatic in their thinking (Svedholm-Häkkinen & Lindeman, 2018). If so, then more open-minded thinkers might make more discriminating and predictive JOSs. Further, because open-minded thinkers may be more open to updating their beliefs about solvability, they may also be more responsive to training with longer-duration anagrams. Future research should test these ideas.

Optimising JOS Discrimination and Predictiveness

My Experiments 2 and 3 investigated whether training via initial blocks of longer-duration anagrams would lead participants to develop more accurate and predictive JOSs. Contrary to that possibility, training did not lead to more discriminating JOSs in the final 2 s block and having more time to develop solvability intuitions did not enhance how well S JOSs predicted problem-solving outcomes. In Chapter 2, I noted some possible explanations for the absence of such training effects. For example, participants in the training group may have shifted their efforts to solving in the longer-duration blocks rather than developing their intuition. I also noted that perhaps more explicit, trial-level feedback about JOS accuracy was needed for participants to effectively regulate their JOSs.

However, in Experiment 4, I found that interleaving JOSs and solving led to more discriminating and predictive S JOSs. A notable difference between interleaved and blocked designs is that the former provides participants with incremental solving experience, which may sponsor greater sensitivity to solvability. Problem-solving expertise is correlated with task experience (Novick & Sherman, 1996, as cited in Novick & Sherman, 2003), and is also related to greater sensitivity to problem solvability (Novick & Sherman, 2003). Given these findings, future research could examine the effects of training via practice problems presented before the JOS task, and whether it improves sensitivity to problem solvability relative to no practice. Such a study would reveal whether JOS accuracy and predictiveness is related to deliberate practice and/or experience with a given reasoning task (Griffin et al., 2009). Furthermore, since cognitive reflection predicts reasoning ability (Patel et al., 2019; Toplak et al., 2014), and more accurate reasoning is related to better meta-reasoning (Borracci & Arribalzaga, 2018; Dunning et al., 2003; Kruger & Dunning, 1999), future studies could measure whether practice/experience improves JOSs for those lower in cognitive reflection.

Other Limitations and Future Directions

My anagram stimuli were designed to minimise potential biasing stimulus features. However, some of the anagrams I used were pronounceable whereas others were not. This may be relevant given that reasoners use processing fluency as a heuristic cue of anagram solvability (Topolinski et al., 2016). Although Topolinski et al. found that JOSs were discriminating whether anagrams were pronounceable or not, my participants' JOSs may have been biased by certain stimulus features (such as anagram pronounceability, bigram frequencies, number of syllables, e.g., Adams et al., 2011; Dominowski & Duncan, 1964; Gilhooly, 1978). Examining the anagram features that inform/bias JOSs was outside of the scope of my thesis, but future studies should explore the impact of biasing stimulus features on JOS discrimination and predictiveness.

Another limitation of my thesis studies was that my measure of JOS discrimination could not separate potential influences of response bias. A bias-free sensitivity measure such as Type 1 d' would have been preferable. Unfortunately, the use of such a measure would have been problematic given the design of my experiments. The experiments in Chapter 2 had only 10 solvable and 10 unsolvable anagrams per block, and these were further subdivided by participants into AS, S, and NS JOSs. Depending on the JOS, hit rates of 1 or false alarm rates of 0 were very common and would often have required correction (and a $1/2n$ correction can be substantial when n is small). To overcome this issue, researchers could include more JOS trials (i.e., 100 solvable and 100 unsolvable anagrams) and present anagrams for durations where S JOSs are still discriminating but AS JOSs are largely absent. However, including more trials may risk a speed-accuracy trade-off if participants become fatigued, and in turn, this may result in less reliable JOSs (Healy et al., 2004). Thus, there are pros and cons to measuring JOS discrimination using hits versus false alarms versus Type 1 d' .

Another reason I could not use Type 1 d' was that my primary interest was whether S JOSs were discriminating. From a signal-detection theory perspective, participants set a higher criterion for an AS JOS and a lower criterion for a S JOS, and anything that falls below the 'solvable' criterion would receive an NS JOS. The area under the signal-detection curve between the AS JOS and S JOS criteria is not defined, thus the Type 1 d' for S JOSs would not provide a valid measure of discrimination. The only workaround here would be to present anagrams briefly enough to eliminate AS JOSs, thus enabling calculation of a valid Type 1 d' for S JOSs. However, as discussed earlier, this approach risks participants reverting to guessing or other non-rational strategies for judging solvability.

Another lingering issue with my studies concerns the assumption JOSs are made begin before any deliberate reasoning occurs (Ackerman & Thompson, 2017). As is true of most meta-reasoning studies, my experiments cannot rule out the possibility that participants deliberately engaged in problem solving during the JOS task—especially when longer-duration anagrams were used. I had anticipated this, hence my inclusion of an AS JOS response option (Topolinski et al., 2016). However, provision of an AS JOS option and longer-duration anagrams may have encouraged participants to try to solve the problems during the JOS, rather than on developing their solvability intuition.

Studies measuring JOSs often distinguish between initial JOSs and final JOSs (Ackerman & Beller, 2017; Ackerman & Thompson, 2017; Lauterman & Ackerman, 2019). The final JOS reflects a reasoner's judgement about whether a problem was solvable, after a solving attempt. Where longer-duration anagrams were provided, the training groups may have captured a JOS more analogous to a final JOS rather than an initial JOS. Even so, the order of the JOS versus solving tasks should still reflect the time course of meta-reasoning, as established by Ackerman and Thompson (2017). That is, collecting the JOS before the

solving phase should capture whether “solvable” JOSs are related to later solving efforts, even if some solving efforts occurred during the JOS task.

Finally, my studies focussed solely on insight problems and solely on anagrams. Future research should explore whether reasoners can make accurate and predictive JOSs for reasoning problems that require greater analyticity to solve, such as mathematic problems or pattern sequencing. Considering the important implications of JOS research for understanding and assisting learners, insight reasoning problems are not representative of the type of reasoning problems learners typically face. So far, only Lauterman and Ackerman (2019) have examined JOSs using analytic reasoning problems. Using a blocked design, they found that JOSs about Raven’s Matrices were discriminating only when the matrix violated fewer rules of solvability, and that “solvable” JOSs led to greater effort investment regardless of the actual solvability of the matrix. Though important, their study left open some interesting questions regarding the task conditions that influence intuitions about solvability. For example, how much time do reasoners need to develop accurate intuitions for analytic problems? Lauterman and Ackerman measured JOSs for Raven’s matrices at 4 s presentation durations, where JOSs were not always discriminating. Perhaps 4 s was not enough time for reasoners to develop accurate intuition for such problems. In Experiment 3, I found that reasoners took longer to abandon unsolvable problems that they judged to be solvable, especially with shorter anagram durations. This result suggests that shorter durations may limit how well people can develop their intuitions regarding a problem’s solvability, which may lead to unreliable JOSs that result in wasted time on futile efforts to solve unsolvable problems.

Conclusions

Understanding the process of meta-reasoning is a relatively new area of research. Relative to reasoning research in general, only a small number of studies have sought to

elucidate meta-reasoning and its initial stage—judging solvability. My thesis highlighted several important measurement and task factors that contribute to whether and how well JOSs discriminate between solvable and unsolvable problems, and whether they predict later problem-solving outcomes. I established that giving reasoners longer-duration problems to develop their intuitions can improve JOS discrimination, but it also increases problem-solving during the JOS task and this needs to be taken into account (e.g., by collecting AS JOSs). My thesis also established that interleaving JOSs with solving attempts (cf. blocking) leads to more accurate discrimination, and more strongly predicts problem-solving performance. Thus, how one studies the first stage of meta-reasoning will impact what one finds, rendering study design a nontrivial decision for researchers to consider. Given that the initial stage of meta-reasoning informs future effort regulation, understanding what drives accurate and predictive solvability judgements is important. The experiments in my thesis represent an important contribution toward this goal.

References

- Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33*(4), 587-605. [https://doi.org/10.1016/S0191-8869\(01\)00174-X](https://doi.org/10.1016/S0191-8869(01)00174-X)
- Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General, 143*(3), 1349-1368. <https://doi.org/10.1037/a0035098>
- Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psychological Topics, 28*(1), 1-20. <https://doi.org/10.31820/pt.28.1.1>
- Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgements during reasoning and memorisation. *Thinking and Reasoning, 23*(4), 376-408. <https://doi.org/10.1080/13546783.2017.1328373>
- Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*(8), 607-617. <https://doi.org/10.1016/j.tics.2017.05.004>
- Ackerman, R., Yom-Tov, E., & Torgovitsky, I. (2020). Using confidence and consensuality to predict time invested in problem solving and in real-life web searching. *Cognition, 199*, 104248. <https://doi.org/10.1016/j.cognition.2020.104248>
- Adams, J. W., Stone, M., Vincent, R. D., & Muncer, S. J. (2011). The role of syllables in anagram solution: A rasch analysis. *The Journal of General Psychology, 138*(2), 94-109. <https://doi.org/10.1080/00221309.2010.540592>
- Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of experimental psychology: General, 136*(4), 569. <https://doi.org/10.1037/0096-3445.136.4.569>

- Arechar, A. A., & Rand, D. G. (2021). Turking in the time of COVID. *Behavior Research Methods*, *53*(6), 2591-2595. <https://doi.org/10.3758/s13428-021-01588-4>
- Arnold, M. M., Chisholm, L. M., & Prike, T. (2016). No pain no gain: The positive impact of punishment on the strategic regulation of accuracy. *Memory*, *24*(2), 146-153. <https://doi.org/10.1080/09658211.2014.990982>
- Arnold, M. M., Graham, K., & Hollingworth-Hughes, S. (2017). What's context got to do with It? Comparative difficulty of test questions influences metacognition and corrected scores for formula-scored exams. *Applied Cognitive Psychology*, *31*(2), 146-155. <https://doi.org/10.1002/acp.3312>
- Arnold, M. M., Higham, P. A., & Martín-Luengo, B. (2013). A little bias goes a long way: The effects of feedback on the strategic regulation of accuracy on formula-scored tests. *Journal of Experimental Psychology*, *19*(4), 383-402. <https://doi.org/10.1037/a0034833>
- Arnold, M. M., & Prike, T. (2015). Comparative difficulty and the strategic regulation of accuracy: The impact of test-list context on monitoring and meta-metacognition. *Acta Psychologica*, *157*, 155-163. <https://doi.org/10.1016/j.actpsy.2015.02.018>
- Ash, I. K., Lee, K. D., & Shurkova, E. Y. (2018). The relationship of working memory span, cognitive reflection test, and compound remote associates performance. the 59th Psychonomic Society Annual Meeting, New Orleans, LA.
- Ash, I. K., & Wiley, J. (2006). The nature of restructuring in insight: An individual-differences approach. *Psychonomic Bulletin & Review*, *13*(1), 66-73. <https://doi.org/10.3758/BF03193814>
- Aziz-Zadeh, L., Kaplan, J. T., & Iacoboni, M. (2009). "Aha!": The neural correlates of verbal insight solutions. *Human brain mapping*, *30*(3), 908-916. <https://doi.org/10.1002/hbm.20554>

- Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92-107.
<https://doi.org/10.1016/j.learninstruc.2014.04.004>
- Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90-109.
<https://doi.org/10.1016/j.cognition.2016.10.014>
- 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. <https://doi.org/10.1080/13546783.2018.1507949>
- Balas, R., Sweklej, J., Pochwatko, G., & Godlewska, M. (2011). On the influence of affective states on intuitive coherence judgements. *Cognition and Emotion, 26*(2), 312-320.
<https://doi.org/10.1080/02699931.2011.568050>
- Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015). Reasoned connections: A dual-process perspective on creative thought. *Thinking & Reasoning, 21*(1), 61-75.
<https://doi.org/10.1080/13546783.2014.895915>
- Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 570. <https://doi.org/10.1037/0278-7393.33.3.570>
- Benjamin, A. S. (2005). Response speeding mediates the contributions of cue familiarity and target retrievability to metamnemonic judgments. *Psychonomic Bulletin & Review, 12*(5), 874-879. <https://doi.org/10.3758/BF03196779>
- Berardi-Coletta, B. (1995). Metacognition and problem solving: A process-oriented approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1). <https://doi.org/10.1037/0278-7393.21.1.205>

- Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of experimental psychology. Learning, memory, and cognition*, 35(2), 317-333. <https://doi.org/10.1037/a0014873>
- Bialek, M., & Pennycook, G. (2017). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-017-0963-x>
- Bjork, R. A. (1994). Memory and Meta-memory Considerations in the Training of Human Beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). The MIT Press.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, 64, 417-444. <https://doi.org/annurev-psych-113011-143823>
- Boekaerts, M. (1999). Self-regulated learning: where we are today. *International Journal of Educational Research*, 31(6), 445-457. [https://doi.org/https://doi.org/10.1016/S0883-0355\(99\)00014-2](https://doi.org/https://doi.org/10.1016/S0883-0355(99)00014-2)
- Bolte, A., & Goschke, T. (2005). On the speed of intuition: Intuitive judgments of semantic coherence under different response deadlines. *Memory & Cognition*, 33(7), 1248-1255. <https://doi.org/10.3758/BF03193226>
- Bolte, A., Goschke, T., & Kuhl, J. (2003). Emotion and intuition: Effects of positive and negative mood on implicit judgments of semantic coherence. *Psychological science*, 14(5), 416-421. <https://doi.org/10.1111/1467-9280.01456>
- Borracci, R. A., & Arribalzaga, E. B. (2018). The incidence of overconfidence and underconfidence effects in medical student examinations. *Journal of Surgical Education*, 75(5), 1223-1229. <https://doi.org/10.1016/j.jsurg.2018.01.015>

- Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, 35(4), 634-639. <https://doi.org/10.3758/BF03195543>
- Bowers, K. S., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology*, 22(1), 72-110. [https://doi.org/10.1016/0010-0285\(90\)90004-N](https://doi.org/10.1016/0010-0285(90)90004-N)
- Burič, R., & Konrádová, L. (2020). Mindware instantiation as a predictor of logical intuitions in the Cognitive Reflection Test. *Studia Psychologica*, 63(2). <https://doi.org/10.31577/sp.2021.02.822>
- Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, 32(4), 460-477. <https://doi.org/10.1080/20445911.2020.1766472>
- Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of personality and social psychology*, 90(1), 60.
- Burton, O. R., Bodner, G. E., Williamson, P., & Arnold, M. M. (2022). How accurate and predictive are judgments of solvability? Explorations in a two-phase anagram solving paradigm. *Metacognition and Learning*. <https://doi.org/10.1007/s11409-022-09313-y>
- Chu, Y., & MacGregor, J. N. (2011). Human performance on insight problem solving: A review. *The Journal of Problem Solving*, 3(2), 6. <https://doi.org/10.7771/1932-6246.1094>
- Chuderski, A., & Jastrzębski, J. (2018). Much ado about aha!: Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of experimental psychology. General*, 147(2), 257-281. <https://doi.org/10.1037/xge0000378>

- Coutinho, M. V., Thomas, J., Alsuwaidi, A. S., & Couchman, J. J. (2021). Dunning-kruger effect: Intuitive errors predict overconfidence on the cognitive reflection test. *Frontiers in psychology*, 1040. <https://doi.org/10.3389/fpsyg.2021.603225>
- de Chantal, P.-L., Newman, I. R., Thompson, V., & Markovits, H. (2020). Who resists belief-biased inferences? The role of individual differences in reasoning strategies, working memory, and attentional focus. *Memory & Cognition*, 48(4), 655-671. <https://doi.org/10.3758/s13421-019-00998-2>
- De keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2019). Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin*, 46(2), 204-215. <https://doi.org/10.1177/0146167219853844>
- De Neys, W. (2006). Automatic–Heuristic and Executive–Analytic Processing during Reasoning: Chronometric and Dual-Task Considerations. *Quarterly Journal of Experimental Psychology*, 59(6), 1070-1100. <https://doi.org/10.1080/02724980543000123>
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106(3), 1248-1299. <https://doi.org/10.1016/j.cognition.2007.06.002>
- De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience*, 10(2), 208-216. <https://doi.org/10.3758/CABN.10.2.208>
- De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, 28(5), 503-509. <https://doi.org/10.1177/0963721419855658>

- 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. <https://doi.org/10.3758/s13423-013-0384-5>
- DeYoung, C. G., Flanders, J. L., & Peterson, J. B. (2008). Cognitive abilities involved in insight problem solving: An individual differences model. *Creativity Research Journal*, 20(3), 278-290. <https://doi.org/10.1080/10400410802278719>
- Dominowski, R. L., & Duncan, C. P. (1964). Anagram solving as a function of bigram frequency. *Journal of Verbal Learning and Verbal Behavior*, 3(4), 321-325. [https://doi.org/10.1016/S0022-5371\(64\)80073-6](https://doi.org/10.1016/S0022-5371(64)80073-6)
- Dougherty, M. R., & Harbison, J. (2007). Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6), 1108. <https://doi.org/10.1037/0278-7393.33.6.1108>
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, 22(4), 271-280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12(3), 83-87. <https://doi.org/10.1111/1467-8721.01235>
- Duttle, K. (2016). Cognitive skills and confidence: interrelations with overestimation, overplacement and overprecision. *Bulletin of Economic Research*, 68(S1), 42-55. <https://doi.org/10.1111/boer.12069>
- Engeler, N. C., & Gilbert, S. J. (2020). The effect of metacognitive training on confidence and strategic reminder setting. *Plos one*, 15(10), e0240858. <https://doi.org/10.1371/journal.pone.0240858>

- Evans, J. R., & Fisher, R. P. (2011). Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Applied Cognitive Psychology, 25*(3), 501-508. <https://doi.org/10.1002/acp.1722>
- Evans, J. S. B., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*(4), 382-389. <https://doi.org/10.1080/13546780542000005>
- Evans, J. S. B. T. (2007). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology, 59*(1), 255-278. <https://doi.org/10.1146/annurev.psych.59.103006.093629>
- Evans, J. S. B. T. (2019). Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning, 25*(4), 383-415. <https://doi.org/10.1080/13546783.2019.1623071>
- Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 238. <https://doi.org/10.1037/0278-7393.33.1.238>
- Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13*(1), 1-17. [https://doi.org/10.1002/\(SICI\)1099-0771\(200001/03\)13:1<1::AID-BDM333>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S)
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*(4), 25-42. <https://doi.org/10.1257/089533005775196732>
- Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—in search of a phenomenon. *Thinking & Reasoning, 21*(4), 383-396. <https://doi.org/10.1080/13546783.2014.980755>

- Gilhooly, K. J. (1978). Bigram statistics for 205 five-letter words having single-solution anagrams. *Behavior Research Methods & Instrumentation*, *10*(3), 389-392.
<https://doi.org/10.3758/BF03205158>
- Gilhooly, K. J., & Fioratou, E. (2009). Executive functions in insight versus non-insight problem solving: An individual differences approach. *Thinking & Reasoning*, *15*(4), 355-376. <https://doi.org/10.1080/13546780903178615>
- Gilhooly, K. J., & Murphy, P. (2005). Differentiating insight from non-insight problems. *Thinking & Reasoning*, *11*(3), 279-302. <https://doi.org/10.1080/13546780442000187>
- Glöckner, A., & Witteman, C. (2010). Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. *Thinking & Reasoning*, *16*(1), 1-25. <https://doi.org/10.1080/13546780903395748>
- Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, *87*(1), B11-B22. [https://doi.org/10.1016/S0010-0277\(02\)00185-3](https://doi.org/10.1016/S0010-0277(02)00185-3)
- Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, *37*(7), 1001-1013.
<https://doi.org/10.3758/MC.37.7.1001>
- Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision making*, *8*(3), 188-201.
- Healy, A. F., Kole, J. A., Buck-Gengler, C. J., & Bourne, L. E. (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied*, *10*(3), 188.
- Higham, P. A. (2007). No Special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of experimental psychology: General*, *136*(1), 1-22. <https://doi.org/10.1037/0096-3445.136.1.1>

- Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review, 17*, 63-84.
<https://doi.org/10.1016/j.edurev.2015.11.002>
- Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters, 110*(2), 97-100. <https://doi.org/10.1016/j.econlet.2010.11.015>
- JASP Team. (2022). *JASP (Version 0.16.2)[Computer software]*. In <https://jasp-stats.org/>
- Johnson, D. M. (1966). Solution of anagrams. *Psychological bulletin, 66*(5), 371-384.
<https://doi.org/10.1037/h0023886>
- Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist, 58*(9), 697. <https://doi.org/10.1037/0003-066X.58.9.697>
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, & Giroux.
- Kahneman, D., Slovic, S. P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.
- Karpicke, J. D. (2009). Metacognitive control and strategy selection: deciding to practice retrieval during learning. *Journal of experimental psychology: General, 138*(4), 469.
<https://doi.org/10.1037/a0017341>
- Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1-24. <https://doi.org/10.1006/jmla.1993.1001>
- Keren, G. (2014). Between-or within-subjects design: A methodological dilemma. In *A handbook for data analysis in the behavioral sciences* (Vol. 1, pp. 257-272). Psychology Press: New York.
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627-633. <https://doi.org/10.3758/BRM.42.3.627>

- Kleitman, S., Hui, J. S.-W., & Jiang, Y. (2019). Confidence to spare: individual differences in cognitive and metacognitive arrogance and competence. *Metacognition and Learning, 14*(3), 479-508. <https://doi.org/10.1007/s11409-019-09210-x>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of experimental psychology: General, 126*(4), 349. <https://doi.org/10.1037/0096-3445.126.4.349>
- Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *The Psychological Review, 103*(3). <https://doi.org/10.1037/0033-295x.103.3.490>
- Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of experimental psychology: General, 135*(1), 36-69. <https://doi.org/10.1037/0096-3445.135.1.36>
- Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of experimental psychology: General, 131*(2), 147. <https://doi.org/10.1037/0096-3445.131.2.147>
- Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*(2), 121-132. <https://doi.org/10.1016/j.learninstruc.2011.08.004>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134. <https://doi.org/10.1037/0022-3514.77.6.1121>

- Lauterman, T., & Ackerman, R. (2019). Initial judgment of solvability in non-verbal problems – a predictor of solving processes. *Metacognition and Learning, 14*(3), 365–383. <https://doi.org/10.1007/s11409-019-09194-8>
- Lee, J. Y., & Hoffman, E. (2020). *The Effect of COVID-19 on Amazon Mturk*. Iowa State University. <https://doi.org/10.2139/ssrn.3712660>
- Leopold, C., & Leutner, D. (2015). Improving students' science text comprehension through metacognitive self-regulation when applying learning strategies. *Metacognition and Learning, 10*(3), 313-346. <https://doi.org/10.1007/s11409-014-9130-2>
- Leutner, D., Leopold, C., & den Elzen-Rump, V. (2007). Self-Regulated Learning with a Text-Highlighting Strategy. *Zeitschrift für Psychologie / Journal of Psychology, 215*(3), 174-182. <https://doi.org/10.1027/0044-3409.215.3.174>
- Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*(2), 149-171. [https://doi.org/10.1016/0030-5073\(80\)90052-5](https://doi.org/10.1016/0030-5073(80)90052-5)
- Lieberman, M. D. (2000). Intuition: a social cognitive neuroscience approach. *Psychological bulletin, 126*(1), 109. <https://doi.org/10.1037//0033-2909.126.1.109>
- Lingel, K., Lenhart, J., & Schneider, W. (2019). Metacognition in mathematics: do different metacognitive monitoring measures make a difference? *ZDM - Mathematics Education, 51*(4), 587-600. <https://doi.org/10.1007/s11858-019-01062-8>
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433-442. <https://doi.org/10.3758/s13428-016-0727-z>
- Livingston, J. A. (1997). Metacognition: An Overview. *ERIC*.
- MacGregor, J. N., & Cunningham, J. B. (2008). Rebus puzzles as insight problems. *Behavior Research Methods, 40*(1), 263-268. <https://doi.org/10.3758/BRM.40.1.263>

- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422-430. <https://doi.org/10.1016/j.concog.2011.09.021>
- Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & Cognition, 43*(4), 681-693. <https://doi.org/10.3758/s13421-014-0488-9>
- Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology, 105*(3), 353. <https://doi.org/10.1037/a0033640>
- Mayzner, M. S., & Tresselt, M. E. (1958). Anagram solution times: a function of letter order and word frequency. *Journal of Experimental Psychology, 56*(4), 376-379. <https://doi.org/10.1037/h0041542>
- Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1. <https://doi.org/10.1037/xap0000040>
- Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(4), 623. <https://doi.org/10.1037/0278-7393.12.4.623>
- Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition, 15*(3), 238-246. <https://doi.org/10.3758/BF03197722>
- Nelson, T. O., & Narens, L. (1990). *Metamemory: A theoretical framework and new findings* (Vol. 26). San Diego, CA: Academic Press.
- Noori, M. (2016). Cognitive reflection as a predictor of susceptibility to behavioral anomalies. *Judgment and Decision making, 11*(1), 114.

- Novick, L. R., & Côté, N. (1992). The nature of expertise in anagram solution. Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,
- Novick, L. R., & Sherman, S. J. (2003). On the nature of insight solutions: Evidence from skill differences in anagram solution. *Quarterly Journal of Experimental Psychology*, *56*(2), 351. <https://doi.org/10.1080/02724980244000288>
- Novick, L. R., & Sherman, S. J. (2008). The effects of superficial and structural information on online problem solving for good versus poor anagram solvers [Article]. *Quarterly Journal of Experimental Psychology*, *61*(7), 1098-1120. <https://doi.org/10.1080/17470210701449936>
- Ohlsson, S. (2011). *Deep Learning: How the Mind Overrides Experience*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511780295>
- Ohlsson, S., Keane, M., & Gilhooly, K. (1992). Information processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking*. London: Harvester-Wheatsheaf.
- Öllinger, M., Jones, G., Faber, A. H., & Knoblich, G. (2013). Cognitive mechanisms of insight: the role of heuristics and representational change in solving the eight-coin problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 931-939. <https://doi.org/10.1037/a0029194>
- Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237-241. <https://doi.org/10.1016/j.tics.2008.02.014>
- Pajares, F. (2006). Self-efficacy during childhood and adolescence. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 339-367). Information Age.
- Patel, N., Baker, S. G., & Scherer, L. D. (2019). Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. *Journal of*

experimental psychology: General, 148(12), 2129.

<https://doi.org/10.31234/osf.io/xeyj8>

Payne, S. J., & Duggan, G. B. (2011). Giving up problem solving. *Memory & Cognition*, 39(5), 902-913. <https://doi.org/10.3758/s13421-010-0068-6>

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, 42(1), 1-10. <https://doi.org/10.3758/s13421-013-0340-7>

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015a). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, 24(6), 425-432. <https://doi.org/10.1177/0963721415604610>

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, 80, 34-72. <https://doi.org/10.1016/j.cogpsych.2015.05.001>

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, 24(6), 1774-1784. <https://doi.org/10.3758/s13423-017-1242-7>

Primi, C., Donati, M. A., Chiesi, F., & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning*, 24(2), 258-279. <https://doi.org/10.1080/13546783.2017.1387606>

Qualtrics. (2019). Qualtrics [computer software]. In Provo, Utah, USA.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, 204, 104381. <https://doi.org/10.1016/j.cognition.2020.104381>

- Salvi, C., Bricolo, E., Kounios, J., Bowden, E., & Beeman, M. (2016). Insight solutions are correct more often than analytic solutions. *Thinking & Reasoning*, 22(4), 443-460.
<https://doi.org/10.1080/13546783.2016.1141798>
- Schuster, C., Stebner, F., Leutner, D., & Wirth, J. (2020). Transfer of metacognitive skills in self-regulated learning: an experimental training study. *Metacognition and Learning*, 15(3), 455-477. <https://doi.org/10.1007/s11409-020-09237-5>
- Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes*, 42(3), 271-283. [https://doi.org/10.1016/0749-5978\(88\)90001-5](https://doi.org/10.1016/0749-5978(88)90001-5)
- Shuffle Characters in Text*. (2010). Retrieved April 10th from <https://www.browserling.com/tools/random-letters>
- Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, 34(3), 619-632.
<https://doi.org/10.3758/BF03193584>
- Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in psychology*, 7, 218. <https://doi.org/10.3389/fpsyg.2016.00218>
- Skinner, N. F. (1979). Learned Helplessness: Performance as a Function of Task Significance. *The Journal of Psychology*, 102(1), 77-82.
<https://doi.org/10.1080/00223980.1979.9915097>
- Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning*, 27(1), 38-68.
<https://doi.org/10.1080/13546783.2019.1708793>

- Stagnaro, M., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision making*, *13*, 260-267. <https://doi.org/10.2139/ssrn.3115809>
- Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Erlbaum.
- Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423-444. <https://doi.org/10.1080/13546783.2018.1459314>
- Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161-188. <https://doi.org/10.1037/0096-3445.127.2.161>
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645-665. <https://doi.org/10.1017/S0140525X00003435>
- Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225-247. <https://doi.org/10.1080/13546780600780796>
- Stanovich, K. E., & West, R. F. (2008). On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking & Reasoning*, *14*(2), 129-167. <https://doi.org/10.1080/13546780701679764>
- Strudwicke, H. W., Bodner, G. E., Williamson, P., & Arnold, M. M. (2022). *Open-minded and reflective thinking predicts reasoning and meta-reasoning: Evidence from a ratio-bias conflict task*. [Manuscript under review]. Department of Psychology, Flinders University.

- Strudwicke, H. W., Christian, M. W., McLean, J. M., & Burton, O. R. (2022). *Bayesian evidence for differences between fast and slow correct responses on the Cognitive Reflection Test*. [Manuscript in preparation]. Department of Psychology, Flinders University.
- 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.
<https://doi.org/10.1080/13546783.2017.1378723>
- Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences*, *61*, 136-150.
<https://doi.org/10.1016/j.lindif.2017.11.015>
- Thompson, V., & Morsanyi, K. (2012). Analytic thinking: do you feel like it? *Mind & Society*, *11*(1), 93-105. <https://doi.org/10.1007/s11299-012-0100-6>
- Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In K. Frankish & J. St B. T. Evans (Eds.), *In Two Minds: Dual Processes and Beyond*. Oxford University Press.
- Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107-140.
<https://doi.org/10.1016/j.cogpsych.2011.06.001>
- Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237-251.
<https://doi.org/10.1016/j.cognition.2012.09.012>

- 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. <https://doi.org/10.3758/s13421-011-0104-1>
- 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. <https://doi.org/10.1080/13546783.2013.844729>
- Topolinski, S., Bakhtiari, G., & Erle, T. M. (2016). Can I cut the Gordian tkok? The impact of pronounceability, actual solvability, and length on intuitive problem assessments of anagrams. *Cognition*, *146*, 439-452. <https://doi.org/10.1016/j.cognition.2015.10.019>
- Topolinski, S., & Reber, R. (2010). Gaining insight into the “Aha” experience. *Current Directions in Psychological Science*, *19*(6), 402-405. <https://doi.org/0.1177/0963721410388803>
- Topolinski, S., & Strack, F. (2009a). The analysis of intuition: Processing fluency and affect in judgements of semantic coherence. *Cognition and Emotion*, *23*(8), 1465-1503. <https://doi.org/10.1080/02699930802420745>
- Topolinski, S., & Strack, F. (2009b). The architecture of intuition: Fluency and affect determine intuitive judgments of semantic and visual coherence and judgments of grammaticality in artificial grammar learning. *Journal of experimental psychology: General*, *138*(1), 39. <https://doi.org/10.1037/a0014678>
- Tversky, B., & Sherman, T. (1975). Picture memory improves with longer on time and off time. *Journal of Experimental Psychology: Human Learning and Memory*, *104*(2), 114-118. <https://doi.org/10.1037/0278-7393.1.2.114>
- Undorf, M., & Zander, T. (2017). Intuition and metacognition: The effect of semantic coherence on judgments of learning. *Psychonomic Bulletin & Review*, *24*(4), 1217-1224. <https://doi.org/10.3758/s13423-016-1189-0>

- Valerjev, P., & Dujmović, M. (2020). The Impact of the Length and Solvability of Anagrams on Performance and Metacognitive Judgments. In A. Tokić (Ed.), *21st Psychology Days in Zadar: Book of Selected Proceedings: International Scientific Conference* (pp. 217-230). University of Zadar, Department of Psychology.
- van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., & Haaf, J. M. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813-826. <https://doi.org/10.3758/s13423-020-01798-5>
- Weisberg, R. W. (1992). Metacognition and insight during problem solving: Comment on Metcalfe. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 426-431. <https://doi.org/10.1037/0278-7393.18.2.426>
- Weisberg, R. W. (2015). Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, *21*(1), 5-39. <https://doi.org/10.1080/13546783.2014.886625>
- Welsh, M., Burns, N., & Delfabbro, P. (2013). The cognitive reflection test: How much more than numerical ability? Paper presented at the Proceedings of the 35th Annual Meeting of the Cognitive Science Society, Berlin, Germany.
- West, J. T., & Mulligan, N. W. (2019). Prospective metamemory, like retrospective metamemory, exhibits underconfidence with practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(12), 2224. <https://doi.org/10.1037/xlm0000708>
- West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of personality and social psychology*, *103*(3), 506-519. <https://doi.org/10.1037/a0028857>

- 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.
- Word frequency: based on 450 million word COCA corpus.* (2016). Retrieved August 28 from <https://www.wordfrequency.info/>
- Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41*(2), 229-241. <https://doi.org/10.3758/s13421-012-0255-8>
- Zedelius, C. M., & Schooler, J. W. (2015). Mind wandering “Ahas” versus mindful reasoning: alternative routes to creative solutions. *Frontiers in psychology, 6*(834). <https://doi.org/10.3389/fpsyg.2015.00834>
- Zhang, Z., Lei, Y., & Li, H. (2016). Approaching the distinction between intuition and insight. *Frontiers in Psychology, 7*, 1195. <https://doi.org/10.3389/fpsyg.2016.01195>
- Zimmerman, B. J., & Schunk, D. H. (2012). *Self-regulated learning and academic achievement: Theory, research, and practice*. Springer Science & Business Media.

Appendices

Appendix A – Anagram stimuli

Solvable anagrams		Unsolvable anagrams	
Solution word	Anagram	Original pseudoword	Anagram
RISKY	YSKRI	HISOB	ISOBH
DRAFT	FADRT	DINIC	IIDCN
CABIN	BIACN	CYOYE	YCOYE
FIGHT	HFTIG	MAWER	EWMRA
UNITY	INTUY	COLER	COERL
MIDST	DSTMI	BEBER	BRBEE
PUNCH	UPHCN	ZEDAN	EZDAN
FAULT	ATLUF	FACIC	CICAF
FLAME	MLEFA	BEZER	ZEREB
DRUNK	UKRND	WOPAN	NWAOP
DEPTH	EPDHT	WESAN	AWSNE
MONTH	NMHTO	RAYER	RYAER
CLOTH	OLHTC	CAXER	ECXAR
CABLE	ELCAB	CAXER	CUSVI
MATCH	TCAHM	STILE	TSELI
CLERK	RLCKE	BONIT	OTBIN
BENCH	EHBCN	MARIC	MCARI
LOWER	ELWOR	FINAR	FIARN
MAKER	AREMK	PAWER	PREAW
PRIZE	EZRIE	WHOTE	OWHET
BLOCK	OCLKB	CALUM	CMAUL
VIRUS	SVRUI	LALER	ALLRE
FRAUD	RFADU	RALUM	MRLUA
TRUNK	KURNT	LEWIC	ECWIL
FROWN	ORFNW	MACOR	CORMA
COUNT	NOTUC	LAKOR	LAROK
GRASP	PASRG	FLISH	ISHLF
PANIC	CPINA	BEFER	REFBE
WHEAT	ETWAH	BEORE	EBOER
WOUND	DUNWO	LIXIC	IXILC
WOMAN	NOWAM	LEYIC	EYLIC
VOCAL	AOVCL	GIMAN	NMAIG
TRICK	CIRTK	GOSAN	GSANO
PLANT	APTLN	RAXER	RREXA
FLASH	HSFAL	FARIM	IAMFR
ABUSE	AESUB	BLICK	KICBL
BLACK	KACBL	PLONT	LNTPO
GRIEF	EFIGR	MINIC	MNIIC
COVER	ORECV	CAOER	AOREC
YOUTH	HOTYU	VIBUS	IBUVS

Appendix B – Experiment 1 & 2: Training group instructions

In Part 1 of the study, you will be shown a series of **5-letter word scrambles** one at a time. Some of the word scrambles can be rearranged to form an **English** word (e.g., **YIRNA = RAINY**, **WRSTI = WRIST**), whereas others cannot be rearranged to form an English word (e.g., **LYSUP**, **KBHEA**). Please keep in mind that none of the word scrambles can be rearranged to form proper nouns (i.e., people's names, names of cities, etc). In Part 2 of the study, you will be asked to try to solve each of the solvable word scrambles under a 45 second time limit.

In Part 3 of the study, you will be asked to solve some problems.

In Part 4 of the study, you will be asked to provide some general information about yourself.

On each trial, a word scramble will be shown briefly. When the word scramble disappears, you will be prompted to make an initial assessment of whether you think that word scramble was solvable (click "YES it is solvable"), or not solvable (click "NO, it is not solvable"). If you managed to find a solution within the time you were given to see the word scramble, click "I have already solved it". Once you respond the next word scramble will appear.

In the first set of trials, each word scramble will be shown for 16 seconds.

In the second set of trials, each word scramble will be shown for 8 seconds.

In the third set of trials, each word scramble will be shown for 4 seconds.

In the fourth set of trials, each word scramble will be shown for 2 seconds.

After the word scramble disappears, the response options will appear on the screen, and you will have 3 seconds to make your initial judgement of solvability. You must make your judgement within this 3 seconds! If you do not, an error message will appear reminding you to respond within 3 seconds.

You will receive 5 practice word scrambles. Each will be shown for 16 seconds, and will then be followed by the 3 response options. When the response options are shown, you must respond within 3 seconds.

In preparation for Part 2 of the study, you will receive 5 practice word scrambles to solve. You have 45 seconds to solve each one.

Please type your answer into the text box on the screen, double check the spelling, and press the arrow button to enter your response.

Appendix C – Experiment 2: No-training group instructions

In Part 1 of the study, you will be shown a series of **5-letter word scrambles** one at a time. Some of the word scrambles can be rearranged to form an **English** word (e.g., **YIRNA = RAINY**, **WRSTI = WRIST**), whereas others cannot be rearranged to form an English word (e.g., **LYSUP**, **KBHEA**). Please keep in mind that none of the word scrambles can be rearranged to form proper nouns (i.e., people's names, names of cities, etc). In Part 2 of the study, you will be asked to try to solve each of the solvable word scrambles under a 45 second time limit. In Part 3 of the study, you will be asked to solve some problems. In Part 4 of the study, you will be asked to provide some general information about yourself.

On each trial, a word scramble will be shown for a brief period. When the word scramble disappears, you will be prompted to make an initial assessment of whether you think that word scramble was solvable (click "YES it is solvable"), or not solvable (click "NO, it is not solvable"). If you managed to find a solution within the time you were given to see the word scramble, click "I have already solved it". Once you respond the next word scramble will appear.

You will complete 4 blocks of judgements about the word scrambles. In each block, the word scramble will be shown to you for only 2 seconds, and will then disappear from the screen. After the word scramble disappears, the response options will appear on the screen and you will have 3 seconds to make your initial judgement of solvability. You must make your judgement within this 3 seconds! If you do not, an error message will appear reminding you to respond within 3 seconds.

You will receive 5 practice word scrambles. Each will be shown for 2 seconds, and will then be followed by the 3 response options. When the response options are shown, you must respond within 3 seconds.

In preparation for Part 2 of the study, you will receive 5 practice word scrambles to solve. You have 45 seconds to solve each one.

Please type your answer into the text box on the screen, double check the spelling, and press the arrow button to enter your response.

Appendix D – Experiment 3: Training group instructions

In Part 1 of the study, you will be shown a series of **5-letter word scrambles** one at a time. Some of the word scrambles can be rearranged to form an **English** word (e.g., **YIRNA = RAINY**, **WRSTI = WRIST**), whereas others cannot be rearranged to form an English word (e.g., **LYSUP**, **KBHEA**). Please keep in mind that none of the word scrambles can be rearranged to form proper nouns (i.e., people's names, names of cities, etc). In Part 2 of the study, you will be asked to try to solve the word scrambles. In Part 3 of the study, you will be asked to solve some problems. In Part 4 of the study, you will be asked to provide some general information about yourself.

On each trial, a word scramble will be shown for a brief period. When the word scramble disappears you will be prompted to make an initial assessment of whether you think that word scramble was solvable (click "YES it is solvable"), or not solvable (click "NO, it is not solvable"). If you managed to find a solution within the time you were given to see the word scramble, click "I have already solved it". Once you respond the next word scramble will appear.

In the first set of trials, each word scramble will be shown for 16 seconds.

In the second set of trials, each word scramble will be shown for 8 seconds.

In the third set of trials, each word scramble will be shown for 4 seconds.

In the fourth set of trials, each word scramble will be shown for 2 seconds.

After the word scramble disappears, the response options will appear on the screen and you will have 3 seconds to make your initial judgement of solvability. You must make your judgement within this 3 seconds! If you do not, an error message will appear reminding you to respond within 3 seconds.

You will receive 5 practice word scrambles. Each will be shown for 16 seconds, and will then be followed by the 3 response options. When the response options are shown, you must respond within 3 seconds.

In preparation for Part 2 of the study, you will now have an opportunity to practice solving the word scrambles. You will be given each word scramble again, and you must try and find the solution for it.

If you find a solution for the word scramble, type your 5-letter answer into the text box on the screen, double check the spelling, and click the arrow button to enter your response.

If you think the word scramble is solvable but you are unable to solve it, type the letter P (to "pass") into the text box and then click the arrow button to enter your response.

If you think the word scramble is not solvable, type the letter N ('not solvable') into the text box and then click the arrow button to enter your response.

Appendix E – Experiment 3: No-training group instructions

In Part 1 of the study, you will be shown a series of **5-letter word scrambles** one at a time. Some of the word scrambles can be rearranged to form an **English** word (e.g., **YIRNA = RAINY**, **WRSTI = WRIST**), whereas others cannot be rearranged to form an English word (e.g., **LYSUP**, **KBHEA**). Please keep in mind that none of the word scrambles can be rearranged to form proper nouns (i.e., people's names, names of cities, etc). In Part 2 of the study, you will be asked to try to solve the word scrambles. In Part 3 of the study, you will be asked to solve some problems. In Part 4 of the study, you will be asked to provide some general information about yourself.

On each trial, a word scramble will be shown for a brief period. When the word scramble disappears you will be prompted to make an initial assessment of whether you think that word scramble was solvable (click "YES it is solvable"), or not solvable (click "NO, it is not solvable"). If you managed to find a solution within the time you were given to see the word scramble, click "I have already solved it". Once you respond the next word scramble will appear.

You will complete 4 blocks of judgements about the word scrambles. In each block, the word scramble will be shown to you for only 2 seconds, and will then disappear from the screen. After the word scramble disappears, the response options will appear on the screen and you will have 3 seconds to make your initial judgement of solvability. You must make your judgement within this 3 seconds! If you do not, an error message will appear reminding you to respond within 3 seconds.

You will receive 5 practice word scrambles. Each will be shown for 2 seconds, and will then be followed by the 3 response options. When the response options are shown, you must respond within 3 seconds.

In preparation for Part 2 of the study, you will now have an opportunity to practice solving the word scrambles. You will be given each word scramble again, and you must try and find the solution for it.

If you find a solution for the word scramble, type your 5-letter answer into the text box on the screen, double check the spelling, and click the arrow button to enter your response.

If you think the word scramble is solvable but you are unable to solve it, type the letter P (to "pass") into the text box and then click the arrow button to enter your response.

If you think the word scramble is not solvable, type the letter N ('not solvable') into the text box and then click the arrow button to enter your response.

You can choose how long you wish to spend solving each word scramble.

Appendix F – Experiment 4: Blocked design instructions

In Part 1 of the study, you will be shown a series of **5-letter word scrambles** one at a time, across **2 blocks**. Some of the word scrambles can be rearranged to form an **English** word (e.g., **YIRNA = RAINY**, **WRSTI = WRIST**), whereas others cannot be rearranged to form an English word (e.g., **LYSUP**, **KBHEA**). Please keep in mind that none of the word scrambles can be rearranged to form proper nouns (i.e., people's names, names of cities, etc) or plurals (e.g., **PILLS**, **BEANS**, **SLIPS**).

In Part 2 of the study, you will be asked to try to solve the word scrambles.

In Part 3 of the study, you will be asked to provide some general information about yourself.

On each trial, a word scramble will be shown for X seconds.

When the word scramble disappears you will be prompted to make an initial assessment of whether you think that word scramble was solvable (click "YES it is solvable"), or not solvable (click "NO, it is not solvable"). If you managed to find a solution within the time you were given to see the word scramble, click "I have already solved it". You will have 3 seconds to make your judgement. Once you respond the next word scramble will appear.

You will receive 5 practice word scrambles. Each will be shown for X seconds, and will then be followed by the 3 response options. When the response options are shown, you must respond within 3 seconds.

In preparation for Part 2 of the study, you will have an opportunity to practice solving the word scrambles. You will be given each word scramble again, and you must try and find the solution for it.

If you find a solution for the word scramble, type your 5-letter answer into the text box on the screen, double check the spelling, and click the arrow button to enter your response.

If you think the word scramble is solvable, but you are unable to solve it, type the letter P to pass and then click the arrow button to enter your response.

If you think the word scramble is not solvable, type the letter N ('not solvable') into the text box and then click the arrow button to enter your response.

You can choose how long you wish to spend solving each word scramble.

Appendix G – Experiment 4: Interleaved design instructions

In Part 1 of the study, you will be shown a series of **5-letter word scrambles** one at a time, across **2 blocks**. Some of the word scrambles can be rearranged to form an **English** word (e.g., **YIRNA = RAINY**, **WRSTI = WRIST**), whereas others cannot be rearranged to form an English word (e.g., **LYSUP**, **KBHEA**). Please keep in mind that none of the word scrambles can be rearranged to form proper nouns (i.e., people's names, names of cities, etc) or plurals (e.g., **PILLS**, **BEANS**, **SLIPS**). In Part 2 of the study, you will be asked to provide some general information about yourself.

On each trial, a word scramble will be shown for X seconds.

When the word scramble disappears you will be prompted to make an initial assessment of whether you think that word scramble was solvable (click "YES it is solvable"), or not solvable (click "NO, it is not solvable"). If you managed to find a solution within the time you were given to see the word scramble, click "I have already solved it". You will have 3 seconds to make your judgement.


After you make your judgement, the word scramble will reappear, and you must try and find a solution for it. If you find a solution for the word scramble, type your 5-letter answer into the text box on the screen, double-check the spelling, and click the arrow button to enter your response. If you think the word scramble is solvable, but you are unable to solve it, type the letter P to pass and then click the arrow button to enter your response.

You will now receive 5 practice word scrambles. Each will be shown for 2 seconds, and will then be followed by the 3 response options. When the response options are shown, you must respond within 3 seconds. After selecting your response, you will be asked to solve each word scramble.

Appendix H – 7-item Cognitive Reflection Test (CRT; Toplak et al., 2014)


1. A bat and a ball cost \$1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost? ____ cents
[Correct answer = 5 cents; intuitive answer = 10 cents]
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? ____ minutes
[Correct answer = 5 minutes; intuitive answer = 100 minutes]
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ____ days
[Correct answer = 47 days; intuitive answer = 24 days]
4. If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together? ____ days
[correct answer = 4 days; intuitive answer = 9]
5. Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class? _____ students
[correct answer = 29 students; intuitive answer = 30]
6. A man buys a pig for \$60, sells it for \$70, buys it back for \$80, and sells it finally for \$90. How much has he made? ____ dollars
[correct answer = \$20; intuitive answer = \$10]
7. 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
 [correct answer = c because the value at this point is \$7,000; intuitive response = b].


Appendix I – Example Judgement of Solvability trial



Flinders University

BIACN

Powered by Qualtrics 




Flinders University

Respond within 3 seconds:


YES it is solvable

NO it is not solvable

I have already solved it

Powered by Qualtrics 


Appendix J – Example solving trials


 Flinders University

UKRND

drunk

→


Powered by Qualtrics 

 Flinders University

EWMRA

N

→

Powered by Qualtrics 

Appendix K – Debriefing form (all experiments)

In the current study we are interested in how well people judge whether a word scramble can be solved. We are also interested if peoples' judgements relate to whether they will solve the word scramble later on. If peoples' judgements do relate to whether they will solve the word scramble, then it will be important for future research to look at why this happens.

If you feel fatigued or discouraged by the task, please keep in mind that this task was intentionally designed to be difficult. Thus, it is normal to feel fatigued, and please do not feel disheartened if you found the task challenging because we expected it to be such.

Important: It is important that you do not share the information we provide in this section with anyone who has not yet completed the project. If someone knows ahead of time what we are looking at, and why, then that will likely change their responses and influence the data in a way that will harm the project.

We thank you for your cooperation in not releasing this information to any potential participants in this study.

Appendix L – Mechanical Turk Virtual Task Description (all experiments)

WORD SCRAMBLES: In this job you will be asked to solve a series of word scrambles.

Your task is to make a solvability judgement, and then to rearrange the letters until you find the solution word. It should take you no longer than 30 minutes to complete. Payment is 25 cents, but if you follow all instructions you will receive a \$1.75 bonus.