DEVELOPING AND APPLYING TOOLS FOR EXPERIMENTAL ECONOMIC RESEARCH

Social cohesion, Other-regarding preferences & Creativity

MALTE BAADER

Thesis submitted to The University of Nottingham
for the Degree of Doctor of Philosophy

School of Economics
University of Nottingham

February 2021

Abstract

This thesis contributes to the evaluation and development of experimental methodologies within experimental economics. Across two parts and five chapters, I outline three novel experimental instruments to measure *social cohesion*, *distributional preferences* and *creativity*. First, I provide a general introduction that gives an overview of the main topic of the thesis and summarises each chapter.

The focus of Part I is social cohesion and distributional preferences. In Chapter 1, we propose a condensed version of a prominent methodology to estimate distributional preferences based on allocation decisions in repeated dictator games (Fisman et al., 2007). We show that we can reduce the total number of decisions by 60% whilst maintaining high accuracy in the estimation of distributional preference parameters. The developed methodology successfully reduces cognitive subject burden and shortens elicitation time by 50%, therefore substantially improving the efficiency of the original instrument.

In Chapter 2, we develop a more nuanced version of the 'Inclusion of Others in Self' (IOS) scale, an established methodology to measure social cohesion. By extending the answer range and creating a computerised interface, we are able to increase measurement accuracy whilst simplifying the experimental implementation. Moreover, we also conduct a detailed replication of Gächter et al. (2015) supporting the robustness of our proposed tool.

In Chapter 3, finally, we apply both instruments developed in Chapter 1 and 2 by investigating the relationship between *social cohesion*, *distributional preferences* and *altruistic giving* in a network of university students. We find that social cohesion significantly affects altruistic giving with distributional preferences serving as a fundamental mediating factor.

Shifting the thesis' focus, Part II explores topics related to experimental *creativity* research. In Chapter 4, we compare five experimental creativity tasks across two studies in a within-subject design. We find no evidence that the examined creativity tasks elicit a *common* underlying creative ability. Moreover, across both studies there is no relationship between survey measures of creativity and performance in experimental tasks.

Chapter 5 builds on these results by proposing a novel creativity task, focusing on *creative associative thinking* and substantially improving experimental properties. Our proposed method elicits two types of associative thinking that we benchmark against two established creativity tasks. We find that performance in our proposed tasks significantly correlates to their established counterpart, while behaviour in the tasks as well as incentive effects differ between the two types of associative thinking.

Finally, Part II concludes with a brief summary of the previous chapters and discusses potential applications of our developed creativity tasks for future research.

Acknowledgements

First and foremost, I want to express my deepest gratitude to my supervisors Chris Starmer and Fabio Tufano. From the first day of my PhD they have supported me with their guidance and expertise. Despite my interest in a topic *uncommon* for a behavioural economist, they encouraged my pursuit and inspired my research. Our countless meetings and informal chats are only a small piece of evidence of the mentoring I received over the course of the years.

I also want to thank Urs Fischbacher and Simon Gächter, both collaborators of chapters within this work. Being exposed to the thought processes of experienced, accomplished academics has developed my scientific mind more than any degree before. This is also why I wish to thank all other CeDEx and NIBS members for providing me with an environment that could not have been more stimulating for my own research and taught me so much as an academic. CeDEx will always be a part of my academic journey and I can call myself lucky that it is.

Without the financial support from the School of Economics, CeDEx and NIBS over the past years, it would have not been possible to complete this thesis. Even though I still do not understand why they provided it, I am very grateful for their funding. I am further thankful to all the friends I made, the laughs we shared and the support we gave each other. Amongst many, I like to thank in particular, Sarah Bowen, Candice Ford, Richard Mills, Jose Guinot Saporta and Yuliet Verbel Bustamante for making my time in Nottingham so memorable.

In closing, I want to thank my parents and siblings for their neverending support that led up to this point. Lastly, I want to thank my partner, Anna Hochleitner, for making me focus on what is truly important, for the encouragement when I was hesitant and enlightening discussions when I felt stuck. She has been the backbone of my work over the years and this thesis would not be how it is without her.

Contents

General introduction 1

Part I 6

Introduction 7

1 Testing a condensed methodology to estimate distributional preferences 9
- 1 Introduction 9
- 2 The original methodology 11
- 3 Finding an optimal design 12
  - 3.1 Simulations 12
  - 3.2 CCEI score 17
  - 3.3 Our proposed implementation 18
- 4 Validating with the *American Life Panel* data 18
  - 4.1 Replicate simulations 19
  - 4.2 Demographic predictors of preference parameters 20
- 5 Own data collection 21
  - 5.1 Design and procedure 21
  - 5.2 Results 23
- 6 Conclusion 26

Bibliography 27

2 Developing an extended version of the 'Inclusion of Others in Self' scale 29
- 1 Introduction 29
- 2 The extended IOS scale 31
- 3 Validating the extended IOS 33
- 4 Conclusion 36

Bibliography 37

3 The role of social cohesion as a determinant of altruistic behaviour 39
- 1 Introduction 39
- 2 Design 41
  - 2.1 Constructing a social network 41
  - 2.2 Measuring cohesion, distributional preferences and altruistic giving 42
  - 2.3 Treatment conditions 42
  - 2.4 Hypotheses 44
- 3 Procedure 45
  - 3.1 Experimental waves 45
  - 3.2 Recruitment 46
- 4 Results 47
  - 4.1 Network characteristics and reported social cohesion 47
  - 4.2 Within-individual robustness of measures 49
  - 4.3 Social cohesion and altruistic giving 50
  - 4.4 Social cohesion and distributional preferences 52
  - 4.5 Distributional preferences and altruistic giving 54
  - 4.6 Mediation analysis 56

4.7 Replication of Leider et al. (2009) 57
5 Conclusion 58
Bibliography 60

PART II **62**

Introduction **63**

4 Investigating experimental measures of creativity **65**
1 Introduction 65
2 Study 1 - Comparing *general* creativity tasks 68
2.1 Design 68
2.2 Results 70
3 Study 2 - Assessing creative problem-solving 73
3.1 Design 73
3.2 Results 76
4 The predictive power of self-reported creativity 78
5 Addressing measurement error 80
6 Conclusion 80
Bibliography 82

5 Measuring creativity: a network exploration task **87**
1 Introduction 87
2 The semantic network 89
3 Developing and benchmarking our associative thinking tasks 92
3.1 Local Search and Unusual Uses Task 93
3.2 Depth Search and Tower of London 95
3.3 Procedure and incentives 99
3.4 Hypotheses 100
4 Results 101
4.1 Local Search and Unusual Uses Task 101
4.2 Depth Search and Tower of London 103
4.3 Correlation across pairs of tasks 106
4.4 The role of incentives 108
5 Conclusion 109
Bibliography 111

Applications **113**

APPENDICES **117**
A Appendix for Chapter 1 117
A.1 Additional analyses 117
A.2 Experimental instructions 121
B Appendix for Chapter 2 127
B.1 Experimental instructions 127
C Appendix for Chapter 3 138
C.1 Additional analyses 138
C.2 Experimental instructions 141
D Appendix for Chapter 4 151

D.1 Additional analyses 151
D.2 Experimental instructions 155
D.3 Creativity tasks 170
E Appendix for Chapter 5 180
E.1 Additional analyses 180
E.2 Experimental instructions 183
E.3 Creativity tasks 192

List of Figures

Part I 6

Figure 1.1 Graphical interface for allocation decisions 11
Figure 1.2 Median euclidean distance for different sub-sample sizes 13
Figure 1.3 Mean accuracy for original estimations of ρ across sub-samples sizes 14
Figure 1.4 Mean euclidean distances of ρ for varying underlying distributions 15
Figure 1.5 Mean euclidean distances for bimodal, distributed draw 16
Figure 1.6 CCEI scores for original, simulated and random data 17
Figure 1.7 Simulation results for different sub-sample sizes (*ALP*) 19
Figure 1.8 Correlation of regression coefficients across methodologies 20
Figure 1.9 Overview of experimental design 22
Figure 1.10 Cumulative distributions of α and ρ parameters across treatments 23
Figure 1.11 *Control* and *Treatment 20 + 30* - pooled data 24
Figure 1.12 *Treatment 20 & 20*: Within-subject comparison 25
Figure 2.1 Graphical interface of the extended IOS scale 31
Figure 2.2 IOS Scores for different relationship levels and elicitation methods 34
Figure 3.1 Hypothesised relationship between cohesion, preferences and giving 39
Figure 3.2 Overview of treatment conditions 43
Figure 3.3 Hypothesis regarding social cohesion, preferences and giving 44
Figure 3.4 Experimental overview 45
Figure 3.5 Line and star network cases of $SD = 2$ 46
Figure 3.6 Elicited social networks 48
Figure 3.7 Heterogeneity of IOS Scores 48
Figure 3.8 Passed tokens in dictator games 50
Figure 3.9 IOS Scores for matched subjects in wave 2 and 3 ($N = 63$) 51
Figure 3.10 Distribution of preference parameters across social distances 53
Figure 3.11 Observed vs predicted dictator game giving 55
Figure 3.12 Overview and results of the mediation analysis 56

Part II 62

Figure 4.1 Relationship between creativity tasks of Study 1 72
Figure 4.2 Example of the Tower of London 74
Figure 4.3 Relationship of *closed* and *open* problem-solving 77
Figure 5.1 The semantic network 90
Figure 5.2 A snapshot of the semantic network 91
Figure 5.3 Proposed associative thinking tasks 92
Figure 5.4 Visual comparison of *Local Search* and *UUT* 95
Figure 5.5 Sub-network creation for *Depth Search* (seed: *River*) 96
Figure 5.6 Visual comparison of *Depth Search* and *ToL* 98
Figure 5.7 Experimental overview 99
Figure 5.8 Most common submissions in the *Local Search* task and the *UUT* 101
Figure 5.9 *Fluency* scores across problems and tasks 102
Figure 5.10 Relationship between the *Local Search* task and the *Unusual Uses Task* 103

Figure 5.11 Completion times across problems in the *Depth Search* task and the *ToL* 104
Figure 5.12 Relationship between the *Depth Search* task and the *Tower of London* 105
Figure 5.13 Submission time patterns across tasks 106
Figure 5.14 Relationship within associative thinking and creativity tasks 107
Figure 5.15 Incentive effects across tasks 108

List of Tables

Part I 6

Table 1.1 Overview of treatment comparisons 24
Table 2.1 *Oneness*: 'Inclusion of the Other in the Self' (IOS) scale & We scale 29
Table 2.2 *IOS scale*: original, extended, recoded 32
Table 2.3 Experimental overview 33
Table 2.4 Correlations to other scales: IOS, Oneness, Extended IOS 35
Table 3.1 Within-subject correlations across preference parameters 49
Table 3.2 The relationship of IOS and other-regarding preferences 53
Table 3.3 The predictive power of other-regarding preferences 55

Part II 62

Table 4.1 Descriptives - Study 1 70
Table 4.2 Correlations across performance metrics in the UUT 71
Table 4.3 Correlation across task duration 71
Table 4.4 Correlation across creative tasks 73
Table 4.5 Descriptives - Study 2 76
Table 4.6 Correlations across performance metrics 77
Table 4.7 Regression model for *closed* and *open* problem-solving 78
Table 4.8 Relationship between creative tasks and self-reports 79
Table 5.1 Problem overview of the *Local Search* task and the *Unusual Uses Task* 94
Table 5.2 Problem overview of the *Depth Search* task and the *Tower of London* 97
Table 5.3 Descriptives - Depth Search and Tower of London 104

General introduction

A common saying states that "*a carpenter is only as good as their tools*". Despite its unclear origin, I strongly think that the saying bares a lot of meaning.[1] As experimental scientists, we frequently employ tools 'off the shelf' in order to best answer our research questions. Every time we design an experimental study, we consider which tools and tasks are best suited for our purpose, often relying on established approaches from the literature. However even if novel experimental methodologies are employed, most studies only argue verbally in their favour and rely on intuitive consensus among readers without providing data driven evidence of the validity of the instruments. The practice can ultimately lead to the utilisation of methodologies without thoroughly questioning their components or exploring possible improvements.

This is of course not true for all methodological tools. One counter example is the elicitation of risk preferences using questionnaires. Succeeding an experimental economic tradition of using incentivised lottery tasks as measures for risk preferences (e.g. Gneezy & Potters, 1997; Holt & Laury, 2002), there has been an increasing number of studies demonstrating that self-reported levels of risk taking correlate significantly with choices in lottery tasks (see Dohmen et al., 2011; Vieider et al., 2015; Falk et al., 2018, among others). This is a great example of how data driven evidence can provide alternatives to comparably budget and time intensive lottery tasks.[2] The process of evaluating and improving established methodologies therefore fundamentally enhances scientists' ability to execute their research designs. Another example of how methodologies are being reviewed in experimental economics, is Charness et al. (2012, 2016, 2017, 2013b,a, 2018, 2021), who, in a series of papers, provide taxonomies of experimental methods for varying purposes. Whilst comparing within- and between-subject designs, incentive mechanisms, framing of instructions as well as proposing extra-laboratory approaches, they also summarise existing tasks to measure risk preferences, effort levels and beliefs. In particular, the last three papers surround measurement instruments and are thus especially relevant for this thesis. In these papers, the authors outline strengths and weaknesses of various experimental tools to ultimately formulate recommendations of their use.[3]

The testing, evaluating and improving of experimental methodologies is also at the heart of this thesis. Based on joint work with Chris Starmer, Fabio Tufano, Simon Gächter and Urs Fischbacher, I develop and validate three new methodological tools and outline applications of them. Overall, the thesis is divided into five chapters across two distinct parts. The first two chapters of each part are methodological contributions for a specific topic. Part I focuses on *social cohesion* and *other-regarding preferences*, while Part II surrounds the study of *creativity*. In addition, the third chapter of Part I provides an application of the tools developed in Chapter

[1] One possible explanation is that it is a variation of the saying "A man is only as good as his tools" (Emmert Wolf).

[2] Due to the fact that questionnaires are not incentivised, there is still an ongoing debate as to what is a better measure. However, existing research clearly suggests a relationship between self-reported and revealed risk preferences.

[3] This relevance is also reflected in other social sciences such as psychology or sociology, both having multiple dedicated journals (e.g. Psychological Methods, Sociological Methods & Research) for developments in methodologies, where in both cases some of these are amongst the journals with the highest impact factors in the discipline.

1 and 2 as a proof of concept. Similarly, to conclude Part II, I summarise three examples of possible applications for our developed method to measure creativity. The following paragraphs summarise the key aspects of each part and chapter. Both parts are self-contained, providing a motivation on the respective topics, outlining the relevant literature and introducing the chapters.

Part I: Social cohesion & Other-regarding preferences

In Chapter 1 we present an improved methodology to estimate distributional preferences based on Fisman et al. (2007). We show that the number of allocation decisions completed by subjects can be reduced by 60% without loss of accuracy in the estimations of two individual preference parameters. We validate our task using simulations based on published data as well as an additional data collection, confirming our proposed instrument. Our tool enables researchers to elicit nuanced measures of distributional preferences in a more time- and cost effective way.

Chapter 2 develops and validates a more efficient instrument to capture *social cohesion* inspired by the IOS scale (Aron et al., 1992; Gächter et al., 2015). We provide compelling evidence that by extending the answer range and adjusting the graphical interface of the instrument, we can elicit *social cohesion* with a higher precision. In particular, we find that our extended IOS scale correlates stronger with a battery of survey measures eliciting *social cohesion* than the original IOS scale.

In Chapter 3, both of our developed tools are applied in a proof-of-concept to investigate the relationship between *social cohesion*, *distributional preferences* and *altruistic giving*. To study the impact of social cohesion, we elicit a real-world friends network and use it to exogenously vary the social distance between players in allocation decisions. We find evidence that social cohesion does affect distributional preferences as well as altruistic giving. Moreover, in a mediation analysis we show that estimated distributional preferences are a crucial mediator between social cohesion and altruistic giving.

Part II: Creativity

In Chapter 4, we use a within-subject design to experimentally compare performances in five established creativity tasks across two studies. In Study 1, we find no systematic relationship between three measures of *general* creativity, suggesting that different creativity tasks do elicit different underlying abilities. In Study 2, we then hold the domain of creativity constant by focusing on creative problem-solving, whilst varying the degree of *task openness* (as in Charness & Grieco, 2019). We again find no within-subject correlation between performances, indicating that *task openness* is a key feature of experimental creativity tasks.

Chapter 5 addresses the results from Chapter 4 by outlining the development of a novel instrument to measure creative ability. We thereby focus explicitly on associative thinking ability as a key domain of creativity. Our new method improves on a number of experimental properties compared to established creativity tasks by utilising a pre-defined solution space. We achieve this by constructing a semantic network that serves as the underlying structure of our instrument. Based on this network, subjects perform two associative thinking tasks, *Local Search* and *Depth Search*. We characterise each by relating it to an established measure

of creativity, finding that performance in our proposed tasks is significantly related to their matched creativity task, whilst improving their implementation and performance evaluation.

Finally, I conclude Part II by providing brief descriptions of three possible applications of our developed associative thinking tasks. The applications demonstrate how the instrument can be applied in a variety of ways and how our tool enables researchers to integrate creativity research with behavioural economic research questions.

Bibliography

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), 596.

Charness, G., Alekseev, A., & Gneezy, U. (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization*, 134, 48–59.

Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, 131, 141–150.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149, 74–87.

Charness, G., Gneezy, U., & Imas, A. (2013a). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, 87, 43–51.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1–8.

Charness, G., Gneezy, U., & Kuhn, M. A. (2013b). Experimental methods: Extra-laboratory experiments-extending the reach of experimental economics. *Journal of Economic Behavior & Organization*, 91, 93–100.

Charness, G., Gneezy, U., & Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189, 234–256.

Charness, G. & Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17(2), 454–496.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3), 522–550.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4), 1645–1692.

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5), 1858–1876.

Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the 'inclusion of the other in the self' scale. *PloS one*, 10(6), e0129478.

Gneezy, U. & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2), 631–645.

Holt, C. A. & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.

Vieider, F. M., Lefebvre, M., Bouchouicha, R., Chmura, T., Hakimov, R., Krawczyk, M., & Martinsson, P. (2015). Common components of risk and uncertainty attitudes across contexts and domains: Evidence from 30 countries. *Journal of the European Economic Association*, 13(3), 421–452.

Part I

Social cohesion & Other-regarding preferences

Introduction

Understanding what makes a group of people cooperate or voluntarily help each other has always been one of the questions at the heart of behavioural economics, as it is a crucial building block that shapes our social communities (Henrich & Henrich, 2006). Previous research has shown that there is substantial heterogeneity in the extent to which individuals display altruistic behaviour towards others (Andreoni & Miller, 2002; Fisman et al., 2007). Understanding what underlies this heterogeneity and identifying its effects on the allocation and distribution of resources is thus crucial for learning about how economic as well as non-economic transfers take place in communities. Despite the experimental and theoretical advances exploring other-regarding preferences, the question of which factors constitute distributional preferences and how these relate to altruism is not yet fully uncovered. One aspect that I consider a key determinant of resource sharing is social cohesion. Social cohesion describes the extent to which individuals in a group feel *close* to each other, and thus to the group. In the case of one-to-one relationships social cohesion is also discussed under the term of *relationship closeness* and for the purpose of this thesis, I will use both terms as synonyms.[1] *Social cohesion* as considered here exceeds the notion of similarity and is a consequence of people's interactions, activities experienced together, and the influence people have on one another (Kelley et al., 1983). Understanding the role of social cohesion as a determinant of altruistic actions among communities and social networks can therefore provide substantial insights into our knowledge of the emergence and termination of altruistic behaviour.

Any experimental investigation into a relationship between distributional preferences and social cohesion, first requires a careful selection and examination of methodologies to elicit each of these elements. To this end, in Chapter 1 we contribute to the current experimental toolbox used to elicit other-regarding preferences. We propose and validate an improved version of the experimental approach introduced by Andreoni & Miller (2002) and later extended by Fisman et al. (2007). Our methodology substantially reduces the time and budget intensity of the original tool whilst accurately estimating distributional preference parameters. Thus, we provide the experimental scientific community with a new instrument that allows researchers to explore nuanced distributional preferences in more environments than currently possible.

Continuing our pursuit of improved measurement tools, in Chapter 2, we develop and validate a novel methodology capturing *relationship closeness*. By extending the established IOS scale (Aron et al., 1992; Gächter et al., 2015), we show significant improvements in its measurement quality whilst maintaining its ease of implementation.

In the subsequent chapters we thus propose and validate more efficient measurements of both concepts of interests. Each of them first describes the underlying existing measure, discusses its strengths and weaknesses and then proposes and validates the new, improved tool of the respective methodology.

[1] See Chan et al. (2006) for a review on the concept of social cohesion.

Finally, as as a proof of concept, we then apply both of these tools in an experimental design in Chapter 3, directly addressing the research question introduced above. Utilising both developed tools to explore the relationship between social cohesion, distributional preferences and altruism does not only contribute to our understanding of these elements but also demonstrates the substance and usefulness of both our tools for researchers in the future. We see these tools not only as assisting us in investigating the relationship between *social cohesion* and *other-regarding preferences*, but also as instruments that can be utilised by experimental researchers to implement their research designs as efficiently as possible. Each of the following three chapters is self-contained including motivation, design, hypotheses and results.

Chapter 1

Testing a condensed methodology to estimate distributional preferences

1 Introduction

Exploring altruistic behaviour and its determinants has a long history in behavioural and experimental economics (see Forsythe et al., 1994; Andreoni & Miller, 2002; Charness & Rabin, 2002; Camerer, 2003; List, 2007; Blanco et al., 2011, among others). Numerous studies have shown that individuals do not only maximise their own income but voluntarily relocate resources to other people. A natural question has thus been how to best capture such preferences theoretically and empirically. One of the most employed theoretical specifications capturing other-regarding preferences is a CES utility function. In its commonly used formulation, it captures two dimensions of preferences, non-selfishness as well as an equity-efficiency trade-off, by means of two independent parameters.[1] Inspired by the seminal work of Andreoni & Miller (2002) a popular methodology to precisely estimate these preference parameters has been put forward by Fisman, Kariv & Markovits (2007, henceforth AM/FKM). In their methodology, the authors extend the experimental design of Andreoni & Miller (2002) and ask subjects to make 50 allocation decisions, where they distribute tokens between themselves and another player on a pre-defined budget line. In each round prices are randomly drawn, thereby affecting the slope of the budget line and varying the cost of passing a token to the other player. Using the allocation decisions, as well as information on prices, Fisman et al. (2007) then calibrate both preference parameters of the CES utility function that best explain the choice data. In their specification, the first parameter, α, captures the relative weight for own earnings and the other, ρ, describes the equity-efficiency trade off. To demonstrate the predictive power of this method, subsequent studies have related both parameters to other individual characteristics as well as exogenous shocks. For example Fisman et al. (2015b) show that more elite students have systematically stronger preferences for efficiency than non-elite students. Moreover, in another study Fisman et al. (2015a) also show that distributional preferences have been significantly affected by the financial crisis in 2008. The same authors further relate estimated distributional preferences to voting behaviour (Fisman et al., 2017), finding more equity concerned individuals to be more likely to vote for the US Democratic party.[2] These studies do

This chapter is based on joint work with Simon Gächter, Chris Starmer & Fabio Tufano. I thank the CeDEx members and the participants of the CCC Conference 2020 for valuable inputs. I am grateful for funding by the British Academy.

[1] In line with the original authors, we use the term efficiency throughout the thesis. However, when examining the exact trade off it is not *efficiency* that is elicited by the tool but rather a *maximisation of the sum of tokens*.

[2] see Choi et al. (2007), Jakiela (2013), Ahn et al. (2014) and Choi et al. (2014) for a number of applications using the methodology.

not only demonstrate the external validity of distributional preferences but more specifically showcase the robustness and relevance of the AM/FKM methodology employed.

Nuanced measurements of individual characteristics have gained in popularity as over the past years experiments in economics have increasingly focused on understanding underlying mechanisms of observed behaviour. One consequence of this development has been a rising need to collect additional individual level characteristics to isolate treatment effects more thoroughly. Thus, including auxiliary measures of distributional preferences in a study could often provide substantial additional insights and even be crucial to identify underlying causal effects. In light of the increasing importance for accurate, auxiliary measures of behavioural mechanisms, in this chapter, we develop and test a condensed version of the AM/FKM methodology that provides a more budget- and time-efficient version of the original whilst sustaining its high degree of measurement accuracy.

Despite the obvious value of such nuanced measures and the proven generalisability and quality of the AM/FKM method, there have not been many studies employing the Fisman et al. (2007) methodology until now.[3] We think this is not driven by the quality of the tool but mainly due to the substantial budget and time requirements of the method. Using 50 independent allocation decisions, the original procedure is simply not suitable to be implemented as an auxiliary experimental measure or as a repeated measure in a within-subject design. Generally, increasing the number of tasks and elicitations in an experiment is directly associated to increases in length, budget and logistics required to execute an optimal experimental design. Thus, optimising methodological tools is a key endeavour to allow researchers to execute their designs exactly as intended. Our goal for this chapter is thus to simplify the AM/FKM method and transform it into an efficient and applicable auxiliary measure of social preferences. Some dimensions to consider when assessing the efficiency of a task are *completion time*, *budget required*, *ease of comprehension*, *feasibility of implementation*, *participant burden*, and the *accuracy of elicitation*. Naturally, the most important feature of an experimental tool is the latter, thus to provide an accurate measure of the underlying concept and an instrument should never be considered efficient without being accurate. That being said, improvements on any of the other dimensions whilst ensuring accuracy can substantially enhance a tool's efficiency.

In the pursuit to propose such an improved tool for the elicitation of other-regarding preferences, we show that we can achieve accurate estimates of α and ρ for as few as 20 allocation decisions, in particular, when more choices are made on relatively steep as well as flat budget lines estimation accuracy improves significantly. Our proposed method therefore reduces the number of allocation decisions from 50 to 20, decreases elicitation time and budget required by about 40% including instructions and further lessens the econometric calibration time by around 15%. To develop and validate our instrument, we follow multiple steps. First, we conduct simulations based on the original data, which we then replicate on data from a repre-

[3] In fact, when inspecting other applications of the methodology, they mostly explore the estimated preferences themselves, rationalisability or relationships to demographics. Thus, they usually do not include any additional experimental measurements or within-subject considerations.

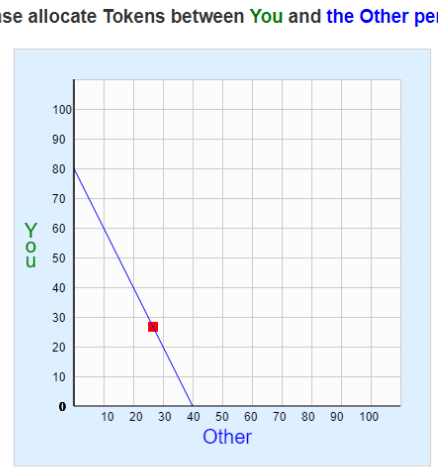
sentative sample of the US. Ultimately, we also provide evidence from an own data collection supporting our proposed condensed methodology.

The remainder of this chapter is structured as follows. In Section 2, we briefly outline the original methodology and the key ingredients of the estimation technique. In Section 3, we present our results from the simulation exercise using the original data by Fisman et al. (2007). In Section 4, we replicate our previous findings using a larger, more diverse and representative sample from the US. In Section 5, we discuss the design and results of our own data collection before concluding remarks in Section 6.

2 The original methodology

The first version of the methodology in question was put forward by Andreoni & Miller (2002). The authors capture other-regarding preferences in line with a CES utility function and explore whether observed choice behaviour is rationalisable. In their study each subject faces either 8 or 11 modified dictator games of the same form. For each decision subjects have to divide m tokens between themselves and another unknown participant such that $\pi_s + p\pi_o = m$, where m is the total endowment, π_s and π_o refer to the payoff for self and other respectively and p is the relative price of π_o. p is pre-determined but varied between each decision such that the authors could explore whether well-behaved preferences can explain the choice data. Later, Fisman et al. (2007) built upon the study of Andreoni & Miller (2002) by shifting their focus away from the question whether altruistic preferences are rationalisable and towards estimating a precise utility function for giving. In their study, the authors increase the number of allocation decisions to 50 for which subjects make a choice on a budget line using an interactive graphical display (see Figure 1.1).[4] Contrasting the Andreoni & Miller (2002) study, prices and thus the slopes of the budget lines in the graphical interface are now drawn at random. The only

Figure 1.1: Graphical interface for allocation decisions



[4] Code and details regarding the implementation are available upon request. The programme and interface is based on previous work by Kyeongtae Lee.

constraint is that at least one maximum possible allocation on one of the axes must be above 50, but no axis can exceed 100 tokens. Given these properties, the slope of the budget line defines the relative price of giving p_o/p_s (labelled the *price ratio*). The changing price ratios imply that in each dictator game subjects have to trade off maximising their own payoff together with equity and efficiency concerns. The authors then move on to use subject's allocation decisions as well as information on prices in a Tobit maximum likelihood estimation to calibrate the CES utility function as specified in Andreoni & Miller (2002)

$$U_s(\pi_s, \pi_o) = [\alpha(\pi_s)^\rho + (1-\alpha)(\pi_o)^\rho]^{\frac{1}{\rho}}. \tag{1}$$

In equation 1, $0 \leq \alpha \leq 1$ describes the relative weight on own payoff, thus the higher α the more an individual values their own payoff. $-\infty \leq \rho \leq 1$ on the other hand captures the equity-efficiency trade-off. A low ρ captures Leontief preferences[5], whereas a high ρ indicates concerns for maximising the sum of payoffs and thus efficiency. Fisman et al. (2007) observe heterogeneous preferences for both parameters where most subjects do not allocate the maximal possible amount to themselves. However, the authors also find no case of $\alpha < 0.5$, indicating that all subjects place as much weight on their own payoff as on the payoff of the other player.[6] Finally, the authors find that the rationalisability of subject's decisions significantly exceeds random choice, thereby confirming Andreoni & Miller (2002) despite the increase in the number of allocation decisions.

3 Finding an optimal design

Staying true to the methodology by Andreoni & Miller (2002) and Fisman et al. (2007), in the following sections we test whether the AM/FKM methodology can indeed be optimised by reducing the total number of decisions following a three step process:

1. Use the original data from Fisman et al. (2007) and simulate task environments that consist of fewer allocation decisions.
2. Replicate all simulations with rich data from the *American Life Panel* (ALP).
3. Collect experimental data to compare our proposed with the original methodology, allowing us to control additional aspects that cannot be addressed in simulations.

In the following sections, we will present our approach as well as the results for each of these steps sequentially.

3.1 Simulations

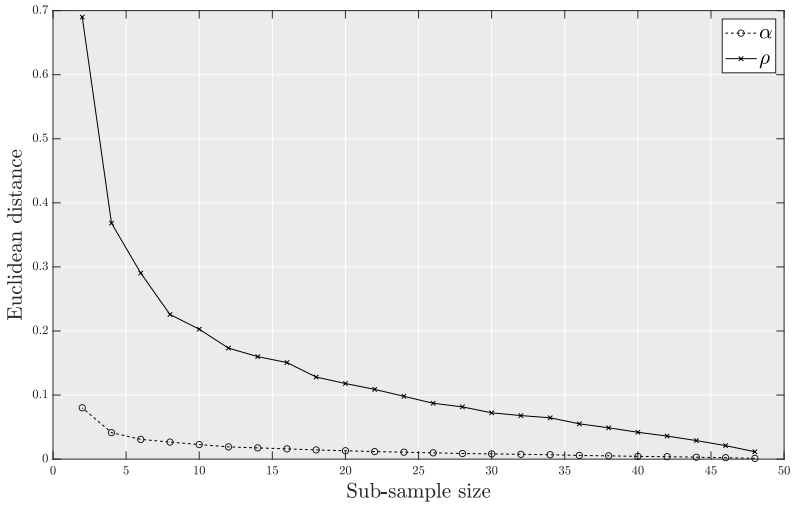
As a first step to explore whether we can reduce the number of decisions whilst ensuring accurately estimated parameters, we use the original data by Fisman et al. (2007) and randomly

[5]Preferences that maximise an equitable outcome, usually illustrated with L-shaped indifference curves.
[6]For more details on the methodology as well as results please see Fisman et al. (2007) and Andreoni & Miller (2002).

draw differently sized sub-samples of the 50 allocation decisions. We thus simulate a scenario where each subject had participated in fewer allocation decisions and estimate both preference parameters based on this subset. We then analyse their accuracy defined as the euclidean distance to the original parameter as estimated by Fisman et al. (2007), hence $accuracy_i = |i_{original} - i_{sub-sample}|,\ i = \alpha, \rho$.[7] Instead of ex-ante specifying a precise number of allocation decisions and evaluating its accuracy, we explore the accuracy of 24 different sub-sample sizes (48 to 2, reducing allocation decisions by 2 in each step). For each subject (N = 76) we draw 100 random sub-samples in each of the sizes and estimate preference parameters for each draw. For example, consider sub-sample size 20, for a given subject we randomly select 20 out of the original 50 allocation decisions and estimate preference parameters based on these 20 decisions. We then repeat this procedure 100 times for each subject and sub-sample size.

Figure 1.2 plots the *median* accuracy for all randomly drawn sub-sample sizes with euclidean distance as our measure of accuracy on the vertical and the sub-sample sizes on the horizontal axis. A lower euclidean distance, thus an observation closer to the horizontal axis indicates higher degrees of accuracy. When examining Figure 1.2, the first result that stands out is that precision increases as the sub-sample size moves closer to the original methodology (n = 50). Moreover, we also see that estimates of α (dashed) are significantly more accurate than the ones of ρ (solid) as we reduce the number of decisions. This is however intuitive, as the range of the α parameter ($0 \leq \alpha \leq 1$) is considerably smaller than the possible range of ρ

Figure 1.2: Median euclidean distance for different sub-sample sizes



Note. Different sub-sample sizes on the horizontal axis (50 corresponds to the AM/FKM method). Euclidean distance to the *original* estimation on the vertical. Each observation captures 76 subjects, each estimated with 100 simulations.

[7]The original data can be downloaded on the website of the American Economic Association under this link: https://www.aeaweb.org/articles?id=10.1257/aer.97.5.1858. Note that this definition of accuracy implies that we treat the parameters estimated in Fisman et al. (2007) as our *benchmark* for simulations. Should we not observe a convergence in accuracy, this would point towards weaknesses of the original methodology. However, since we do find a convergence of estimations in our simulations, it does not seem that more dictator games are likely to yield different estimation results, however this might be worth exploring in a subsequent study.

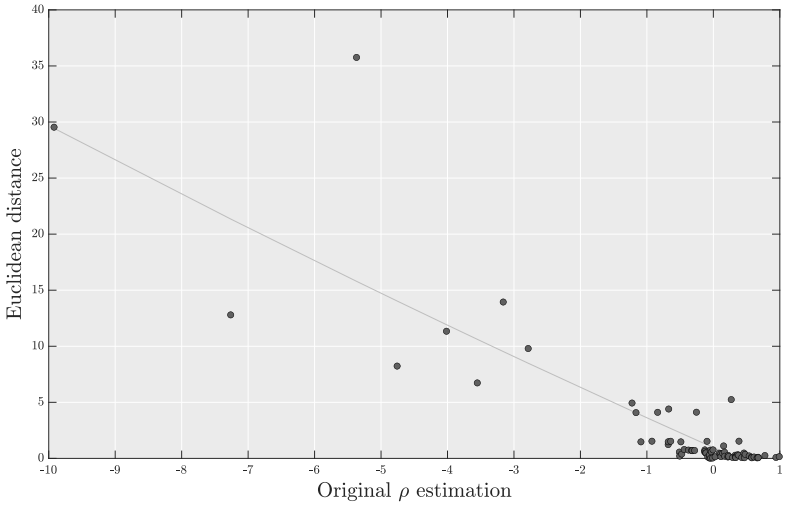
$(-\infty \leq \rho \leq 1)$. All in all, from Figure 1.2 we conclude that accuracy for both parameters is still relatively high, also when reducing the number of decisions. For instance, at a sub-sample size of 20, *median* euclidean distance for ρ is 0.12, and for α even as low as 0.022.

That being said, focusing on the mean instead of median accuracy (Appendix A, Figure A.1.1), we find that estimates for ρ appear to be subject to substantial outliers. To gain more insights into these outliers, we explore whether there are specific values of ρ for which the estimations are particularly inaccurate. Figure 1.3 plots evidence on this, as it shows mean euclidean distances on the vertical axis for all original AM/FKM estimates of ρ on the horizontal. While euclidean distance is marginal for for ρ around 0, we find that extreme outliers result from estimating parameters that originally had a large negative value of ρ. Nonetheless, since $\rho \geq -\infty$, there is an infinite space that captures Leontief preferences. In fact, in Fisman et al. (2015b) the authors consider all individuals with $\rho < 0$ as *equality-focused* suggesting that for example a ρ estimation of *-2* or *-10* can be considered equivalent. However, in our simulations we find that estimates of ρ with large negative values are not only sometimes estimated with other large negative values of ρ but jump between $\rho < -1$ and $\rho = 1$. For the CES utility function, this implies that the same individual is sometimes estimated as strongly equity concerned ($\rho < -1$) and at other times as entirely efficiency concerned ($\rho = 1$). Moreover, these switches do not seem to be a result of the reduction in the number of decisions but already occur at sub-samples sizes close to the original methodology, leading to our first main result.

Result 1. *Using simulations we show that estimations of individual preference parameters with* $\rho \leq -1$ *or* $\rho = 1$ *are not robust irrespective of the sub-sample size.*

This is also confirmed by the fact that once we exclude estimations of $\rho < -1$ and $\rho = -1$ we now find high *mean* accuracies across sub-sample sizes (Appendix A, Figure A.1.2). Thus,

Figure 1.3: Mean accuracy for original estimations of ρ across sub-samples sizes



Note. Estimates with the AM/FKM method on the horizontal axis. Mean euclidean distance to the *original* estimation on the vertical. Each observation captures one subject, each estimated with 100 simulations for 24 sub-sample sizes.

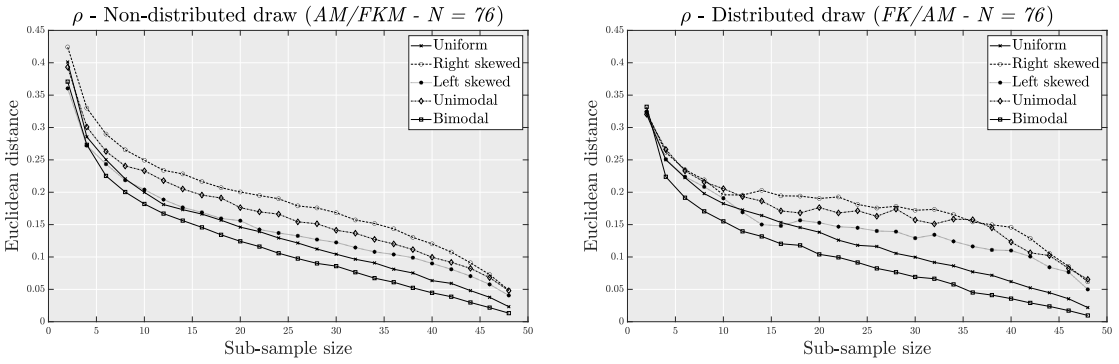
across all subsequent analyses we only include estimations of ρ within the robust range $[-1, 1)$. However, despite these exclusions the estimates of ρ are still less accurate than the ones of α.

To investigate whether there are other possibilities to improve our estimation accuracy, we now turn to the budget lines. As ρ is capturing the trade-off between equity and efficiency, it is plausible that certain slopes of budget sets provide more information for the econometric calibration than others. Consider a scenario where in 50 rounds each randomly generated price ratio is exactly identical and a subject consistently makes the same choice. This would fail to provide any variation and thus make it impossible to accurately calibrate preference parameters. We therefore consider a broader representation of budget sets by ordering all 50 allocation decisions by their price ratio for each subject and then draw sub-samples of allocation decisions according to distinct underlying distributions.

We again follow an exploratory approach here and investigate a variety of underlying distributions (i.e., *left skewed*, *right skewed*, *unimodal*, and *bimodal*).[8] These distributions ensure an oversampling of steep, flat, intermediate or steep and flat budget lines respectively. We again draw 100 sub-samples per subject and size. The left graph of Figure 1.4 plots *mean* euclidean distances for ρ across sub-sample sizes according to distinct underlying distributions. We can see that in fact different distributions clearly affect accuracy. Some distributions perform worse (*right -, left skewed, unimodal*) than the *random* (uniform) draw, whereas the distribution yielding the most accurate estimations is *bimodal*. This implies that when oversampling low as well as high price ratios (flat and steep budget lines) we can significantly improve our accuracy of ρ estimations. This seems intuitive as a bimodal distribution oversamples budget sets with a high tension between equity and efficiency and thus likely provides the most information for the estimation. In fact, for a sub-sample size of 20, a bimodal draw improves accuracy by almost 15%.

Nonetheless, the bimodal draw as outlined here might still oversample *only* low or *only*

Figure 1.4: Mean euclidean distances of ρ for varying underlying distributions



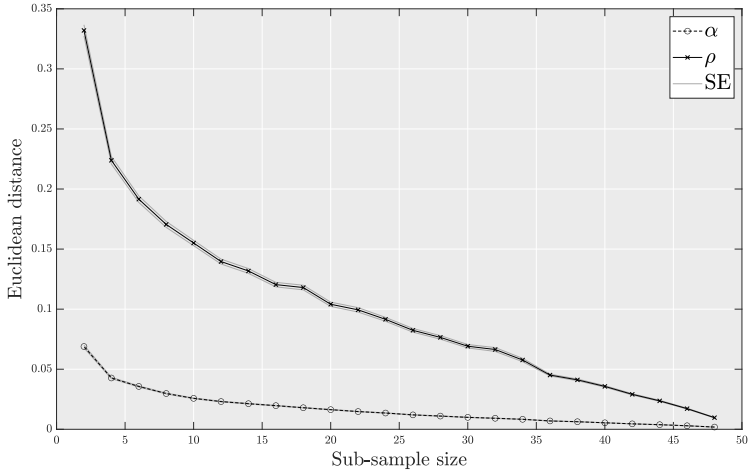
Note. Left: Mean euclidean distance for *non-distributed* draw of ρ. Right: Mean euclidean distance for *distributed draw*. Each observation captures 76 subjects, each estimated with 100 simulations for a specific underlying distribution. Fewer observations according to exclusion criteria for ρ ($-1 \leq \rho < 1$).

[8]Note that so far we have drawn budget sets at random, which is equivalent to drawing from a uniform distribution.

high price ratios without ensuring sufficient numbers of both. Consider the case for a sub-sample of 2 allocation decisions. There is a considerable likelihood that both budget sets are drawn either from low or from high price ratios, but not one from each. To address this issue in an additional step, we draw price ratio such that they are distributed between steep and flat slopes, whilst accounting for underlying distributions. For example, for a sub-sample size of 2, we ensure to select one budget set with a flat and one with a steep budget line. The result of these simulations is shown in the right graph of Figure 1.4.[9] We again observe that a bimodal draw provides the most accurate estimations of ρ. Moreover, when comparing the left and right graphs in Figure 1.4, we find additional improvements in the estimation accuracy of ρ on the right. Taking again a sub-sample of 20 decisions as an example, the accuracy is increased by another 16% in the distributed (right) relative to the non-distributed (left) draw. Thus, relative to the benchmark of a non-distributed random draw we find considerable accuracy improvements of 29% when estimating ρ.

All in all, using a distributed bimodal draw we only observe minor inaccuracies (0.104 points) when estimating ρ for 20 allocation decisions. However, despite these improvements in the estimate of ρ we still need to ensure accurate estimates of α as well. In fact, we find that its estimate also improves when moving to our optimal way of drawing budget sets. While accuracy was 0.0204 points for a sub-sample of 20 decisions under the random draw, this number improves to 0.0175 under a bimodal and 0.0163 with a distributed bimodal draw. Figure 1.5 plots the mean accuracy of both, α and ρ, using the preferred distributed bimodal draw. The graph depicts remarkably accurate estimations for both parameters. When examining the accuracy at 20 decision problems once more for $\alpha \in [0, 1]$ and $\rho \in [-1, 1)$, we find mean

Figure 1.5: Mean euclidean distances for bimodal, distributed draw



Note. Mean euclidean distance for *best* methodology. Each observation captures 76 subjects, each estimated with 100 distributed, bimodal draws. Fewer observations according to exclusion criteria for ρ ($-1 \leq \rho < 1$).

[9] For this we again order all price ratios and then split the data into n differently sized bins. For the bimodal distributions the bins for low and high price ratios are narrow, whereas the bins for intermediate price ratios are wider. Following this, we randomly pick one budget set per bin.

euclidean distances relative to the original method of 0.0163 (1.6%) and 0.104 (5.2%) estimation points respectively.[10]

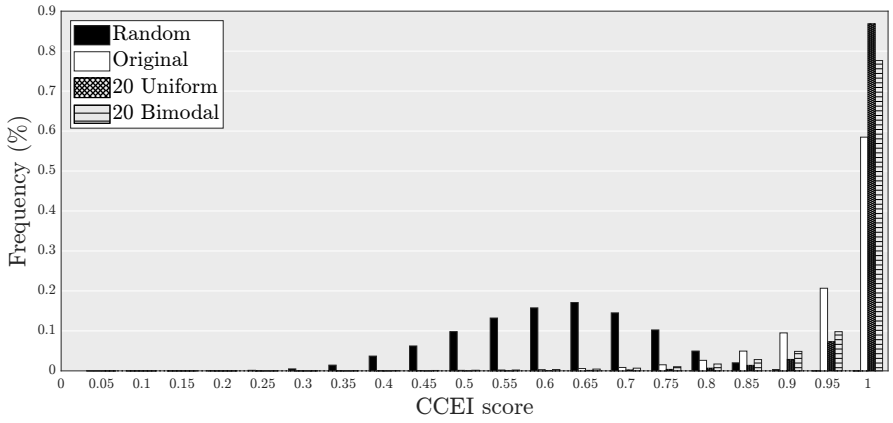
After having established a strategy to draw budget sets in a way that ensures high levels of accuracy, we have to choose, which number of decisions provides a good trade off between the practical cost-effectiveness of a reduction in decisions and loss of accuracy. This decision is ultimately somewhat arbitrary, however, when examining Figure 1.5, we can see that most substantial improvements in accuracy from marginal changes in the number of decisions occur at sub-sample sizes below 15. After which the improvements are almost linear. However, as we still find higher levels of inaccuracies with a sub-sample of 15 and aim to provide a robust tool to estimate distributional preferences, we ultimately consider 20 decision allocations (a reduction of 60%) a good trade-off between making the methodology more efficient, whilst ensuring high levels of accuracy.

Result 2. *Based on our simulations using the AM/FKM data, we find that a sub-sample of 20 budget sets, where price ratios are drawn according to a distributed bimodal distribution, allow an estimation of distributional preference parameters with high accuracy.*

3.2 CCEI score

While we just showed that both α and ρ can be estimated with high accuracy for a sub-sample size of 20, we need to examine additional dimensions to ensure the robustness of our proposed preference elicitation, namely rationalisability of choices. In order to do so, we follow the literature and calculate the Critical Cost Efficiency Index (CCEI) developed by Afriat (1972) to examine GARP violations in our data. Figure 1.6 shows distributions of CCEI scores from the original methodology (white) as well as the condensed simulations with a random (*uniform*,

Figure 1.6: CCEI scores for original, simulated and random data



Note. Distributions of CCEI scores. Compare original AM/FKM (white), 20 budget sets uniform (cross), 20 budget sets bimodal (horizontal lines) relative to a random choices (black). All observed behaviour greatly exceeds random decisions

[10]The percentage accuracies relate to the range of $\alpha \in [0, 1]$ and $\rho \in [-1, 1)$.

crossed lines) and the distributed bimodal draw (horizontal lines). In line with the literature, the figure also depicts a hypothetical benchmark of CCEI scores from 10,000 simulated individuals making random choices in 50 allocation decisions (black). A CCEI score always lies between 0 and 1, where higher values refer to higher rationalisability, thus fewer GARP violations. Note that by construction, simulating a reduction in the number of allocation decisions can never decrease the CCEI score for a subject, as there are fewer opportunities for violations. Therefore, like AM/FKM we find CCEI scores that are significantly higher than random choices and demonstrate a high degree of rationalisability. Moreover, we find that the CCEI score of the distributed bimodal draw is marginally closer to the original method than the uniformly drawn decisions. This, however, is not surprising as, intuitively, oversampling *steep* and *flat* budget lines makes violations of GARP more likely since subjects face more choices that can be inconsistent with other elicited behaviour. Thus, we conclude that also when examining the CCEI score, reducing the number of decisions to 20 provides robust evidence for rationalisability.

3.3 Our proposed implementation

From the simulation results above we arrive at the following proposal to improve the methodology put forward by AM/FKM.

1. Subjects face 20 modified dictator games with price ratios drawn bimodally distributed across the range of all budget sets. For implementation, split all possible price ratios ($0.1 \leq PR \leq 10.0$) into 20 distinct bins, where each bin size is defined as 5% of the area of a beta distribution with $\alpha = \beta = 0.5$.[11]

2. For each of the 20 allocation decisions a price ratio is selected randomly from one bin as defined in (1). Bins are selected in a random order.[12]

3. Instructions, incentives and econometric calibration are identical to AM/FKM. To ensure robustness only include values of $-1 \leq \rho < 1$ for all analyses.

After having identified a methodology that condenses the estimation of altruistic preference parameters, we now turn to exploring its validity in two ways. First, we investigate whether we can replicate our simulation findings with substantially more observations, using data from the *American Life Panel* (ALP). Following the replication, we then move on to collect our own data and test the proposed methodology with a new subject pool.

4 Validating with the *American Life Panel* data

The *American Life Panel* is a survey conducted online with more than 5000 adult Americans. The data we use was collected in 2013 by Fisman, Jakiela & Kariv and later utilised in a number of subsequent publications (Fisman et al., 2015b, 2017). The methodology employed is almost

[11] A figure of the beta distribution and the corresponding bins can be found in Appendix A, Figure A.1.3.
[12] Please contact the authors for code to run a web implementation of the task.

identical to the original task in Fisman et al. (2007), with the only difference that respondents answered the survey online instead of in a laboratory environment. Following invitations via e-mail, 1257 ALP subjects completed all of the 50 modified dictator games. Fisman et al. (2014) show that the composition of the ALP subsample that completed the survey, is consistent with a representative sample of the US population and the overall ALP sample. For the simulations below we utilise all 1257 subjects.

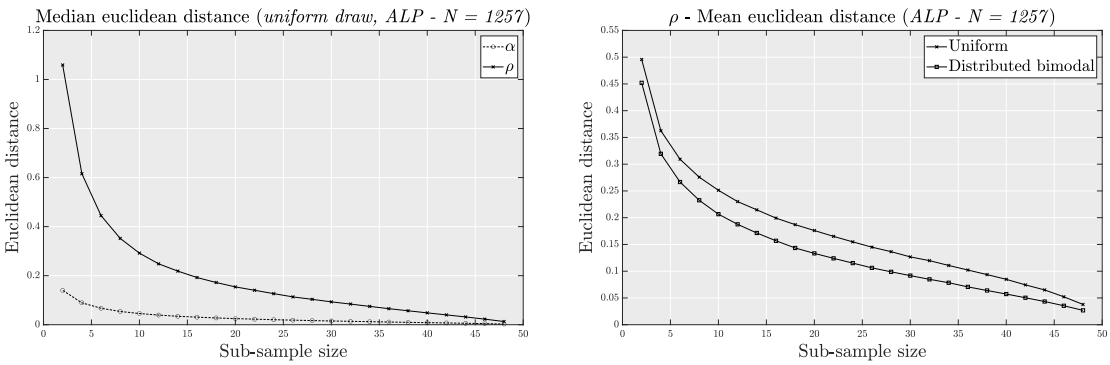
4.1 Replicate simulations

In a first step to test our previous findings, we run the same simulation exercises as above, focusing explicitly on our proposed distributed bimodal draw as well as the uniform draw as a benchmark. Overall, we are again exploring 24 different sub-sample sizes with 100 draws per individual, but this time for 1257 subjects. This leaves us with just over 6 million α and ρ parameter estimations in total. Moreover, we once more employ euclidean distance to the original estimates as our measure of accuracy.

The main simulation results can be found in both graphs presented in Figure 1.7. The left graph replicates Figure 1.2, by plotting median accuracy levels for both parameters across sub-sample sizes using the uniform draw. The graph confirms our findings from above by showing that also for the ALP, median accuracy for both parameters is relatively high for sub-sample sizes of 15 and higher. Upon closer inspection, however, accuracy is slightly lower using the ALP data compared to the original AM/FKM. Previously, for a sub-sample size of 20, we found euclidean distances of 0.022 (0.12) for α (ρ), now we find 0.0249 (0.15) for the two parameters. Thus, the differences appear to be marginal.

Moreover, in the left graph, when examining both parameters, we again observe lower levels of accuracy for ρ compared to α. This is once more especially apparent when focusing on mean instead of median euclidean distance due to extreme outliers for $\rho \leq -1$ (Appendix A, Figure A.1.4). Following the same procedure as before, after excluding all observations that are estimated with $\rho < -1$ or $\rho = 1$ we find high *mean* accuracies for estimations of ρ. In fact, these are presented in the right graph of Figure 1.7 as the *mean* euclidean distance of the uniform

Figure 1.7: Simulation results for different sub-sample sizes (*ALP*)



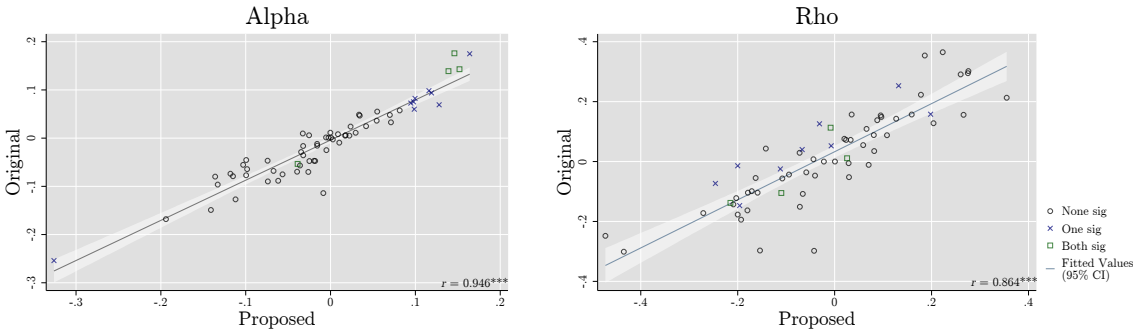
Note. Left: median euclidean distance for uniform draw. Right: distributed bimodal, and non-distributed uniform draw for $-1 \leq \rho < 1$. Each observation corresponds to 1257 subjects, each estimated with 100 simulations.

draw. The same graph serves an additional purpose to compare the uniform draw with our recommended *distributed bimodal* draw. Euclidean distance for the *distributed bimodal* draw is again lower for all sub-sample sizes, therefore we replicate our previous simulations that a *distributed bimodal* draw improves accuracy significantly. Once again, we find slightly lower levels of accuracy compared to the AM/FKM data, but still observe a low mean euclidean distance for ρ at a sub-sample of 20 of 0.13 (6.5%) estimation points, compared to 0.104 or 5.2% previously.

4.2 Demographic predictors of preference parameters

Using the ALP sample also provides us with the additional opportunity to investigate the relationship between demographics and estimated distributional preferences using our proposed as well as the original methodology. This analysis is in part motivated by work by Fisman et al. (2015b, 2017) and Kerschbamer & Müller (2020) who find that social preferences are significantly related to voting behaviour, being an elite student and other demographics such as age and gender. The information available in the ALP data is *age, gender, ethnicity, education, born in the US, living situation, family income, work status and category of occupation.* For the purpose of validating our proposed methodology, we can examine four regression models. Two models include the original and proposed estimates of α respectively, whereas the other two compare the methodologies for ρ as dependent variables and employing all demographic variables as regressors (Appendix A, Table A.1.1). We can then compare the significance and size of each coefficient between the original AM/FKM and our proposed methodology. If our proposed methodology provides robust estimates of preference parameters, we should find similar regression coefficients between the original and the propose instrument. Irrespective of statistical significance, the sign and magnitude of the estimates should closely correlate to support our proposed methodology. To provide this comparison, we plot each coefficient estimated in the regression model in Figure 1.8. Thus each of the 63 observation in each graph of Figure 1.8 corresponds to a pair of coefficients estimated once with the original (vertical) and

Figure 1.8: Correlation of regression coefficients across methodologies



Note. Parameters estimated with AM/FKM on the vertical axis and our proposed methodology on the horizontal. Each observation in the Figure corresponds to one coefficient from regressing a number of demographics on the respective parameters (α left, ρ right). Demographics are *age, gender, ethnicity, education, born in the US, living situation, family income, work status* and *category of occupation*

once with our proposed methodology (horizontal).[13] The left graph shows the relationship between coefficients of the two methodologies for α, whereas the right graph plots the same for ρ. The marker shape and colour indicate whether a pair of coefficients was significant at the 10%-level in both, in one, or in neither model. Both graphs show a high and significant correlation between coefficients, providing strong support that the relationship between estimated preference parameters and demographics is consistent between both methodologies ($r_\alpha = 0.946$; p-value < 0.01 and $r_\rho = 0.864$; p-value < 0.01).

Moreover, with respect to demographic information across the regression models, we find evidence that women as well as lower levels of family income have lower levels of α. The former finding is in line with results by Kerschbamer & Müller (2020), who also find higher preferences for equal payoffs for women. Moreover, the latter result supports Fisman et al. (2015b), finding more selfish preferences for elite (presumably higher income) individuals. Contrasting this, we cannot replicate the literature's findings with respect to ρ. We neither find higher preferences for efficiency for high-income individuals (Fisman et al., 2015b), nor that older subjects are more equality concerned (Kerschbamer & Müller, 2020). All other demographic information available does not show any significant relationship to estimated preference parameters.

After having successfully replicated our initial findings with a more diverse and larger sample as well as provided additional evidence regarding the consistency of coefficient estimates, we now move on to our last validation step and test the proposed methodology by means of a new experiment.

5 Own data collection

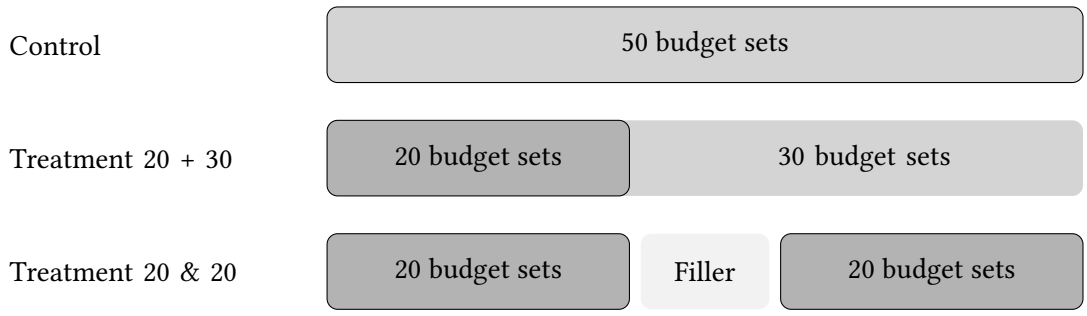
5.1 Design and procedure

For the final validity test, we design a study spanning three distinct treatments. These allow us to address additional potential differences between the original and our proposed methodology, namely *fatigue*, *the difference in probability of payoff relevance for each decision*, and *being instructed in the beginning of the experiment that it consists of 20 decisions*. Our proposed methodology should on improve aspects regarding *fatigue*, as it is a concern that due to the cognitive load of 50 allocation decisions, subjects get tired and change their choices in some form towards later rounds. With respect to the second dimension, in AM/FKM subjects have a $1/50$ probability of each decision to be payoff relevant. By reducing the number of dictator games to 20, this probability is now increased, which could lead to a change in behaviour. Lastly, once being instructed that the one will face 50 (20) decisions, it is possible that subjects adjust their behaviour according to the expectation of the number of decisions they will have to make. All of these three concerns cannot be accounted for when using simulations, which motivated us to control for them in our experimental design (see Figure 1.9) for an overview).

As a first treatment, we replicate the original method exactly as in Fisman et al. (2007)

[13] We estimate 63 coefficients for 9 demographic variables as there are various categorical variables, which levels are included as dummies.

Figure 1.9: Overview of experimental design



Note. Each of the treatments is conducted between-subject. The black outline in the figure indicates budget sets that we utilise for estimations in the analysis. 20 budget sets follow the bimodal, distributed draw of price ratios.

(*Control*). Our second treatment (*Treatment 20 + 30*), holds the total number of decisions, the time taken, and the expected value of each choice constant, while using our proposed way of drawing budget sets. We ask subjects to make decisions in 50 modified dictator games, where we draw budget sets according to our proposed methodology in the first 20 rounds, followed by 30 rounds as in AM/FKM. Thus, from a subjects perspective the first and second treatment are exactly identical, however the latter allows us to estimate preference parameters based on our proposed methodology. Lastly, to investigate a *pure* version of our new method in a third treatment, we ask subjects to complete 20 allocation decisions with our proposed method twice (*Treatment 20 & 20*) with a brief filler task between the two sets of tasks. Treatment 3 allows us to hold the overall time of the experiment constant in order to avoid selection effects into a specific treatment. Moreover, the treatment also enables us to explore within-subject stability of the elicited preference parameters providing us with even more insights into the reliability of our modified task as well as the general stability of estimated preferences.[14]

Referring to the three potential concerns raised above, *Treatment 20 & 20* addresses the element of fatigue, which when at play would imply different choice behaviour during the second of the two elicitations compared to the first. Moreover, the probability of payoff relevance is controlled for comparing the *Control* and *Treatment 20 + 30*, as in both cases subjects complete 50 allocation decisions of which one is selected at random. Lastly, in *Treatment 20 & 20*, subjects are initially only informed about the first set of 20 decision allocations, thus we can compare these estimated preference parameters with *Treatment 20 + 30*, to explore whether the expectation of a number of allocation decisions has an impact on estimated preferences.

Incentives in all treatments are identical and follow the original study. First, each subject is matched with one other participant. Then, one allocation decision is randomly chosen for payment with subjects receiving the tokens they allocated to *self* and the money the matched

[14] As both sets of tasks were completed only separated by a filler task, the test-retest reliability is not comparable to studies with a larger time span between the tasks. Moreover, the tasks entail randomly drawn budget sets, which means the 2 sets of 20 budget sets were not *exactly* identical. Thus the second set of 20 budget sets can alternatively also be used for out-of-sample predictions.

participant allocated to *other* in their chosen round. In *Treatment 20 & 20* we paid out one round of each of the 20 sets of decisions.

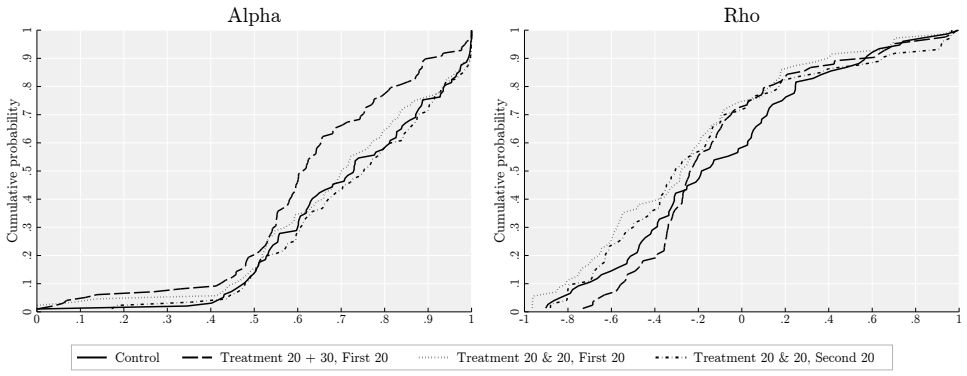
The experiment was approved by the Nottingham School of Economics' Research Ethics Committee. We conducted the study in January 2021, programmed the experimental interface using LIONESS Lab (Giamattei et al., 2020) and recruited subjects using Amazon MTurk.[15] We sampled a total of 282 subjects, distributed across treatments, the average duration was around 25 minutes and subjects received a $1 participation fee plus a $1 bonus for 25 tokens earned. This payoff structure maintained the proportionate payment between participation and bonus identical to AM/FKM, resulting in an average payment of about $10/hr.[16]

With respect to our hypotheses, by randomisation, we expect identical distributions of preference parameters across treatments, if our proposed methodology is identical to the original AM/FKM.

5.2 Results

Overall, our data supports the conclusion that our proposed methodology allows for an accurate estimation of parameters. Figure 1.10 shows the different cumulative parameter distributions obtained from our three treatments. It also plots separate preference parameter distributions for the first and second elicitation in *Treatment 20 & 20*. Visually, across both graphs we can observe no differences in distributions of both parameters between the *Control* and the two elicitations in *Treatment 20 & 20*. However, we do find a significant difference in the distributions of α (and to a lesser extent ρ) between *Treatment 20 + 30* and *Control*. This visual evidence is also statistically confirmed by the results of Kolmogorov-Smirnov tests presented in Table 1.1. The Table reports p-values of comparisons between each treatment and *Control* for both estimated parameters. We find that once we restrict ρ to the robust range

Figure 1.10: Cumulative distributions of α and ρ parameters across treatments



Note. Cumulative distribution of α (left) and ρ (right). For ρ we exclude subjects for whom $-1 \leq \rho < 1$.

[15] Instructions can be found in Appendix A.2.

[16] Due to the double bonus paid in *Treatment 20 + 20*, hourly payment was higher for this treatment ($13.62/hr) and lower for the other treatments ($8.09/hr), which however was not announced beforehand to the subjects.

Table 1.1: Overview of treatment comparisons

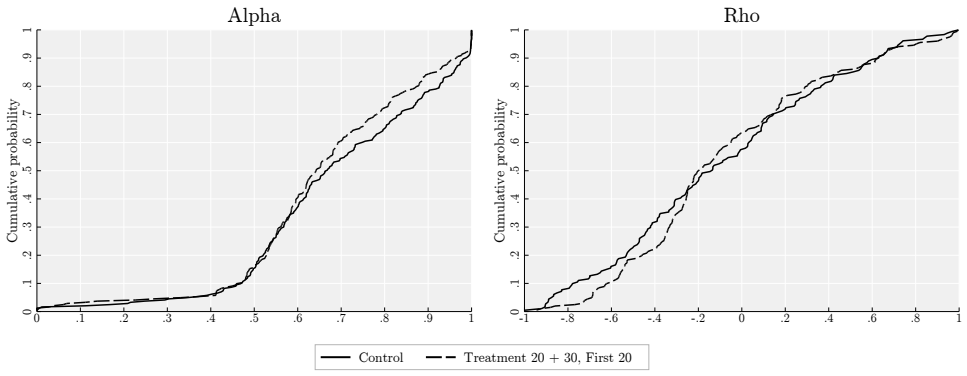
	α	ρ	$-1 \leq \rho < 1$
Control vs. Treatment 20 + 30	0.028**	0.090*	0.200
Control vs. Treatment 20 & 20 - First	0.965	0.787	0.125
Control vs. Treatment 20 & 20 - Second	0.958	0.659	0.298

Note. All values reported are p-values of Kolmogorov-Smirnov tests. The results are also robust when employing Wilcoxon rank-sum tests.

$-1 \leq \rho < 1$ (column 4), we can no longer find significant differences in distributions between the *Control* and *Treatment 20 + 30*.

The observed differences in the distributions of α is somewhat surprising as, especially for α, our simulations throughout show a remarkably high accuracy when reducing the number of decisions. Thus, we contemplate whether this specific result might be due to sampling error as we conducted treatments on subsequent days instead of implementing a clean randomisation within the same session. We examine comparisons of demographics and other observables to investigate potential randomisation failure, but cannot find any differences. Nonetheless, to explore whether this finding is in fact due to sampling error or a robust result, we pre-registered a replication of our *Control* as well as *Treatment 20 + 30* now employing randomisation into treatment within the same session.[17] Moreover, using our obtained results from above we ensured 90% power for statistical inferences, recruiting an additional 291 subjects across both treatments. The replication results do point towards sampling error in our first data collection as they are now in line with our initial hypothesis. We find no significant difference between the *Control* and our proposed methodology. To confirm this even further, Figure 1.11 shows pooled cumulative distributions of both parameters for the two treatments of interest, thereby substantially increasing our statistical power ($N = 486$). Visually, as well as

Figure 1.11: *Control* and *Treatment 20 + 30* - pooled data



Note. Cumulative distribution of α (left) and ρ (right) for *pooled* data (N = 486). For ρ we exclude subjects for whom $-1 \leq \rho < 1$.

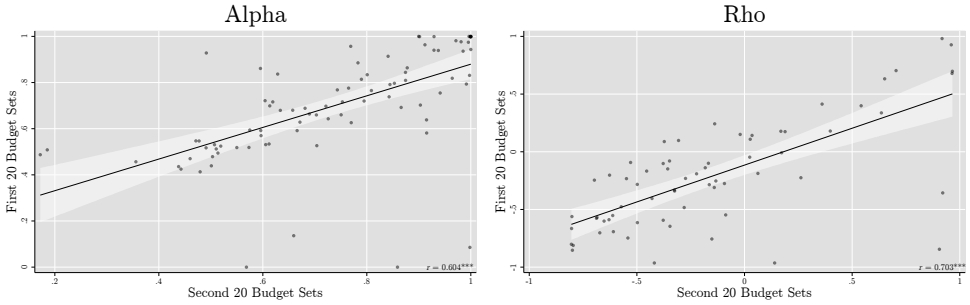
[17]Registration number AEARCTR-0007217. See https://www.socialscienceregistry.org/trials/7217.

using Kolmogorov-Smirnov tests, we fail to detect a difference in distributions of parameters between the treatments.[18]

Finally, to further support the robustness of our estimated parameters, we also find a high within-subject consistency in *Treatment 20 & 20* (see Figure 1.12). Each graph plots parameters based on the first elicitation on the vertical and based on the second elicitation on the horizontal axis. The fitted line is the result of a univariate regression and correlation coefficients are provided in the bottom right corner, showing substantial correlations for both parameters ($r > 0.60$; p-value < 0.01). This suggests that our proposed methodology is not only robust in comparison to the original tool but it also consistently estimates parameters for the same individual. Moreover, despite previously finding less robust results when estimating ρ, for the within-subject stability, we now find a somewhat stronger correlation for ρ than for α, supporting that the proposed methodology does in fact also robustly estimate ρ. With respect to our three possible dimensions of concern, fatigue, probability of payoff relevance and instructions, we conclude that these do not substantially affect choice behaviour as we find no significant difference between any of the pooled parameter distributions.

A main motivation for exploring a condensed version of AM/FKM was to significantly reduce the time it takes subjects to complete the experiment. In our online setting, without the time taken for instructions and comprehension questions, subjects took ~ 17 min to complete the *Control*. Our condensed task on the other hand took subjects ~ 8 min to complete.[19] While still yielding accurate estimations, our proposed methodology therefore reduces overall completion times by around 50% relative to the original procedure. This makes it possible to conduct an entire preference elicitation in around 13 minutes including instructions, comprehension questions and practice, making it now viable as an auxiliary measure to be used in combination with other experimental tasks.

Figure 1.12: *Treatment 20 & 20*: Within-subject comparison



Note. Within-subject estimations of α (left) and ρ (right). For ρ we exclude subjects for whom $-1 \leq \rho < 1$. Tasks were separated by a filler task. First elicitation on vertical and second elicitation on horizontal axis.

[18] Please see Appendix A, Table A.1.2 for Kolmogorov-Smirnov tests.

[19] These numbers are rather an upper bound as we included a delay of 2 seconds before subjects could make their decision in order to give them time in order to engage with the graphical interface. Moreover, for both treatments the instructions were identical and took ~ 6 min to complete.

6 Conclusion

In this chapter we show that distributional preference parameters can be accurately estimated using a condensed version of the AM/FKM methodology. Results from simulations indicate that α can always be estimated with high precision, whereas ρ is only accurate for $-1 \leq \rho < 1$. Moreover, we can substantially improve estimation accuracy when oversampling flat and steep budget lines. Using this approach we find strong evidence that 20 allocation decisions are sufficient for accurate estimations. These findings are further supported by a replication exercise using the larger, more diverse ALP sample as well as an experimental study, where we could not identify any statistically significant differences between the original and the condensed methodology. Altogether, the results convincingly show that a condensed version of the AM/FKM method still allows an accurate estimation of distributional preferences. We therefore provide researchers with a more time- and cost-effective tool that can be easily integrated in experimental designs to explore other-regarding behaviour.

Bibliography

Ahn, D., Choi, S., Gale, D., & Kariv, S. (2014). Estimating ambiguity aversion in a portfolio choice experiment. *Quantitative Economics*, 5(2), 195–223.

Andreoni, J. & Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2), 737–753.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), 596.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2), 321–338.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Chan, J., To, H.-P., & Chan, E. (2006). Reconsidering social cohesion: Developing a definition and analytical framework for empirical research. *Social Indicators Research*, 75(2), 273–302.

Charness, G. & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3), 817–869.

Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, 97(5), 1921–1938.

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who is (more) rational? *American Economic Review*, 104(6), 1518–50.

Fisman, R., Jakiela, P., & Kariv, S. (2014). *The distributional preferences of Americans*. Technical report, National Bureau of Economic Research.

Fisman, R., Jakiela, P., & Kariv, S. (2015a). How did distributional preferences change during the great recession? *Journal of Public Economics*, 128, 84–95.

Fisman, R., Jakiela, P., & Kariv, S. (2017). Distributional preferences and political behavior. *Journal of Public Economics*, 155, 1–10.

Fisman, R., Jakiela, P., Kariv, S., & Markovits, D. (2015b). The distributional preferences of an elite. *Science*, 349(6254).

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5), 1858–1876.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), 347–369.

Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the 'inclusion of the other in the self' scale. *PloS one*, 10(6), e0129478.

Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). Lioness lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1), 95–111.

Henrich, J. & Henrich, N. (2006). Culture, evolution and the puzzle of human cooperation. *Cognitive Systems Research*, 7(2-3), 220–245.

Jakiela, P. (2013). Equity vs. efficiency vs. self-interest: on the use of dictator games to measure distributional preferences. *Experimental Economics*, 16(2), 208–221.

Kelley, H. H., Berscheid, E., Christensen, A., Harvey, J. H., Huston, T. L., Levinger, G., McClintock, E., Peplau, L. A., & Peterson, D. R. (1983). Analyzing close relationships. *Close Relationships*, 20, 67.

Kerschbamer, R. & Müller, D. (2020). Social preferences and political attitudes: An online experiment on a large heterogeneous sample. *Journal of Public Economics*, 182, 104076.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3), 482–493.

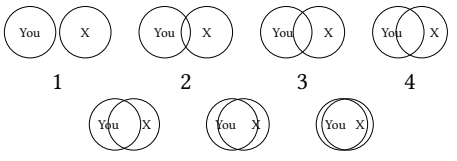
Chapter 2

Developing an extended version of the 'Inclusion of Others in Self' scale

1 Introduction

After having proposed a condensed tool to elicit distributional preferences, we now turn towards our second methodology of concern, the elicitation of social cohesion. As outlined in the previous chapter, experimental tools can be evaluated and compared along a number of dimensions. If an experimental tool is more time- and cost-effective, easy to explain and understand for subjects whilst maintaining its accuracy of measurement we should consider it to be superior relative to a compared tool. Above we showed that we can in fact develop a tool to elicit other-regarding preferences that is more time- and cost-effective whilst still being accurate. In this chapter, we outline a similar approach for a methodology measuring levels of relationship closeness, proposing an improved version of the 'Inclusion of the Other in the Self' scale (IOS) (developed by Aron et al., 1992).[1] In a previous study Gächter, Starmer and Tufano (2015, henceforth GST) find compelling evidence that the IOS scale is a cost efficient tool to measure social cohesion by extracting precise information about relationship closeness between individuals (see Table 2.1, LHS). The tool can be administered in under a minute and

Table 2.1: *Oneness*: 'Inclusion of the Other in the Self' (IOS) scale & We scale

IOS scale	**'We' scale**
"*In the following figure we ask you to consider which of these pairs of circles best represents your relationship with X. By selecting the appropriate number please indicate to what extent you and X are connected.*"	"*Please, select the appropriate number below to indicate to what extent you would use the term "WE" to characterize you and X.*"
	Not at all Very much so 1 2 3 4 5 6 7

Note. The arithmetic mean of both scores constitutes the level of *oneness*. The origin of the 'We' scale is not provided in Cialdini et al. (1997).

This chapter is based on joint work with Simon Gächter, Chris Starmer & Fabio Tufano. I want to thank Ben Beranek and Geoffrey Castillo whose modification of the IOS scale has inspired this research. I am grateful for funding by the British Academy.

[1] Note that as defined in the introduction, for the purpose of this thesis we consider social cohesion and relationship closeness as synonyms.

due to its graphical interface is highly intuitive.

That being said, when evaluating the scale, across a series of studies GST find that a superior method to capture relationship closeness is to combine the IOS with the so called 'We' scale (see Table 2.1, RHS). Taken together those two scales form a measure called *oneness* (Cialdini et al., 1997). In order to evaluate respective performances, GST correlate a variety of established tools designed to capture relationship closeness with the IOS and *oneness* measure. They find robust evidence that *oneness* is in fact outperforming the IOS scale as it shows higher cross-correlations with other elicited benchmarking tools. Since the publication, many studies have referred to their evidence to motivate the inclusion of the IOS as an experimental measure of social cohesion. However most of these studies do not employ the combined *oneness* scale (see Tarr et al., 2016; Westlund et al., 2018; Pellencin et al., 2018; Molleman & Gächter, 2018; Dimant, 2021; Robson, 2021, among others). Thus, despite the knowledge of the advantages, authors seemingly prefer implementing the IOS scale by itself rather than *oneness*.[2] We think that even though both elements of the measure are short, having to administer two instead of a single scale to elicit *oneness* might not be perceived as desirable, even if it provides a better measure of social cohesion. In addition, when implementing repeated measures of *oneness*, employing two scales for each repetition exacerbates possible concerns. As for the AM/FKM task the lack of use of the *oneness* measure seems thus to be driven by pragmatic design concerns and not by a doubt towards the quality of the tool.

Before, we outline how we propose to improve the tool, let us first hypothesise why *oneness* might be a more accurate measure than IOS. Having one additional scale likely provides more information about relationship closeness. It is plausible to assume that the addition of the 'We' scale allows subjects to provide more nuance to their IOS score, as the original IOS scale is restricted to exactly 7 levels of relationship closeness. Thus, in case subjects feel only slightly closer to one than another individual, they can provide an identical IOS score, but increase the response in the 'We' scale. If this holds true, then adding this nuance might also be possible without a second independent scale but by simply extending the number of pairs of circles in the original IOS scale.

In line with this thought we propose an extended IOS scale, as a computerised version of the original incorporating more levels and thus providing subjects with the opportunity to indicate more nuance in their reported social cohesion. We then validate our proposed scale by precisely replicating GST using our novel methodology.[3] The following chapter therefore combines a validation of our extended IOS scale with a detailed replication of GST.

The chapter is structured as follows. In Section 2.2, we explain our approach of extending the IOS scale and outline how it can be compared to the original IOS measure. In Section 2.3, we show the results of validating the new, and benchmark our findings against the original work by Gächter et al. (2015). In Section 4 we provide concluding remarks.

[2]This is even true for one of the authors involved in GST (see Molleman & Gächter, 2018).
[3]In particular, we replicate their *Study 3* in which they correlate *oneness* and the IOS scale with a series of other measures of relationship closeness.

2 The extended IOS scale

As mentioned above, we decided to extend the original IOS scale in the hope that it will render the inclusion of the 'We' scale redundant. However, when extending the number of relationship levels from which subjects can choose two challenges arise. First, how can we visualise an increased number of overlapping circles as they cannot be easily presented on a single screen as the original IOS scale. Secondly, it is not trivial to decide by how many options the answer range should be extended.

Given that most experiments nowadays are computerised we decided to create an interactive screen that allows subjects to intuitively adjust the degree to which circles overlap. The final layout is displayed in Figure 2.1.[4] Subjects can move the slider below a box in order to adjust the degree to which the circles overlap. As long as experiments are conducted using computers or mobile devices, the changes to the original scale should not affect the portability or ease of explanation and do not increase the time it takes to complete the scale. Using this graphical interface, there is theoretically no restriction on how much we can extend the answer range. This leads us to the second consideration. Moving to the extreme case of a con-

Figure 2.1: Graphical interface of the extended IOS scale

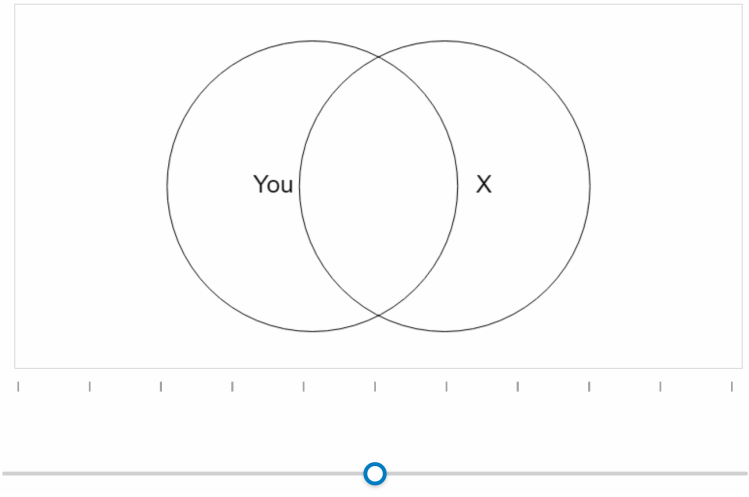
"*Once you move the slider below, a pair of circles will appear in the box. The position of the slider will determine the extent to which the circles overlap. When the slider is all the way to the left, the circles will look like this* 

. When the slider is near the middle, the circles look like this 

. With it all the way to the right the circles look like this 

. You should interpret the degree of overlap as representing the relationship between you and X.

Please position the slider so that the circles indicate to what extent you and X are connected."



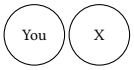
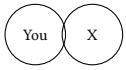
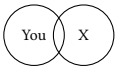
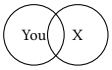
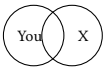
Note. The initial screen the subjects see when entering the elicitation is blank. For illustration purposes we are depicting the slider at a central position in this figure.

[4] You can access the interface using the following link:
https://nottingham.qualtrics.com/jfe/form/SV_6DzwI76k1IDtcb4

tinuous scale might introduce noise as subjects are less likely to explore all possible options. We thus decided to stick with a discrete version of the task. To be comparable to previous studies we moreover decided that the maximum and minimum overlap of circles should match the original methodology by Aron et al. (1992).

We then chose the number of levels such that the change in distance between the centres of the two circles is roughly linear and that the original IOS scale is a subset of the extended version.[5] This leaves us with 11 levels of social cohesion as shown in Table 2.2 that shows how our extended version compares to the original relationship levels. To directly compare scores in both scales we recode the extended IOS, matching the 7-point scale from the original method (see column 4). The recoding ensures that all pairs of circles that are identical between the two scales are assigned the same score and only the four additional items are assigned scores that lie between the existing ones.

Table 2.2: *IOS scale*: original, extended, recoded

	Original	**Extended**	**Extended Recoded**
	1	1	1
		2	1.5
	2	3	2
		4	2.5
	3	5	3
	4	6	4
	5	7	5
		8	5.5
	6	9	6
		10	6.5
	7	11	7

Note. 'X' serves as a placeholder for the initial of the person considered. The original scale does no reduce the distance between circles linearly. Thus, we extend our scale in the ranger [1,3] and [5,6] to yield a linear change in overlap.

[5] We deliberately decided to extend the scale to 11 items instead of 13 to ensure the almost linear change in distance between the two circles across items. There is no explanation in Aron et al. (1992) regarding the chosen degree of overlap between the circles, however intuitively ensuring a somewhat linear change matches the assigned scores.

3 Validating the extended IOS

To test the predictive power of our extended IOS task, we follow the same validation employed by GST and compare our extended IOS tool to the performance of the original IOS scale as well as *oneness*. For a more detailed description of the original validation please inspect GST, we will however summarise the key elements of the design below.

To evaluate the quality in measurement we employ a between-subject treatment, where subjects either perform the original IOS and 'We' scale or complete our extended IOS measure. We then explore the within-subject correlation of each measure of interest to a series of other established surveys designed to capture relationship closeness. The different scales that we use are the *Relationship Closeness Inventory (RCI)*, the *Subjective Closeness Index (SCI)* (both by Berscheid et al., 1989), the *Love and Liking scale* (Rubin, 1970) as well as the *Personal Acquaintance Measure (PAM)* (Starzyk et al., 2006).

Note that some of these measures are specifically constructed to capture social closeness for specific degrees of relationship closeness (e.g. the *RCI* explicitly refers to romantic relationships, whereas the *PAM* is designed for acquaintances). However, from an experimental scientist's perspective, an ideal measurement tool should capture the whole range of relationships. For that reason, GST employ a between-subject variation where participants are asked to either think of a very close person, a friend, or an acquaintance for all questions within the study. For the purpose of validating the extended IOS we replicate this dimension, allowing us to assess the validity of our proposed tool across a range of relationship levels. To ensure salience of the considered person throughout the study, we again follow GST and ask subjects in the beginning of the experiment to provide the initials of the person they are thinking of. These initials are then inserted in all parts where the instructions explicitly refer to another person.[6] The structure of the design allows us to evaluate the measurement quality of the original IOS, *oneness*, and the extended IOS scale independently as well as compare our results to GST. Table 2.3 summarises the design.

We pre-registered our study and collected data online in July 2021 using Qualtrics (Qualtrics,

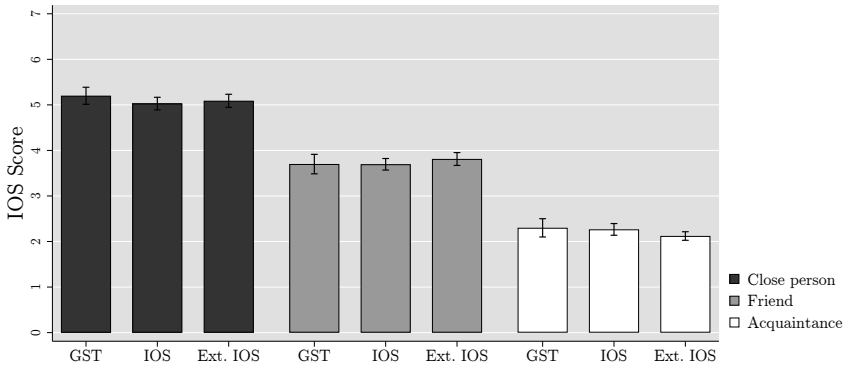
Table 2.3: Experimental overview

	Focus person		
	Close person	Friend	Acquaintance
IOS/Oneness	*RCI, SCI, Love, Like, PAM*	*RCI, SCI, Love, Like, PAM*	*RCI, SCI, Love, Like, PAM*
Extended IOS	*RCI, SCI, Love, Like, PAM*	*RCI, SCI, Love, Like, PAM*	*RCI, SCI, Love, Like, PAM*

Note. The vertical dimension of the matrix compares the original IOS scale and *oneness* (top) to our proposed extended IOS scale (bottom). The horizontal axis indicates the person subjects were asked to consider to vary relationship closeness. Both are manipulated according to a between-subject protocol. The scales that are completed within-subject are included in each cell.

[6] For example, in the IOS and 'We' scale we substitute the 'X' with the provided initial. All instructions and scales utilised can be found in Appendix B.1.

Figure 2.2: IOS Scores for different relationship levels and elicitation methods



Note. Recoded IOS scores for the extended scale. Relationship levels indicated by colour. Close person (black), friend (grey), Acquaintance (white). Each methodology in the set of bars. GST (left), original IOS (center), extended IOS (right).

2021) to program the survey.[7] The study was approved by the Nottingham School of Economics' Research Ethics Committee and, in total, we recruited 753 subjects evenly distributed across all cells ($N \approx 125$ per cell) using Prolific's UK sample.[8] This is about the same number of observations as *Study 3* in GST. We paid a flat fee of £1.20 per subject and the study took about 15 minutes to complete.

Before diving into the comparisons between different scales let us first examine the descriptives of the different treatments and methods. All analyses that follow only utilise the recoded scores as shown in Table 2.2 to allow for direct comparisons between methodologies and the replication. Figure 2.2 plots the elicited IOS scores for different relationship levels and methods used. The colours indicate whether the person thought of was a *close person*, a *friend* or an *acquaintance*. The different elicitations (*GST*, *original IOS*, *extended IOS*) are then presented in each bar from left to right for a respective relationship level. Figure 2.2 shows distinct reported *closeness* for different relationship levels. In line with GST, subjects reported the highest scores for a close person (black), intermediate levels for a friend (grey) and the lowest levels of social cohesion for an acquaintance (white). Moreover, the figure also demonstrates that reported IOS scores between studies and methods are remarkably similar. Not only do we replicate the scores from GST, but also in our extended IOS scale average reported scores are very similar to the original IOS scale.

Result 1. *Our proposed extended IOS scale leads to a remarkably similar assessment of relationship closeness in terms of the overlap of circles as the original IOS scale. This is true for each examined degree of relationship level.*

Following GST, we now examine the correlation of both IOS scales as well as *oneness* with respect to other measures capturing relationship closeness. Each cell in Table 2.4 reports within-subject Spearman's correlations between the different scales. We present the scores

[7]Registration number AEARCTR-0007947. See https://www.socialscienceregistry.org/trials/7947
[8]See www.prolific.co for more information on their services.

Table 2.4: Correlations to other scales: IOS, Oneness, Extended IOS

	GST		Replication		
	IOS	**Oneness**	**IOS**	**Oneness**	**Ext. IOS**
RCI	0.65***	0.68***	0.60***	0.69***	0.68***
RCI Frequency	0.51***	0.53***	0.46***	0.52***	0.52***
RCI Diversity	0.57***	0.60***	0.56***	0.61***	0.64***
RCI Strength	0.62***	0.68***	0.60***	0.70***	0.65***
SCI	0.82***	0.84***	0.79***	0.84***	0.84***
Love scale	0.79***	0.83***	0.73***	0.78***	0.79***
Liking scale	0.56***	0.60***	0.53***	0.51***	0.58***
PAM	0.71***	0.75***	0.72***	0.76***	0.77***
IRC	0.82***	0.86***	0.81***	0.86***	0.86***

* p < 0.1, ** p < 0.05, *** p < 0.01.
Note. Spearman's rank correlations. *RCI*: Relationship Closeness Index, *SCI*: Subjective Closeness Index, *PAM*: Personal Acquaintance Measure, *IRC*: Index of Relationship Closeness. All scales can be found in Appendix B.1.

for each benchmark scale as well as a decomposition of the *RCI* into its three subcomponents (*frequency, diversity* and *strength*) in column 1. Columns 2 and 3 show the original findings by GST, whereas columns 4 to 6 display our results for the original IOS, *oneness*, and the extended IOS respectively.

When comparing both columns for GST (columns 2-3) and the same measures in our replication (columns 4-5) in Table 2.4, we find a higher correlations between *oneness* and all benchmark scales than for IOS ($p < 0.05$). Thus, we confirm the finding by GST that *oneness* outperforms the original IOS scale as a measure of relationship closeness. When examining our proposed extended IOS scale, by contrast, the difference to *oneness* disappears (columns 5-6). Correlation coefficients of our extended IOS scale are statistically indistinguishable from *oneness*, whilst performing significantly better than the original IOS scale.

Lastly, we also conduct a principal component analysis to compute an *Index of Relationship Closeness* (IRC) and report the correlation for each methodology of interest in the last row.[9] Extracting a principal-component across all benchmark scales provides a single nuanced index of relationship closeness as elicited by the entire collection of employed scales. Thus, instead of examining the scales independently as above, we can investigate the performance of IOS, *oneness* and extended IOS with respect to the *Index of Relationship Closeness* (IRC). The correlations in Table 2.4 confirm our results from the previous paragraph, extended IOS correlates more strongly to the IRC than the original IOS scale and the same as *oneness*.

Result 2. *Using the Index of Relationship Closeness (IRC) as a benchmark, the extended IOS scale measures relationship closeness significantly better than original IOS and as well as oneness.*

Finally, Table 2.4 shows that we replicate the findings from GST remarkably well (columns

[9]See Gächter et al. (2015) for details on the computation of the principal component.

2-5). In our replication we observe correlation coefficients that mimic the original results almost to the second decimal. The same accuracy in replication also holds true for all additional analyses conducted in GST.[10] A noteworthy aspect of this finding is that we utilised a different subject population (US vs. UK) and a significant amount of time has passed since the original study (2014 vs 2021). Despite these differences we successfully replicate all results from GST to an astounding extent, which leads us to our third and last result.

Result 3. *We successfully replicate all findings as well as effect sizes from Gächter, Starmer and Tufano (2015).*

4 Conclusion

In this study we propose a new measurement tool for *relationship closeness* that improves the accuracy of the IOS scale, while retaining its easy implementability. Following the same empirical strategy as GST, we find compelling evidence that our extended IOS scale does indeed capture social cohesion to the same extent as *oneness*. It seems that subjects are not influenced by the extended range of answers and score their subjective relationship solely based on the overlap of the circles. We therefore confirm our initial conjecture that extending the range of answers is sufficient to improve the quality of measurement. This allows us to propose a single measurement tool that is computerised, portable, and easy to implement whilst ensuring high accuracy in measurement.

Moreover, we also replicate the original study by GST reproducing their results with a remarkable degree of accuracy. Despite using a different subject population and collecting data multiple years later, we almost precisely replicate all coefficients. Thus in addition to validating our proposed tool, we also provide more evidence demonstrating the validity and reliability of using the (extended) IOS scale as a measure of social cohesion.

[10] The only exception is that the extended IOS scale only marginally correlates with the 'We' scale. This however supports the redundancy of the 'We' scale when using the extend IOS tool.

Bibliography

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), 596.

Berscheid, E., Snyder, M., & Omoto, A. M. (1989). The relationship closeness inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology*, 57(5), 792.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy–altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology*, 73(3), 481.

Dimant, E. (2021). Hate trumps love: The impact of political polarization on social preferences.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, 18(2), 192.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4), 1645–1692.

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5), 1858–1876.

Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the 'inclusion of the other in the self' scale. *PloS one*, 10(6), e0129478.

Molleman, L. & Gächter, S. (2018). Societal background influences social learning in cooperative decision making. *Evolution and Human Behavior*, 39(5), 547–555.

Paulhus, D. L. & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, 60(2), 307.

Pellencin, E., Paladino, M. P., Herbelin, B., & Serino, A. (2018). Social perception of others shapes one's own multisensory peripersonal space. *Cortex*, 104, 163–179.

Qualtrics (2021). *The code for this paper was generated using Qualtrics software. Copyright © 2021 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com.*

Robson, M. (2021). Inequality aversion, self-interest and social connectedness. *Journal of Economic Behavior & Organization*, 183, 744–772.

Rubin, Z. (1970). Measurement of romantic love. *Journal of Personality and Social Psychology*, 16(2), 265.

Starzyk, K. B., Holden, R. R., Fabrigar, L. R., & MacDonald, T. K. (2006). The personal acquaintance measure: a tool for appraising one's acquaintance with any person. *Journal of Personality and Social Psychology*, 90(5), 833.

Tarr, B., Launay, J., & Dunbar, R. I. (2016). Silent disco: dancing in synchrony leads to elevated pain thresholds and social closeness. *Evolution and Human Behavior*, 37(5), 343–349.

Westlund, J. M. K., Park, H. W., Williams, R., & Breazeal, C. (2018). Measuring young children's long-term relationships with social robots. In *Proceedings of the 17th ACM conference on interaction design and children* (pp. 207–218).

Chapter 3

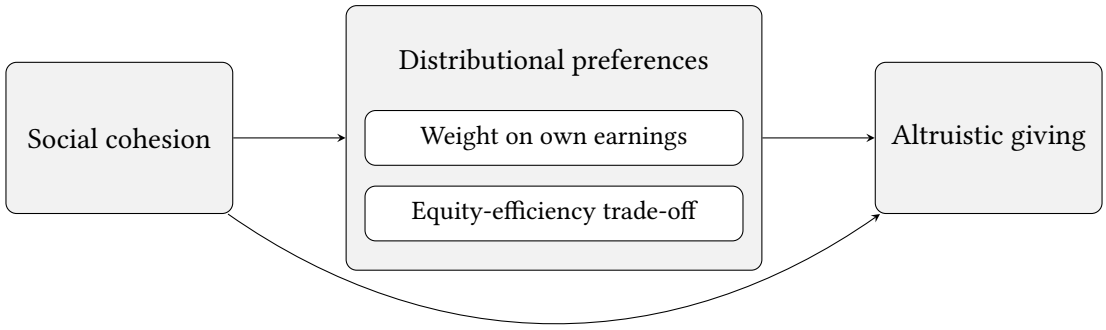
The role of social cohesion as a determinant of altruistic behaviour

1 Introduction

Having outlined and developed two methodologies in the previous chapters, we now apply both of these by returning to the research question introduced in the beginning of Part I. As a recap, we consider *social cohesion* to play a crucial role in understanding heterogeneous altruistic behaviour. Moreover, we regard distributional preferences to serve as a fundamental mediating factor between these two dimensions. For this reason, in this chapter, we isolate the underlying mechanism between social cohesion, distributional preferences and altruistic giving. To do so we first elicit a cardinal measure of relationship closeness between university students using our *extended IOS scale* developed in Chapter 2 and construct a *weighted altruistic friends network*. Using this network, we estimate precise distributional preference parameters with our *condensed AM/FKM* tool developed in Chapter 1 for different levels of social cohesion and test whether they mediate the relationship between *relationship closeness* and altruistic giving.

In particular, we expect social cohesion to affect altruistic giving both, directly and indirectly, via the channel of distributional preferences. Our hypothesis is thus that the latter is a crucial mediator in explaining how relationship closeness translates into altruistic giving. Figure 3.1 illustrates the hypothesised relationship between our key dimensions of interest. Our design allows us to identify each element and relationship depicted in Figure 3.1 with high accuracy, providing us with a deeper understanding of the underlying mechanisms. Our

Figure 3.1: Hypothesised relationship between cohesion, preferences and giving



This chapter is based on joint work with Simon Gächter, Chris Starmer & Fabio Tufano. I thank members of CeDEx for comments and inputs. The studies in this chapter are funded by the British Academy.

study draws on Leider et al. (2009) who ask subjects to perform allocation decisions for different levels of social distances and degrees of anonymity. In their study *social distance* is defined as the distance between two individuals in the social network measured as the number of link between them. Anonymity on the other hand refers to anonymous payments and not to the identity of the recipient. This methodology allows us to disentangle different dimensions of altruism, namely *baseline altruism*, *directed altruism* and the effect of a *future prospect of interaction*. Whilst following their design by employing a series of within-subject measures, we are utilising a much more nuanced tool to identify individual distributional preferences. As described in Chapter 1, the *condensed* AM/FKM tool allows to not only capture individual's subjective weight on other's earnings but also their preference when trading off equity and efficiency. Especially in light of economic networks and interactions between socially close individuals the latter dimension seems crucial, as one could imagine that on the one hand more cohesive networks will share resources more equally, but on the other hand their ability to coordinate can ensure that they better exploit potential efficiency gains. For this reason we think our methodology introduced in Chapter 1 is of particular purpose within our set-up.

We find that subjects in our elicited network have substantial heterogeneity in social cohesion as well as altruistic preferences and giving. Overall, we find strong evidence for robust individual altruistic considerations, where subjects that have higher levels of *baseline altruism*, also show higher levels of *directed altruism*. Moreover, we find social distance directly affects distributional preferences as well as generosity, however we only find a small effect of IOS as our cardinal measure of social cohesion. Lastly, we find that distributional preferences do serve as a substantial mediating factor in the relationship between social cohesion and altruistic giving.

Our study contributes to various strands of the existing literature. It directly relates to the extensive research on distributional preferences (see Forsythe et al., 1994; Hoffman et al., 1994, 1996; Andreoni & Miller, 2002; Charness & Rabin, 2002; Fisman et al., 2007; Charness & Gneezy, 2008, among others) and their relevance in economic networks (Leider et al., 2009; Goeree et al., 2010; Brañas-Garza et al., 2010; Ligon & Schechter, 2012; Bourlès et al., 2017, 2021). Moreover, we extend previous studies, by including the dimension of measured social cohesion to further understand behaviour between individuals (Gächter et al., 2019; Robson, 2021). As we are especially interested in the role of underlying distributional preferences, our study is strongly inspired by the seminal work of Andreoni & Miller (2002) and Fisman et al. (2007). However, as we are interested in preferences individuals hold towards *identifiable* others, we slightly adjust their method and instead of matching them with a stranger, we do reveal the identities of recipients to elicit *directed altruism*.

Exploring preferences towards identifiable others relates to work by Bohnet & Frey (1999), Charness et al. (2007), Charness & Gneezy (2008), Leider et al. (2009), Goeree et al. (2010) and Brañas-Garza et al. (2010), who all, in various ways, find that decreasing anonymity or social distance generally results in higher levels of generosity towards recipients. In particular, Leider et al. (2009) as well as Goeree et al. (2010) are relevant for our study as they also examine altruistic giving in an elicited real-world network. Both studies find that generosity increases as a function of reduced social distance. However they do not provide additional measures of

social cohesion or the estimation of distributional preference parameters. We close this gap and thus allows for a deeper understanding of what determines altruistic giving.

By looking at a social network, our study also relates to work by Brañas-Garza et al. (2010), who observe that the position of individuals within a network significantly relates to revealed altruistic actions. This is also in line with recent theoretical contributions by Bourlès et al. (2017, 2021), who present an equilibrium analysis of altruistic transfers in economic networks. In their model, the embeddedness of an agent within a network, as well as the altruistic weight to their neighbours explain transfer levels. Their model showcases the importance of exploring heterogeneity in altruistic considerations between individuals to understand other-regarding behaviour and resource allocations. As outlined above, one dimension that we deem crucial for these altruistic considerations is social cohesion. Demonstrating the impact of social cohesion, as measured by *oneness* (Cialdini et al., 1997; Gächter et al., 2015), Gächter et al. (2019) find that higher levels of group cohesion are a significant component in overcoming coordination failure between individuals. In their study Gächter et al. (2019) also provide some evidence that social preferences increase with higher levels of cohesion, pointing towards a relationship between social cohesion and preferences.

Taken together the studies discussed so far suggest, that social cohesion, distributional preferences and altruistic giving in economic networks are closely tied. However, thus far there has only been one piece of research bringing parts of these dimensions together within a similar design to ours. In a recent publication, Robson (2021) explores the relationship between relationship closeness and distributional preferences, finding a significant relationship between the two components. Like our study, he utilises a methodology inspired by Fisman et al. (2007) as a measure of social preferences exploring a three-person set-up in a lab-in-the field experiment in Uganda. However, the environment in our study is more complex as we elicit a whole social network instead of only looking at the relationship between separate individuals. Our findings, thus both shed light on the robustness of his results, and extend them by integrating an economic network as well as employing independent elicitations of distributional preferences for the same individual with varying recipients.

The following chapter is structured as follows. In Section 2, we provide a detailed overview of our experimental design. We thereby first explain how we elicit the social network, followed by the measurement of social cohesion, distributional preferences and altruistic behaviour. Following that, we present our treatment conditions and hypotheses. In Section 3.1, we describe our experimental procedure and data collection. In Section 4, we present our results and in Section 5 we provide concluding remarks.

2 Design

2.1 Constructing a social network

The first element required to answer our research question, is to identify different degrees of social cohesion. As we aim to explore the real-world impact of social cohesion on altruistic actions, we want to identify social cohesion between individuals as it has naturally emerged. To do so, we follow Leider et al. (2009) and elicit a real-world friends network. This means,

we abstain from inducing any levels of cohesion by means of making similarities salient or using minimal group paradigms. Instead, all observed differences in social cohesion are consequences of real world interactions and experiences (or the lack thereof) among individuals before participating in our study.

To construct our network, we follow a standard methodology from previous research (see Leider et al., 2009; Goeree et al., 2010; Krishnan & Sciubba, 2009; Brañas-Garza et al., 2010; Banerjee et al., 2013, among others).[1] We approach a large body of students from the University of Nottingham, asking each to list 10 names of fellow students within their study programme and year. Even though this cannot be guaranteed, we assume that students will have some close relationships to others within their study programme.[2] Having elicited this information, we can subsequently link students that named each other during the elicitation within a network structure and therefore construct and visualise the students' friendship network for a given programme and year.

2.2 Measuring cohesion, distributional preferences and altruistic giving

Contrasting previous network elicitations we also ask subjects to provide an IOS Score for each of the 10 names using the extended IOS scale developed in the previous chapter. The IOS rating then serves as the cohesion weight of each link in our elicited social network, allowing us to obtain considerably more information than previous network elicitations and identify a *weighted* friendship network (Goyal, 2005). Since we are asking for a total of 10 names we ensure a sufficient degree of variation in the IOS scores to explore the effect of different social cohesion levels on altruistic giving.[3]

Having outlined our approach to identify different levels of social cohesion we next describe how we elicit preference parameters and altruistic giving. To test whether other-regarding preferences mediate the effect of social cohesion on altruistic giving, we estimate precise preference parameters for each subject, using our condensed AM/FKM methodology presented in Chapter 1. In addition, to explore altruistic giving, all subjects participate in three dictator games as in Leider et al. (2009) with the exchange rates between own and other's tokens being either 3:1, 1:1 or 1:3.

2.3 Treatment conditions

Our treatments follow a 2x3 within-subject design, varying both the social distance between senders and receivers in the AM/FKM task and Leider et al. (2009) dictator games, as well as addressing reciprocity by manipulating the *prospect of future interaction.* Let us first consider the dimension of social distance.

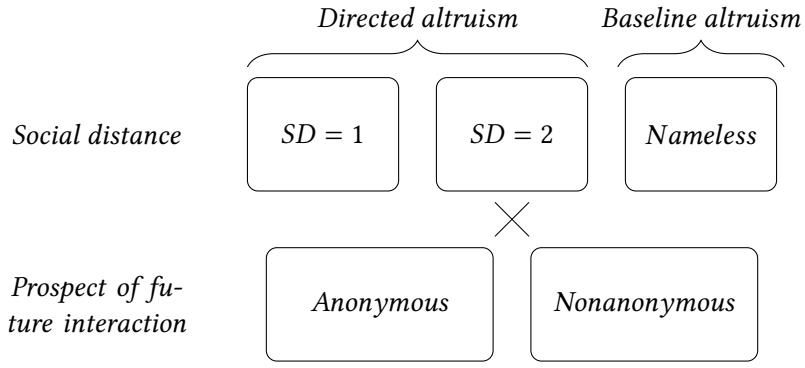
The elicited network allows us to match subjects for the AM/FKM task and dictator games across different degrees of social distance, creating an exogenous variation in social cohesion.

[1] Brañas-Garza et al. (2017) provide an overview of methodologies to elicit social networks.
[2] In case a subject cannot list 10 names, we randomly select other students for the remaining ones such that each student is ultimately linked to 10 others.
[3] We find large heterogeneity in IOS scores as will be shown in the results.

We thereby replicate Leider et al. (2009) who use social distance as their proxy of social cohesion. As in their work we define a social distance equal to 1 ($SD = 1$) if sender and receiver are *directly* linked within the network. Following that, a friend of a friend is defined to have a social distance of 2 ($SD = 2$), and so on. As we are particularly interested in the effect of social cohesion between directly linked individuals using our IOS Score, our study especially focuses on $SD = 1$. However, to compare our more nuanced IOS measure with social distance as a proxy of social cohesion, we replicate Leider et al. (2009) and explore $SD = 2$ as well. Both $SD = 1$ and $SD = 2$ capture altruism towards named individuals (*directed altruism*). To complement this and also get a baseline measure for altruism and social preferences as well as to be able to compare our findings with the literature (in particular AM/FKM), we also match each participant with a *nameless* other student (*baseline altruism*). Each subject in our study thus participates in three allocation scenarios with distinct partners. One direct friend ($SD = 1$), one friend of a friend ($SD = 2$), and one *nameless* partner.

As mentioned above our second treatment dimension is the *prospect of future interaction*. Given that participants know each other and are likely to interact outside the study, it is important for us to control for this aspect. We do so by varying the degree of information that is shared with the recipient (*anonymous* vs. *non-anonymous*).[4] In the *anonymous* treatment the recipients are neither informed about the tokens allocated by the sender, nor who the sender is. In contrast, in the *non-anonymous* situation both of these details are disclosed.[5] As a consequence, recipients could reward/punish sender's decision in any potential future interaction. Relaxing the degree of anonymity does not only allow us to assess the effect of a *potential future interaction*, but also captures altruistic relationships as they occur in real-life. In everyday interactions, full anonymity is barely present, especially when considering altruistic actions between individuals connected in a social network. Using a real social network and relaxing the degree of anonymity are thus both elements of our design that help us to generalise

Figure 3.2: Overview of treatment conditions



[4] As in Leider et al. (2009) anonymity here does *not* refer to the disclosure of identities. Instead, it refers to information regarding choices and payments when receiving the bonus paid for the study.

[5] As subjects complete a series of allocation decisions and one is randomly selected for payment it is impossible to infer the tokens allocated directly from the payment received. Thus, in the *non-anonymous* treatment we inform recipients about the identity of their sender, the tokens allocated and the corresponding earnings when processing the payments.

findings to real-world interactions.

To summarise, in total each subject completes the *condensed* AM/FKM task and three dictator games for six different scenarios that vary both, the social distance between sender and recipient ($SD = 1$, $SD = 2$ and *nameless*) and the *prospect of future interaction* (*anonymous*, *non-anonymous*). Figure 3.2 provides an overview of all treatments. To avoid overwhelming participants with the number of required choices, we structure our data collection across three waves. The exact procedure of each wave will be explained in more detail in Section 3.1.

2.4 Hypotheses

As mentioned above, we expect both a significant direct and indirect impact of social cohesion on altruistic giving. Moreover, we consider distributional preferences to play a substantial mediating role in that relationship. In line with our schematic overview provided in the introduction (Figure 3.1), we now formulate hypotheses regarding each of the relationships below.

1. We expect higher levels of social cohesion (as measured by IOS or proxied as social distance) to correspond to higher generosity in the dictator games.

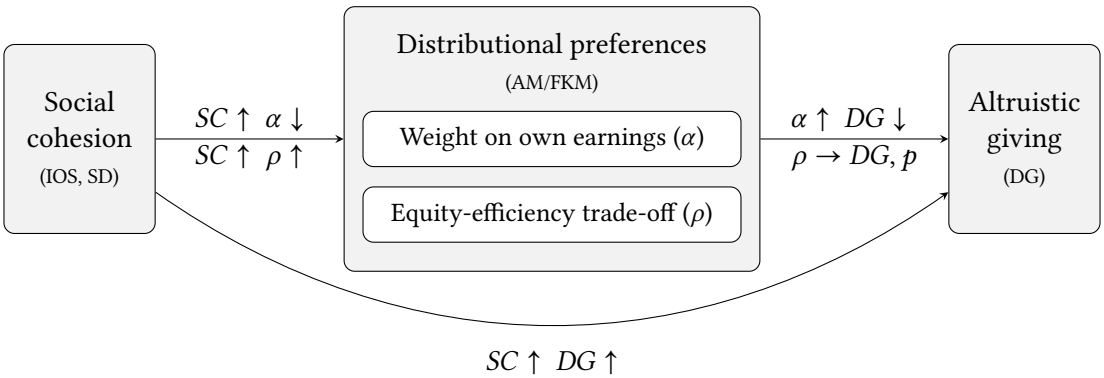
2. Increased social cohesion is associated with stronger other-regarding preferences. Specifically, for socially close individuals, we expect lower values of α and higher values of ρ, which is due to successfully exploiting efficiency gains.

3. With respect to the relationship between distributional preferences and altruistic giving, we expect lower levels of α to correspond to increased altruistic giving in the dictator games. ρ, on the other hand, explains diverse choices in the dictator games across the three exchange rates.

4. The *prospect of future interaction* leads to higher altruistic preferences as well as more generosity in the dictator games across all social distances.

Furthermore, we expect to replicate all the elements we borrow from Leider et al. (2009). Figure 3.3 provides the same schematic overview as in the introduction, now including our

Figure 3.3: Hypothesis regarding social cohesion, preferences and giving



Note. SC = Social cohesion as measured by IOS or Social Distance (SD). DG = Dictator game giving. p = three different price levels in the dictator games. Arrows correspond to increases or decreases for each relationship.

Hypotheses 1-3. Each arrow in Figure 3.3 thereby presents a relationship between *social cohesion*, *distributional preferences* and *altruistic giving*. As for the hypotheses on the respective arrows, $SC \uparrow \ \alpha \downarrow$, for example, captures that we expect increased social cohesion (SC) to be associated with lower levels of α.

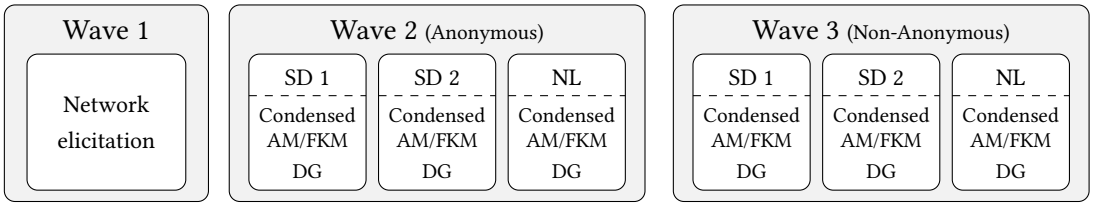
3 Procedure

3.1 Experimental waves

As the experimental components of this study ask subjects to make many independent allocation decisions, we decided to span the data collection across three distinct waves over a 4-week period. In the first wave, we elicit the social network as described in Section 2. Subjects are incentivised to list names of friends that are also likely to name themselves in return. Every time two subjects list each other during the network elicitation, each of them receives a bonus of £0.50 with probability $1/2$.

After constructing the network we re-invite subjects twice more to complete the actual experimental tasks. In both Wave 2 and Wave 3, which took place one week apart, subjects perform the AM/FKM task and dictator games for all three levels of social distance outlined in Section 2. What varies between the two waves is the *prospect of future interaction*, we thereby randomise whether participants first take decisions in the anonymous or non-anonymous treatment.[6] Figure 3.4 gives an overview of the different waves.

Figure 3.4: Experimental overview



Note. We randomise whether subjects first face the anonymous or non-anonymous treatment. SD = Social Distance. NL = Nameless. Condensed AM/FKM describes our proposed methodology developed in Chapter 1.

To match senders and receivers across the different levels of social distance, we proceed as follows. For a direct friend, we randomly choose one of the 10 names listed during the elicitation. Identifying friends of friends (SD = 2) is somewhat more complicated, as depending on the completeness of the network there might not be another participant that is a friend of a friend. Therefore, we first identify whether we can find a line network of friends. If this does not exit, we look for a star network instead. The difference between these network is depicted in Figure 3.5 below with the arrow indicating which subject listed the other.

[6] Meaning that if a subject was assigned to the *anonymous* treatment in Wave 2, they now complete the *non-anonymous* treatment in Wave 3 and vice versa.

Figure 3.5: Line and star network cases of $SD = 2$

For our study, we consider both of these possible scenarios cases of $SD = 2$, but if available, we match a partner based on the *line network* case. As for $SD = 1$, if multiple matches exist, we use a random number generator to select one. Altogether, in each wave subjects participate in a total of 69 allocation decisions for three distinct partners, two of which are *named* and one is *nameless*.

To calculate payments at the end of each wave, we follow the standard procedure of AM/FKM. First, for each individual we draw one social distance at random. We then randomly select one allocation decision[7] and implement the decisions for the matched partner. All tokens are then converted at a rate of 20 tokens = £1. We ensure that each subject is exactly once a sender and once a receiver.

3.2 Recruitment

We present the results from a pre-registered study, administered in November and December 2021.[8] To elicit our social network, we invited all 2^{nd}- and 3^{rd}-year students from the Business School of the University of Nottingham via email. We thus obtain two separate networks, one for 2^{nd}-year and one for 3^{rd}-year students. The students were able to complete wave 1 at any point over a 10-day span (Nov, 8^{th} to Nov, 17^{th}), using their personal computer or mobile devices. We informed subjects at the beginning of Wave 1 that all payments will be processed after the end of all three waves, irrespective of their participation in Wave 2 and 3. For each wave subjects received a participation fee of £3, plus bonus payments as described above. To minimise attrition, we moreover set up a lottery at the end of Wave 3, where students could win one of five £100 cash prizes. For the completion of each wave students thereby received an increasing number of lottery tickets. Wave 1 yielded one ticket, whereas completing Wave 2 won students another five, and Wave 3 another 15 lottery tickets. For the latter two waves, we only invited students back that completed Wave 1.

The experiment was approved by the Nottingham School of Economics' Research Ethics Committee and programmed using LIONESS Lab (Giamattei et al., 2020), providing subjects with a simple interactive interface for all tasks.[9] Payments were made using PayPal and students had to register with the CeDEx participant database in order to complete the study.[10]

[7] Out of the 23 total dictator games.
[8] Registration number AEARCTR-0008485. See https://www.socialscienceregistry.org/trials/8485.
[9] Instructions for all waves can be found in Appendix C.2.
[10] On top of obtaining informed consent during the study itself, the registration with the CeDEx database ensures an additional level of informed consent.

In addition to the tasks described in Section 2, the study also included some additional elements, that were independently incentivised and which are not part of this thesis.[11] Across all waves the median completion time per subject was 53 minutes and median payment excluding the five cash prizes was £11.[12] Overall, 118 students completed the network elicitation. A further 63 completed Wave 2 and 56 ultimately participated in Wave 3, implying an overall attrition of about 55% despite our use of lottery ticket incentives. Due to budgetary restrictions we were not able to recruit a larger sample at this stage, however, upon access to additional funds, we plan to extend the data collection in spring 2022.[13]

4 Results

Our results are structured as follows. We first outline network characteristics and elicited social cohesion for 2^{nd} and 3^{rd}-year students. Following that we examine the within-subject robustness across our measures, as we elicit distributional preferences as well as altruistic giving for three different levels of social distance and two degrees of *anonymity*. We then turn to our main research question by examining the relationship between *social cohesion*, *distributional preferences* and *altruistic giving* as depicted in Figure 3.1. For that we provide results for each relationship independently, before addressing them simultaneously by conducting a mediation analysis.

4.1 Network characteristics and reported social cohesion

Let us first outline descriptives of the elicited networks for 2^{nd}-year and 3^{rd}-year students. Figure 3.6 illustrates both networks graphically, where students who participated in the study are coloured in red, and students that were *only named* are depicted in black. Within 2^{nd}-year, 77 students participated in our experiment and listed 362 other students. Compared to that, 41 3^{rd}-year students participated and named another 229 students.

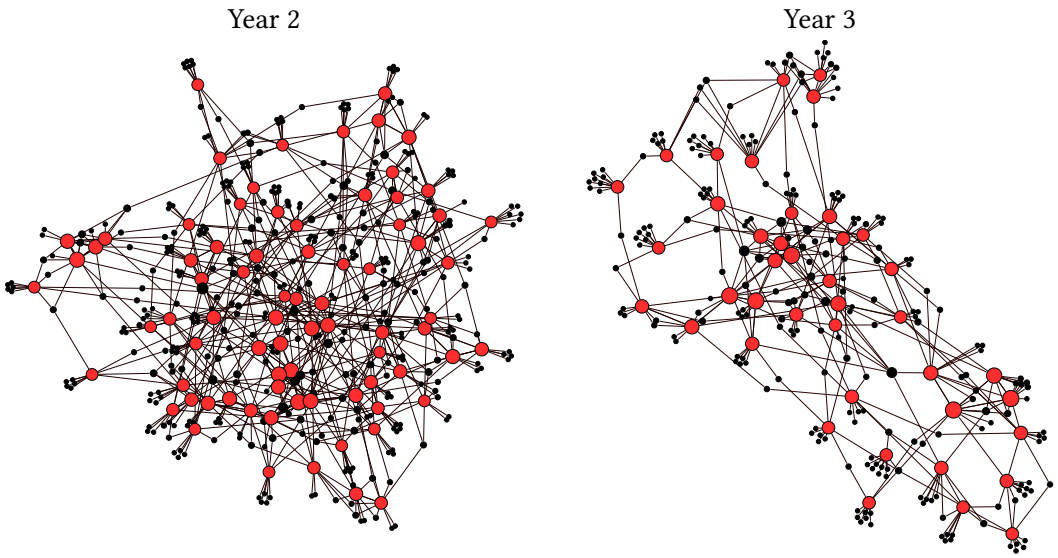
When examining the networks in Figure 3.6, we can see that the network of 2^{nd}-year students is much more tightly connected, which is mainly driven by the different number of participants. Our elicited networks only cover a small subset of the actual student population (2^{nd}-year: 663, 3^{rd}-year: 524), implying that most *named* students did not participate in the experiment themselves. This is also reflected in the low average number of links per students of 1.75 and 1.52 for the two respective networks. However, despite this there are still some students that were named up to 9 (7) times. Thus, already within these relatively small networks, we observe quite some heterogeneity in terms of how connected students are within their programme of study. This heterogeneity is also present when exploring reported IOS scores, which can be considered the *weight* of each network link. We find that IOS scores are

[11] In particular, we elicited transfers in more complex network structures and normative evaluations of giving. The analysis of these parts will constitute a separate project.

[12] With the inclusion of cash prizes, the median payment was £15.40.

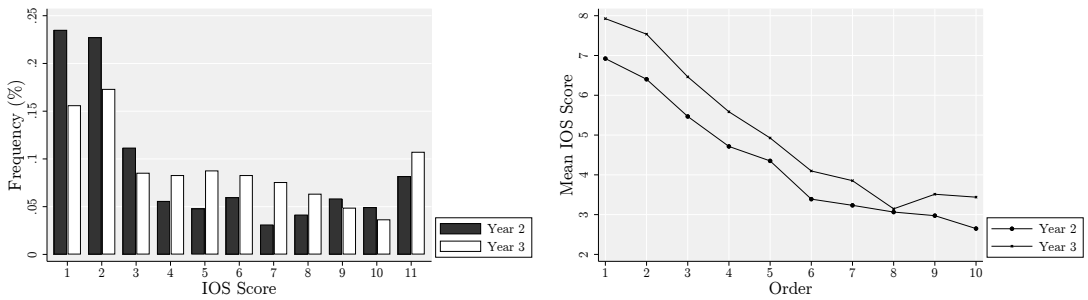
[13] For the extended data collection, we will thus consider different channels through which we can motivate more students to participate in the elicitation, for example by more frequent reminders in lectures and tutorials. Moreover, we plan to provide additional time for the network elicitation, and increase incentives for subsequent waves.

Figure 3.6: Elicited social networks



somewhat bimodal, meaning subjects tend to name a close friend with a high IOS score as well as list students with whom they have low levels of cohesion (see left graph, Figure 3.7). This is also supported when examining the trend of IOS scores across the 10 listed names. The right graph of Figure 3.7 plots the average IOS score as subjects reported a score for each of their listed names. The figure clearly shows, that subjects on average begin to name other students they consider closest and divert to lesser known students towards the end of their selection. This supports our design decision that 10 names are sufficient to elicit a substantial part of a student's social network within their respective programme and year. Moreover, despite offering students the opportunity to list fewer than 10 names, we find that subjects on average list 8.33 names with 70% of the sample opting to list all 10. Lastly, from Figure 3.7 we can also infer that on average 3^{rd}-year students report higher levels of cohesion than 2^{nd}-year students. This also seems intuitive, as 3^{rd}-year students had more time to develop relationships amongst each other and they also experienced their first year before the Covid-19 pandemic, which seems to be reflected in their reported levels of social cohesion.

Figure 3.7: Heterogeneity of IOS Scores



Note. Left: IOS scores from our extended IOS scale across years. Right: Mean IOS scores as listed during the task. Order corresponds to whether a name was listed earlier (left) or later (right).

4.2 Within-individual robustness of measures

After having outlined network properties and having described varying levels of social cohesion within the network, we now explore our measures of distributional preferences and altruistic giving. Since we elicit preference parameters as well as altruistic giving within-subject, we can analyse their correlation across different social distances and anonymity levels (see Table 3.1). On the one hand, this analysis provides us with information regarding the robustness of our measurements and on the other hand we can explore within-subject stability of altruism. Table 3.1 shows Pearson's correlation coefficients for α, ρ and dictator game giving across social distances (top panel) and anonymity (bottom panel). Altogether, Table 3.1 shows significant and substantial within-subject correlations for both treatment dimensions, social distance as well as the prospect of future interaction. Since distributional preferences and altruistic giving are elicited for different partners and anonymity levels, the significant correlations do provide evidence in favour of stable types of altruism. In particular, the significant coefficient for $SD = 1$ and $SD = 2$ with respect to *nameless* shows a significant relationship between *baseline* and *directed* altruism. Thus, in line with Leider et al. (2009), we therefore conclude that

Result 1. *Individuals with higher baseline altruism, also have higher levels of directed altruism independent of the prospect of future interaction.*

Furthermore, the significant correlations across anonymity levels support our proposed methodology. As these parameters are elicited for the same sender-receiver pairs, one week apart, these correlations encompass some degree of test-retest reliability in distributional preferences. The significant coefficients therefore provide additional evidence of the robustness of our *condensed* AM/FKM task.

Demonstrating the robustness of our measures, we now turn to our main research ques-

Table 3.1: Within-subject correlations across preference parameters

	SD = 1			SD = 2		
	α	ρ	DG	α	ρ	DG
SD = 2	0.381***	0.850***	0.560***	1	1	1
	(97)	(65)	(338)	(101)	(68)	(339)
Nameless	0.282***	0.787***	0.514***	0.333***	0.842***	0.664***
	(99)	(68)	(346)	(94)	(63)	(334)
	Anonymous					
	α	ρ	DG			
Non-Anonymous	0.335***	0.606***	0.553***			
	(127)	(86)	(156)			

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Within-subject correlations between preference parameters and dictator game giving for social distances and anonymity. In each cell, α is displayed in the first column, ρ in the second and dictator game giving in the third. The top half of the table pools observations across anonymity, the bottom pools across social distance. All results also hold when decomposing the sample across both dimensions. Results for ρ are only shown for cases of $-1 \leq \rho \leq 1$ in line with evidence from the previous chapter. Number of observations in parenthesis. For brevity, we exclude columns that show a perfect correlation between both *nameless* treatments, as well as both *non-anonymous* treatments.

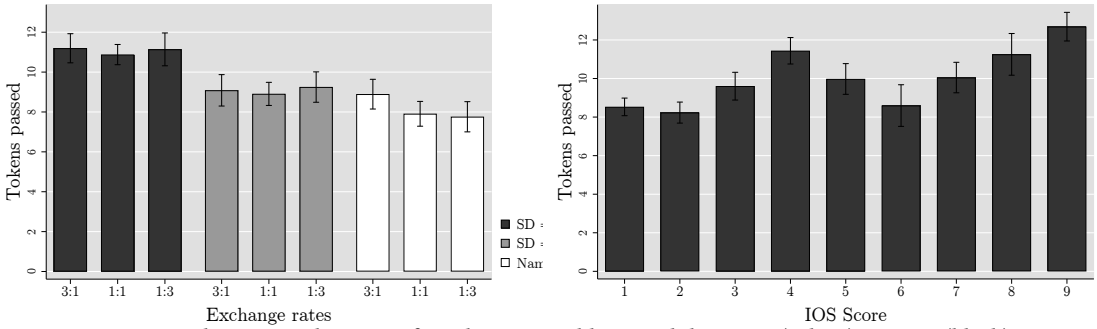
tion. In line with our hypotheses, we structure our analysis closely following the considered schematic relationships between social cohesion, distributional preferences and altruistic giving (see Figure 3.1). We examine each relationship sequentially, beginning with the effect of social cohesion on altruistic giving below. This is followed by exploring the role of social cohesion as a determinant of distributional preferences. For both analyses, we first consider social distance as a proxy for social cohesion and then examine IOS scores. As we only measure IOS scores for $SD = 1$, our analyses of IOS levels as nuanced measures of social cohesion can be considered a decomposition of $SD = 1$. After both of these sections we then analyse the impact of distributional preferences on altruistic giving, before bringing all elements together by means of a mediation analysis.

4.3 Social cohesion and altruistic giving

Recall that each subject faced three dictator games with varying exchange rates (3:1, 1:1, and 1:3) for each social distance and anonymity level. To ensure that efficiency concerns are easily comprehensible and entirely transparent to all participants, we displayed all decisions in the same graphical interface as the AM/FKM allocations. Mimicking the choices by Leider et al. (2009), each subject could pass on a maximum of 30 tokens at each distinct exchange rate.[14] Thus, in case of an equal split of tokens between sender and receiver, we would observe 15 tokens passed.

As introduced above, let us first consider social distance as a proxy of social cohesion in Figure 3.8, left graph. The graph shows average number of tokens passed for all social distances (indicated by colour) and exchange rates. We can observe that higher levels of social cohesion (lower social distance) directly relate to an increase in tokens passed.[15] We can however, find no difference in generosity across the three efficiency levels. If at all, we find a reduction in passed tokens for an inefficient exchange rate (1:3) for a *nameless* partner, which contradicts the finding by Leider et al. (2009), who show increased generosity when efficient. Thus, pool-

Figure 3.8: Passed tokens in dictator games



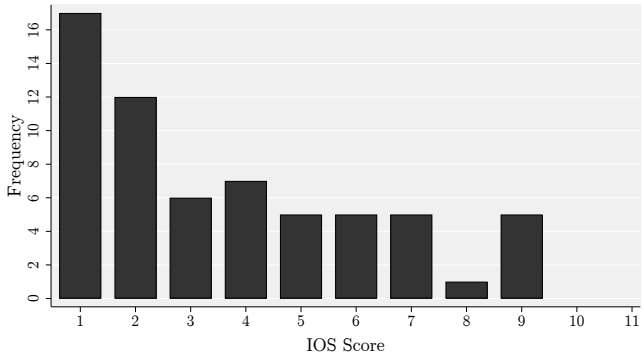
Note. Maximum tokens passed is 30. Left: Tokens passed by social distances (colour). SD = 1 (black), SD = 2 (grey), Nameless (white). Exchange rates on horizontal axis. Right: Tokens passed by IOS level only for SD = 1.

[14] Resulting in maximum tokens to self and others of 90:30, 60:60 and 30:9 for 3:1, 1:1 (2:2) and 1:3, respectively.
[15] See Appendix C, Table C.1.1 for results of Wilcoxon Rank Sum tests.

ing across efficiency levels, subjects on average pass 11.07 tokens (37%) to direct friends which is significantly more than they pass to friends of friends (9.08 tokens) or *nameless* partners (8.18 tokens). In fact, the difference in tokens passed between $SD = 2$ and *nameless* partners is considerably smaller than between $SD = 1$ and $SD = 2$. As we are expecting a reduction in tokens passed for lower social cohesion, the results suggest that subjects consider *named* recipients with $SD = 2$ almost identically to *nameless* students. This also supports our design decision to not explore any further social distances beyond $SD = 2$.[16]

Let us now focus on our more nuanced measure of relationship closeness within $SD = 1$. The right graph in Figure 3.8 can thus be considered a decomposition of $SD = 1$, plotting average numbers of tokens passed across different levels of IOS. All in all, the graph only shows marginal effects of IOS scores on altruistic giving. We do find a positive coefficient ($\hat{\beta} = 0.27$) in a regression model, when estimating the effect of IOS on tokens passed, however this is statistically insignificant ($p = 0.18$).[17] A potential reason for the insignificance is a lack of variation in our observed IOS scores. After attrition between the first and the subsequent waves, we only observe 9 out of 11 distinct IOS levels (1 to 9), with a highly right skewed distribution as can be seen in Figure 3.9. When comparing this to the left graph in Figure 3.7 that shows the initial distribution of IOS scores in Wave 1, we can see that matches in Wave 2 and 3 were indeed drawn randomly as the two distributions closely resemble each other. However, this led to a lack of observations for most IOS levels (except IOS = 1) suggesting insufficient statistical power to analyse *closer* relationships. In particular, the large number of observations with the smallest possible IOS score of 1 are worth noting. When examining the data more closely, we can see frequent cases where participants evaluated all listed names with an IOS score of 1, whilst still selecting specific names from the list of students. This might

Figure 3.9: IOS Scores for matched subjects in wave 2 and 3 ($N = 63$)



Note. No observations for IOS scores of 10 and 11. Distribution similar to underlying distribution in Figure 3.7.

[16] In Leider et al. (2009) the authors explore social distances all the way up to $SD = 5$ finding if at all marginal effects in generosity between high social distances.
[17] The regression result can be found in Appendix C, Table C.1.2.

indicate a lack of comprehension or motivation to complete the experiment thoroughly.[18] In fact, when repeating the regression analysis from above excluding IOS scores equal to 1, we find a doubling in the coefficient from 0.26 to 0.55, which is now also statistically significant ($p < 0.05$) (Appendix C, Table C.1.3). This is striking, as we find a stronger relationship between social cohesion and altruistic giving when exogenously reducing the already limited variation in IOS scores even further. As this result is rather counter-intuitive it suggests that in fact observations with an IOS score of 1 do entail additional degrees of noise. Moreover, this analysis also suggests that the statistically insignificant results from above are indeed a result of a lack of power.

With respect to the *prospect of future interaction*, we find increases in tokens passed when payments are *non-anonymous*. The effect is stronger for $SD = 2$ and *nameless*, which can likely be explained by a ceiling effect. Even though subjects could in theory pass 30 tokens, in practice the ceiling is likely an even split of 15 tokens and for $SD = 1$ subjects already pass 10.4 tokens when payments are *anonymous*.[19] To summarise, we do find evidence that higher levels of social cohesion do in fact lead to more generosity in dictator games. We replicate the findings by Leider et al. (2009) that lower social distances as proxies for social cohesion significantly relate to a reduction in tokens passed. This finding is further decomposed by the use of IOS scores as more nuanced measures of *relationship closeness*. Elicited IOS scores also indicate some evidence for increased tokens passed for *closer* relationships, even though this evidence is not as robust as our result for social distance.

Result 2. *Higher levels of social cohesion correspond to increased levels of altruistic giving. This is robust for different social distances but only weakly present across IOS scores within $SD = 1$.*

4.4 Social cohesion and distributional preferences

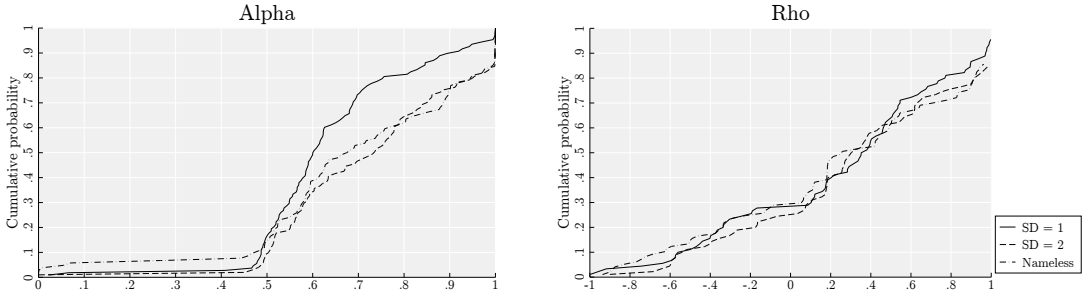
Let us now turn to elicited distributional preferences. Overall, we hypothesise similar findings for distributional preferences as we observed for altruistic giving, however we can now decompose the effect of social cohesion with respect to both, α and ρ. Recall that α captures the relative weight of own earnings and ρ preferences for equality or efficiency. As in the previous section, let us first examine the effect of social distances on the two preference parameters. Figure 3.10 presents cumulative distributions for both parameters (α left, ρ right) across social distances.

As a first observation, across all social distances we replicate the finding from our methodological contribution as well as Fisman et al. (2007) that most participants have an $\alpha < 0.5$, indicating that subjects place at least an equal relative weight on own earnings. This also supports our hunch above, that indeed 50% of tokens passed can be considered a ceiling. Moreover, continuing with the analysis of α, we can see a different distribution of parameters with re-

[18] This is however unlikely to only come from the fact that the IOS scale is not incentivised. If this were the case, then previous studies would not have found robust results using the same measure. It might be a combination of incentives and subjects facing a new somewhat tedious task as they had to indicate 10 distinct IOS levels on the same screen.

[19] 11.9 tokens when payments are *non-anonymous*.

Figure 3.10: Distribution of preference parameters across social distances



Note. Estimated parameters by social distance, α (left) and ρ (right). Observations restricted to $-1 \leq \rho < 1$.

spect to the closest social distance ($SD = 1$). For direct friends we observe more values close to 0.5 than in the case of a friend of a friend ($SD = 2$) or a *nameless* partner. This is in line with our hypothesis that increased social cohesion reduces the relative weight of own payoff. Similar to the reduced effect sizes found in the previous section, when examining distributional preferences we find no differences in α between $SD = 2$ and a *nameless* partner. We observe a different result for ρ, where we find no significant differences in distributions between any of the social distances. This is not in line with our expectation as we predicted higher levels of preferences for efficiency (higher ρ) for increased social cohesion.[20]

Table 3.2 again decomposes $SD = 1$ more finely using IOS scores by providing the results of a regression analysis estimating the relationship of IOS scores for each estimated preference parameter (α in models 1-3, ρ in models 4-6). Contrasting the results for social distance, we cannot find any effect of IOS on either dimension of distributional preferences. None of the estimated models show a relationship between the two measures, however for both parameters the signs of the pooled coefficients are in line with our initial hypotheses. They suggest a reduced weight on own earnings and higher preferences for efficiency with increased *rela-*

Table 3.2: The relationship of IOS and other-regarding preferences

	α			ρ		
	(1) Pooled	(2) Anony.	(3) Non-Anony.	(4) Pooled	(5) Anony.	(6) Non-Anony.
IOS Score	-0.000419	0.00399	-0.00538	0.00142	0.00862	0.0169
	(0.00776)	(0.0103)	(0.0110)	(0.0246)	(0.0248)	(0.0503)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	108	60	48	88	52	36
R^2	0.075	0.102	0.210	0.308	0.298	0.528

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Controls are *age, gender, ethnicity, study year, pro-social survey measures, IOS score for stranger, and experimental wave.* Pooled regression with standard errors clustered at the individual level. The difference in observations is due to a reduction in sample for ρ as we exclude and $\rho < -1$ in line with the previous chapter.

[20] This as well as all other findings from Figure 3.10 are also statistically confirmed using Kolmogorov-Smirnov tests in Appendix C, Table C.1.4.

tionship closeness. Nonetheless, in particular the coefficient on α is not robust as it switches signs when examining the effects for different levels of *prospect of future interaction* (models 2 and 3). Restricting our analysis again to IOS scores > 1, as motivated above, we do also in this case find a substantial increase in effect size from -0.0004 to -0.012, which however, also due to the reduced number of observations ($N = 76$), is still statistically insignificant ($p = 0.25$) (Appendix C, Table C.1.5).

In addition, exploring the direct effects of a *prospect of future interaction* on distributional preferences we cannot find statistically significant differences. This contradicts the result observed for altruistic giving, pointing towards the fact that estimated preferences using the AM/FKM methodology might in fact capture more primitive distributional preferences than observed using a few dictator game decisions. Taken together, we find mixed evidence regarding the relationship between social cohesion and distributional preferences, hinting towards a lower relative weight on own payoff for increased *relationship closeness.*

Result 3. *Lower levels of social distance correspond to a reduced relative weight on own earnings (α), but have no effect on preferences for trading of equity and efficiency (ρ). This result does not hold for IOS scores when decomposing the results for SD = 1.*

4.5 Distributional preferences and altruistic giving

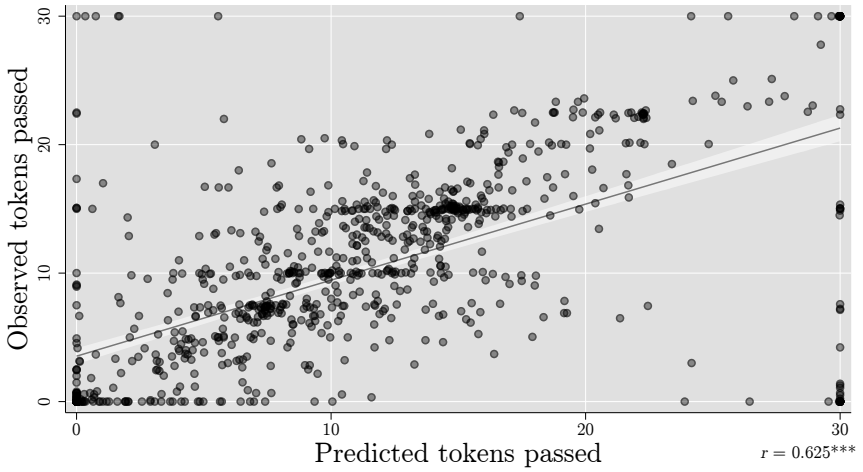
When exploring the relationship between estimated preferences and altruistic giving we can utilise the individually fitted utility functions allows us to make out of sample predictions based on the estimated parameters. Thus, for each level of social distance and anonymity we can construct a CES utility function for each subject. If the estimated utility function is meaningful and relates to altruistic giving in other dictator games, we should be able to predict choices of individuals only based on the estimated parameters and compare these to observed choices. To do so, we consider the following optimisation problem

$$\begin{aligned} \max_{\pi_s,\pi_o} \quad & [\alpha(\pi_s)^\rho + (1-\alpha)(\pi_o)^\rho]^{\frac{1}{\rho}} \\ \text{s.t.} \quad & \pi_s + p\pi_o = m \quad for \quad p \in \{1/3, 1, 3\}, \end{aligned} \tag{1}$$

that allows us to obtain a point prediction of tokens passed (π_o), for each of the three exchange rate levels. Figure 3.11 plots the comparison of predicted and observed choices. On the vertical axis we plot the tokens passed as observed within our study, whereas the horizontal axis plots the predicted tokens according to the fitted utility function. The fitted line is again the result of a univariate regression between the two components. Figure 3.11 shows a strong and significant relationship ($r = 0.625$ p-value < 0.01) between observed and predicted tokens passed. This finding also holds when examining social distance and anonymity levels independently (Appendix C, Figure C.1.1). Altogether Figure 3.11 provides clear support that individually estimated utility functions entail substantial information regarding altruistic giving.

However, from Figure 3.11 alone we cannot determine the exact effect size of α or ρ on altruistic giving. We therefore also provide regression analyses in Table 3.3 including dummy variables for the different exchange rates. In line with our previous result, we find significant predictive power of distributional preferences on altruistic giving, where a change from selfish

Figure 3.11: Observed vs predicted dictator game giving



Note. Predicted tokens on horizontal axis, observed tokens on vertical axis. Each observation is a comparison of an individuals fitted utility function and their out-of-sample transfer.

preferences ($\alpha = 1$) to an equal relative weight on payoffs ($\alpha = 0.5$) corresponds to an average increase of 8.38 tokens passed. This effect is even stronger (12.32 tokens) when examining *non-anonymous* payments separately (see model 3). We therefore, confirm our hypothesis and provide evidence for a considerable reduction in tokens passed with an increased weight on others payoff. Moreover, we also find that an increase in preferences for efficiency (higher ρ) leads to a significant reduction in tokens passed. As we can find no effects of exchange rates on giving and the three dictator games are symmetric in terms of efficiency gains, it is not entirely clear why higher values in ρ should be associated to lower generosity. Moreover, contrasting

Table 3.3: The predictive power of other-regarding preferences

	Tokens passed		
	(1) Pooled	(2) Anonymous	(3) Non-Anonymous
α	-16.76***	-13.16***	-24.63***
	(2.718)	(3.321)	(2.093)
ρ	-5.013***	-4.798***	-5.683***
	(1.069)	(1.559)	(1.189)
Exchange Rate (Reference: 3:1)			
1:1	-0.120	-0.799	0.715
	(0.566)	(0.587)	(0.793)
1:3	0.671	-0.422	1.978
	(1.166)	(1.095)	(1.728)
Controls	Yes	Yes	Yes
Observations	840	458	382
R^2	0.225	0.230	0.283

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Controls are *age, gender, ethnicity, study year, pro-social survey measures, IOS score for stranger, and experimental wave.* Standard errors are clustered at the individual level.

our expectations, we also fail to find evidence of a relationship between ρ and choices in the three different dictator games. Lastly, with respect to demographic controls, we do not find any systematic effects on altruistic giving, except for some evidence that male subjects pass more tokens in the *anonymous* treatment relative to other reported genders.

Both analyses above demonstrate findings that are in line with our expectation that there is a significant relationship between elicited distributional preferences and altruistic giving.

Result 4. *Using our condensed AM/FKM methodology, we find that estimated preference parameters entail substantial predictive information regarding altruistic giving. A lower relative weight on own payoff (lower α) and more equity concerns (lower ρ) are associated to an increase in generosity.*

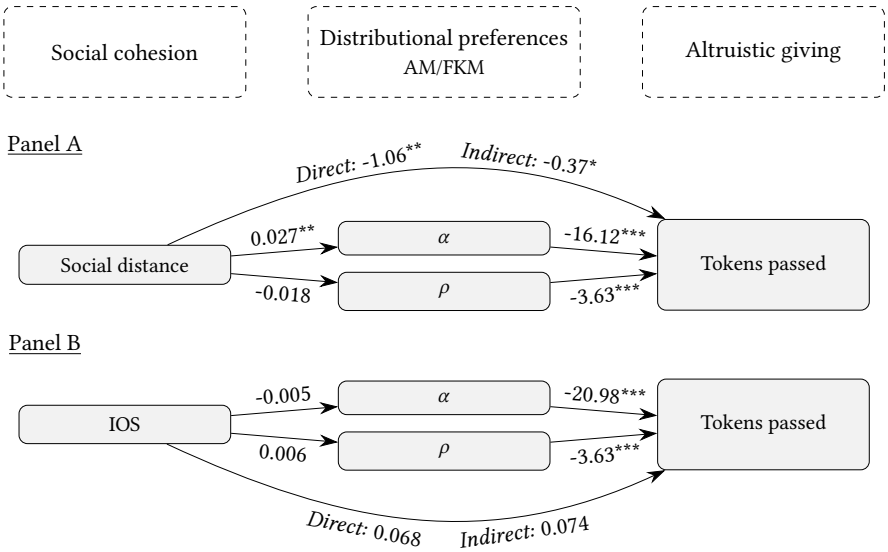
4.6 Mediation analysis

We now combine the analysis of the three previous sections into a single mediation analysis using a structural estimation approach (see MacKinnon et al., 2007; Iacobucci et al., 2007). Before presenting the results, we can consider our previously depicted schematic overview as a set of regressions to illustrate the estimated models

$$\begin{aligned} Tokens\ passed &= \beta_0 + \beta_1 * \alpha + \beta_2 * \rho + \beta_3 * SocialCohesion + \epsilon_1 \\ \alpha &= \gamma_0 + \gamma_1 * SocialCohesion + \epsilon_2 \\ \rho &= \delta_0 + \delta_1 * SocialCohesion + \epsilon_3 \\ for\ SocialCohesion &\in \{Social\ Distance,\ IOS\}. \end{aligned} \tag{2}$$

Social cohesion as proxied by social distance or elicited using the IOS scale is considered a direct predictor of altruistic giving (tokens passed) as well as an ingredient in explaining dis-

Figure 3.12: Overview and results of the mediation analysis



Note. Mediation analysis using a structural model estimation. Panel A shows social distance as a proxy for social cohesion. Panel B captures IOS scores. Direct and indirect effects on arrows from social cohesion to altruistic giving.

tributional preferences (α and ρ), see *Results 2* and *3*. Moreover, both estimated preference parameters also directly affect tokens passed, as found in *Result 4*. As in the previous sections we therefore conduct an independent mediation analysis for each measure of social cohesion (SC). For ease of presentation, the considered relationships and their corresponding effects are best outlined using a diagram in Figure 3.12. Panel A demonstrates the results of our mediation analysis using social distance, whereas Panel B shows the results for IOS scores.

Our structural estimations confirm the results obtained and outlined in the previous sections. In both panels we can observe significant direct effects of distributional preference parameters on altruistic giving. Moreover, as our analysis of the IOS is again restricted to $SD = 1$, we find that α has a stronger effect on altruistic giving for $SD = 1$ than for larger social distances.[21] With respect to social cohesion, we again find significant effects for social distance, but not for IOS levels, with all the signs of coefficients being in line with our hypotheses.[22] Lastly, we observe a difference in magnitude between the significant *direct* and *indirect* effects of social distance on altruistic giving. This difference in magnitude between *direct* and *indirect* effects is not observed for IOS, where, though insignificant, both coefficients are comparable in size.

4.7 Replication of Leider et al. (2009)

As outlined in the introduction our study closely follows the work by Leider et al. (2009) including an exact replication of some of their elements. We thus want to use this section to summarise parts of their study that we succeeded to replicate as well as pointing out aspects that we were unable to. We do confirm their *Result 1* that *baseline altruism* and *directed altruism* are significantly correlated (see Table 3.1). Furthermore, we replicate that social distance significantly affects the degree of *directed altruism*, by observing different generosity levels for distinct social distances. Like Leider et al. (2009), with respect to the effect of a *prospect of future interaction*, we find that altruism increases in *non-anonymous* conditions. However contrasting their results, we cannot replicate a stronger effect for the prospect of future interaction for friends than for nameless partners. If at all, we find the opposite, where the generosity with SD = 2 and *nameless* partners is more affected by lifting anonymity. Moreover, across the board we cannot identify any effect of exchange rates in the dictator game decisions.

All in all, we are able to confirm most of the findings by Leider et al. (2009), except for the impact of exchange rates. One reason for this result might be that we utilised a different interface to present the impact of relative prices. Using a budget set interface as in AM/FKM transparently allowed subjects to choose an allocation and immediately observe how the exchange rates affect final payoffs. In Leider et al. (2009) on the other hand, subjects were presented with all dictator game decisions on the same screen and were asked to pass 50 tokens in total for each choice. Exchange rates were then communicated such that each token was worth

[21] This is also the case when examining social distance and restricting the sample to $SD = 1$.
[22] Note that higher social distance corresponds to lower levels of cohesion, resulting in opposite signs than for IOS scores.

varying points for a subject and their matched partner. Thus, any impact of exchange rates had to be mentally calculated. In addition, having all decisions on the same screen may have resulted in a demand effects or an increased salience of exchange rates that was not present in our study.

5 Conclusion

In this study, we contribute to the understanding of underlying mechanisms that shape other-regarding preferences as well as predict altruistic giving. We thereby focus explicitly on social cohesion as a key determinant in explaining other-regarding behaviour. We hypothesise that the subjective *closeness* individuals perceive to one another directly affects directed altruism both directly and indirectly via distributional preferences. We confirm Leider et al. (2009) by finding significant effects of social distance on altruistic giving and extend their findings by showing that this effect is mediated by changing distributional preferences. In particular, we do find that a higher estimated weight on own earnings significantly explains lower generosity in out-of-sample allocation decisions. Moreover, higher preferences for efficiency are also associated with fewer tokens passed. However, when using the IOS scale as a more nuanced measure of *relationship closeness*, we cannot find systematic evidence for a relationship between social cohesion and altruistic giving, which might be due to a lack of statistical power. Finally, conducting a mediation analysis, we only observe a significant *direct* and *indirect* effect of social distance on altruism, whereas IOS does not relate to preferences or generosity.

In addition to closeness, we also investigate the role of *a prospect of future interaction* (anonymous vs. non-anonymous) to account for reciprocal concerns. Here we find that the prospect of future interaction significantly increases generosity. In addition, by employing a within-subject design we can test for stability in distributional preferences. Subjects that have higher levels of *baseline* altruism towards a *nameless* person also have higher preferences for *directed* altruism towards *named* people. This underlines again the stability of preferences and provides support for the utilised methodology. Altogether our study contributes substantially to our understanding of the interactions between social cohesion, distributional preferences and altruistic giving. Extending the data collection will allow us to further dive into the significance of social cohesion and conclusively infer the impact of of relationship closeness on distributional preferences and altruistic giving.

Throughout Part I of this thesis, we have developed and validated novel experimental methodologies to elicit distributional preferences as well as social cohesion. Both proposed tools improve on shortcomings of the original methodologies whilst maintaining the quality of measurement. We therefore provide researchers with two thoroughly validated, concise and efficient experimental instruments that can be used in future research. Moreover, the last chapter where we explore the interplay between social cohesion, distributional preferences and altruistic giving serves as a proof of concept, underlining the applicability of our tools. As the study involves a combination of different measures, as well as a repetition across waves, time and budget considerations were crucial making it reliant on our developed methodologies. Within chapter 3, we thus present a concrete example of how the proposed tools can

assist researchers to execute their experimental designs. Continuing our pursuit of examining experimental methodologies, Part II now shifts the focus away from social preferences and cohesion and towards the topic of *creativity*.

Bibliography

Andreoni, J. & Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2), 737–753.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144).

Bohnet, I. & Frey, B. S. (1999). Social distance and other-regarding behavior in dictator games: Comment. *American Economic Review*, 89(1), 335–339.

Bourlès, R., Bramoullé, Y., & Perez-Richet, E. (2017). Altruism in networks. *Econometrica*, 85(2), 675–689.

Bourlès, R., Bramoullé, Y., & Perez-Richet, E. (2021). Altruism and risk sharing in networks. *Journal of the European Economic Association*, 19(3), 1488–1521.

Brañas-Garza, P., Cobo-Reyes, R., Espinosa, M. P., Jiménez, N., Kovářík, J., & Ponti, G. (2010). Altruism and social integration. *Games and Economic Behavior*, 69(2), 249–257.

Brañas-Garza, P., Jiménez, N., & Ponti, G. (2017). Eliciting real-life social networks: a guided tour.

Charness, G. & Gneezy, U. (2008). What's in a name? anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, 68(1), 29–35.

Charness, G., Haruvy, E., & Sonsino, D. (2007). Social distance and reciprocity: An internet experiment. *Journal of Economic Behavior & Organization*, 63(1), 88–103.

Charness, G. & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3), 817–869.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy–altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology*, 73(3), 481.

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5), 1858–1876.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3), 347–369.

Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the'inclusion of the other in the self'scale. *PloS one*, 10(6), e0129478.

Gächter, S., Starmer, C., & Tufano, F. (2019). *The surprising capacity of the company you keep: revealing group cohesion as a powerful factor of team production.* Technical report, CeDEx Discussion Paper Series.

Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). Lioness lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1), 95–111.

Goeree, J. K., McConnell, M. A., Mitchell, T., Tromp, T., & Yariv, L. (2010). The 1/d law of giving. *American Economic Journal: Microeconomics*, 2(1), 183–203.

Goyal, S. (2005). Strong and weak links. *Journal of the European Economic Association*, 3(2-3), 608–616.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3), 346–380.

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review*, 86(3), 653–660.

Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology*, 17(2), 139–153.

Krishnan, P. & Sciubba, E. (2009). Links and architecture in village networks. *The Economic Journal*, 119(537), 917–949.

Leider, S., Möbius, M. M., Rosenblat, T., & Do, Q.-A. (2009). Directed altruism and enforced reciprocity in social networks. *The Quarterly Journal of Economics*, 124(4), 1815–1851.

Ligon, E. & Schechter, L. (2012). Motives for sharing in social networks. *Journal of Development Economics*, 99(1), 13–26.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593–614.

Robson, M. (2021). Inequality aversion, self-interest and social connectedness. *Journal of Economic Behavior & Organization*, 183, 744–772.

Part II

Creativity

Introduction

A recent documentary about *The Beatles*[1] depicts a scene of a young Paul McCartney sitting in a chair and composing their hit song '*Get Back*' out of seemingly thin air. This scene has received an incredible amount of attention and viewers were fascinated by seeing this happen. Creative feats such as this often leave observers wondering where they emerged from and how some individuals can accomplish them whereas others do not. There is a natural intrigue to understand creative achievements and to learn what makes people more or less creative. This has led to substantial research amongst psychologists over the last century, ultimately resulting in dedicated journals for the study of creativity.[2] However, it would be wrong to associate creativity only to the arts as above, since it also plays a crucial role in a variety of other contexts. For example, it is undeniable that creativity substantially affects economies through facilitating new innovations (Sarooghi et al., 2015).[3] Due to increased automation and competition, firms must innovate to achieve comparative advantages. More importantly, in light of increasing societal as well as environmental challenges, generating creative solutions to combat these is as important as ever. Therefore, identifying mechanisms that underlie successful innovation and creative ideation is a crucial endeavour not only from an economic but also from a societal perspective.

Despite the relevance of creativity for many facets of our lives, the scientific study of it is non-trivial. When examining the experimental creativity literature it seems that almost every study is accompanied with a different experimental task to measure creativity. This exorbitant number of tasks and methodologies reflects a lack of consistency when it comes to the measurement and ultimately to the conceptualisation of creativity itself. Before studying determinants of creativity, it is crucial to move one step back and first examine approaches to measure creativity. This necessity is also apparent when inspecting the emerging economic literature that experimentally investigates potential mechanisms that underlie creative ability (see Chen et al., 2012; Eckartz et al., 2012; Erat & Gneezy, 2016; Laske & Schröder, 2017; Charness & Grieco, 2019; Bradler et al., 2019; Attanasi et al., 2019b; Charness & Grieco, 2021; Gneezy et al., 2021, among others). Most of these studies focus on financial incentives for creative performance, finding remarkably mixed results. A number of studies find evidence that incentives enhance creative output, whereas some find no effect of financial incentives. Another smaller set of studies even finds detrimental effects of financial incentives on creativity, especially when stakes are high.[4] While incentive schemes are relatively comparable across these studies, they vary with respect to their measures of creative performance, raising the question whether these measures capture fundamentally different abilities or aspects of creativity.

In the next two chapters, I contribute to our understanding of experimental creativity tasks and the literature above by first examining the correlation between existing experimental

[1] 'The Beatles: Get Back' directed by Peter Jackson.
[2] See *Creativity Research Journal*, *Journal of Creativity* and *The Journal of Creative Behavior*.
[3] In their meta-analysis Sarooghi et al. (2015) extensively review the role of creativity in innovation.
[4] See Appendix D, Table D.1.1 for an overview of all studies examining incentives for creativity and their results.

tools to measure creativity, before developing and testing a new methodology. In Chapter 4, we shed light onto the question whether described differences in the effects of incentives for creativity could be caused by differences between tasks employed to measure creativity. To explore this notion we compare the within-subject validity of five distinct creativity tasks. Our results show no significant correlation in performance across tasks, suggesting that they do indeed measure different aspects of creative ability. This finding provides an intuitive possible explanation for the mixed results found in the literature.

As a consequence of this finding, in Chapter 5 we first narrow our attention to a specific domain of creativity: associative thinking. Next, we propose two novel measures of associative thinking that address existing measurement issues and benchmark them against established measures of creativity. In particular, we develop two associative thinking tasks based on an underlying semantic network that provides a pre-defined solution space. We find that our tasks correlate with well-established measure of creativity, while increasing experimental control and providing objective measures of performance.

Once again, each chapter is self-contained and provides all details regarding research questions, hypotheses, designs and findings. Moreover, we conclude part II with a discussion of three possible applications of the developed methodology. The applications demonstrate the variability of our tasks and how creativity research can be combined and integrated with broader economic research questions.

Chapter 4

Investigating experimental measures of creativity

1 Introduction

As argued in the introduction to this part, understanding mechanisms underlying creativity is crucial from an economic perspective, as it is a key component in successful innovations. In response to this, an emergent literature has recently focused on understanding the role of incentives for creativity (see Chen et al., 2012; Eckartz et al., 2012; Erat & Gneezy, 2016; Laske & Schröder, 2017; Charness & Grieco, 2019; Bradler et al., 2019; Attanasi et al., 2019b; Charness & Grieco, 2021; Gneezy et al., 2021, among others). However, when inspecting this literature it becomes apparent that findings of these studies are rather mixed and inconclusive. Using a variety of incentive schemes and creativity tasks, some find positive, some negative and some no effects of financial incentives on creative performance for individuals as well as groups.[1]

As the different studies are generally applying comparable incentive structures (see Attanasi et al., 2021, for an overview), a crucial question is what the reason for the mixed results is. One plausible explanation for the inconsistent findings is their use of distinct creativity tasks that might measure different creative abilities, which in turn respond differently to incentives. More precisely, upon reflection of this literature, it appears that there is still a lack of understanding as to what differentiates and links different creativity tasks. In fact, the only systematic experimental comparison of two different creativity tasks from the studies mentioned above is by Charness & Grieco (2019) comparing *closed* and *open* creativity. They find incentives enhance creativity in *closed* tasks, but have no effect on *open* creativity tasks.

In this chapter, we therefore provide a methodical investigation and comparison of various experimental measures of creative ability in a within-subject design. By improving the understanding of experimental creativity tasks, our study sheds light on the generalisability of previous work on creativity and can highlight potential reasons for the mixed results found. To be more specific, across two studies we explore the role of *domain specificity* and *task openness* using five different experimental creativity tasks. *Domain specificity* describes the notion that distinct domains of creativity, such as artistic or mathematical creativity, should be considered as independent components. Therefore, a task eliciting creativity in one domain might not entail any information on creative ability in another. Following the conceptualisation of Unsworth (2001), *task openness* on the other hand describes how well-defined a task is before

This chapter is based on joint work with Urs Fischbacher, Chris Starmer & Fabio Tufano. I thank Andrew Schotter, Werner Güth, Thomas Dohmen, Daniel Seidmann, all other CeDEx members and the participants of the CCC Conference 2018 for discussions and comments. I am also grateful for funding by the Thurgau Institute of Economics (Kreuzlingen), CeDEx and the Network for Integrated Behavioural Sciences (NIBS). Study 2 of this chapter was conducted as part of my MSc dissertation.

[1] A review of results and the corresponding methods utilised can be found in Attanasi et al. (2021).

a subject begins the process of completing it.[2] While in "*...closed creativity tasks, there is a specific and delineated goal*" (Charness & Grieco, 2019, p.2) open creativity tasks "*...represent unfettered thinking outside the box without any obvious underlying ex-ante goal or direction*" (Charness & Grieco, 2019, p.3). An example for a *closed* creativity task would thus be for instance to program a computer software for a specific purpose, whereas composing a piece of music would fall under the category of an *open* creativity task.

In Study 1, we compare three established creativity tasks that have previously been utilised as measures for *general* creative capability: the *unusual uses task* (Guilford, 1967; Bradler et al., 2019), *a word task* (Eckartz et al., 2012) and *a figural task* (Torrance, 1962, 1966). The three tasks vary in both, their domain of creativity and their degree of openness. If these experimental measures indeed extract *general* creative ability we should find a significant within-subject relationship in performance between them. Contrasting that, our results show no systematic statistical relationship between the three creativity tasks that have been utilised to measure *general* creativity. This suggests that high performance in different creative domains and levels of task openness require different sets of creative skills.

In Study 2, we thus fix the domain of creativity and only vary the degree of task openness. Apart from Charness & Grieco (2019) there has been little research exploring the role of task openness and its potential relevance. This seems rather striking as creative processes are frequently described as fundamentally associative and divergent, thereby always including a degree of *openness*. The two tasks we utilise in Study 2 are the *Tower of London* (Krikorian et al., 1994) and a *Hiking task* (Myszkowski et al., 2015). Both tasks focus on the domain of creative problem-solving but are distinct in their degree of openness. However, even after holding the domain of creativity constant, we find that performance in the two tasks does not systematically correlate. In addition, we also find no relationship between self-reported measures of creativity and performance in creativity tasks across both studies, thereby questioning the appropriateness of either self-reports or experimental creativity tasks as measures for creative ability.

However, before investigating characteristics of creativity tasks, let us first define the concept of creativity. In a review, Runco & Jaeger (2012) find that the *commonly used* definition of creativity dates back to Stein (1953), who defines it as spanning two dimensions: *novelty* and *usefulness*. In light of this definition, there has been a substantial number of experimental measures of creativity proposed in the psychological literature beginning with the *Unusual Uses Test of Divergent Thinking* (Guilford, 1950, 1967) and the *Torrance Test of Creative Thinking* (1962; 1966).[3]

In parallel to the development of creativity tasks emerged a debate concerning the in-

[2] In the creative problem-solving literature problems are commonly defined from the perspective of a subject that is facing a problem for the first time. Thus, before attempting to solve it. See Runco (1994) or Charness & Grieco (2019).

[3] In addition, more recent experimental and clinical methodological developments also include Simonton's histometric approach (Simonton, 1997), case study approaches (Gruber & Wallace, 1999) or the use of neuroscientific tools (see Andreasen, 2005). As the range of creativity tasks greatly exceeds the capacity of this work, please see Hennessey & Amabile (2010) and Runco (2004) for more comprehensive reviews of creativity research.

herent nature of creativity. Some scholars such as Guilford (1950, 1967) or Torrance (1962, 1966, 1974) consider creativity as content general, suggesting that creative ability correlates between different domains, such as verbal, musical or artistic.[4] This however has been recently criticised by a growing literature favouring domain specificity (see Brown et al., 1989; Baer, 1994c,b,a; Cropley, 2000; Kaufman & Baer, 2002, 2004, 2005; Kaufman et al., 2010; Cropley et al., 2011; Kaufman, 2012, among others). In a prominent conceptual framework by Baer & Kaufman (2005), the authors describe a number of domains of creativity, including *creative* problem-solving, which we explicitly focus on in Study 2.

Turning to the emerging creativity literature in economics, most work concerns examining incentives for creativity. With respect to *creative problem-solving*, using various methodologies, many studies find detrimental effects of incentives (Ariely et al., 2009; Chen et al., 2012; Artes et al., 2019; Kleine, 2021), whereas some find no effect (Eckartz et al., 2012; Artes et al., 2019) and others even beneficial impacts of incentives (Ariely et al., 2009; Toubia, 2006; Chen et al., 2012). Similar mixed findings can also be observed when employing tasks that are utilised to measure *general* creativity. Bradler et al. (2019), Artes et al. (2019) and Laske & Schröder (2017) find positive incentive effects using variants of the unusual uses task,[5] whereas Kachelmeier et al. (2008), Kachelmeier & Williamson (2010) and Erat & Gneezy (2016) find positive, negative as well as no effects of incentives when asking subjects to design rebus puzzles. Using verbal, drawing and mathematical creativity tasks Charness & Grieco (2019) and Attanasi et al. (2019a, 2020) both find that incentives do increase creative performance. Moreover, as mentioned above, Charness & Grieco (2019) compare their findings to *open* creativity tasks for which they cannot find an effect of financial incentives.

In addition to individual creative performance, there is also some literature on group creativity (see Chen et al., 2012; Ramm et al., 2013; Attanasi et al., 2019b; Grözinger et al., 2020; Charness & Grieco, 2021; Gneezy et al., 2021). Comparing individuals with groups as well as comparing different incentive schemes, the studies again find inconsistent results. Using the same drawing, verbal and mathematical tasks as in previous studies, Charness & Grieco (2021) find cooperative group incentives to enhance creativity, whereas competitive incentives have no effect. This is contrasted by Attanasi et al. (2019b), who utilise the same tasks, but find the opposite to be true. While competitive group incentives stimulate creative output, cooperative incentives do not. Examining a non-routine task, Englmaier et al. (2021) also find that incentives improve group performance for solving escape room games. Comparing individual and group creativity, Ramm et al. (2013) find higher performance in problem-solving by groups, whereas Gneezy et al. (2021) cannot find a difference between individuals and groups when asking subjects to create titles for videos.

[4] It is however important to note that researchers who support content generality do not refer to a perfect correlation between creative abilities across different domains. Their perspective has to be interpreted in the sense that there is an aspect of creativity that is inherent to every individual and affects creative performance across all activities (Plucker, 1998). This is similar to the general intelligence factor (g) in intelligence research (Spearman, 1904; Deary, 2000).

[5] The task by Laske & Schröder (2017) is not the traditional unusual uses task but rather a variant of it. Using a set of provided objects subjects have to generate many different and original words.

A table summarising all experimental studies described above, including creativity tasks, treatments and results can be found in Appendix D, Table D.1.1. The mixed results outlined in the previous paragraphs, make it almost impossible to conclude a systematic effect of incentives on creative performance. This is precisely where our studies connect to the existing research. Contrasting previous work, we do not focus on incentive effects for creative performance but systematically shed light on several different creativity tasks and examine whether they elicit the same creative ability.

The remainder of this chapter is structured as follows. In Section 2, we present the design, procedure and results of Study 1. In Section 3, we outline the same aspects for Study 2. In Section 4, we present evidence on the predictive power of survey measures. In Section 5, we address potential concerns of measurement error before we provide concluding remarks in Section 6.

2 Study 1 - Comparing *general* creativity tasks

2.1 Design

The goal of Study 1 is to investigate different creativity tasks commonly employed as measures of *general* creativity. If these tasks are indeed measures of *general* creativity, we should observe better performances by more creative individuals across tasks, independent of differences in domains or openness. The tasks we employ are a *word task*, the *unusual uses task* and a *figural task*, which are all borrowed from previous research. The reason we decided to focus on these three tasks is multifaceted. First, the *word task* entails a simple implementation and would therefore be a strong candidate for a recommended measure of creativity. The unusual uses task on the other hand is one of the most widely employed creativity tasks and thus provides a meaningful benchmark for the other two tasks. Lastly, the figural task moves away from the language based domain of the other tasks and approaches creativity in an entirely different, artistic, way. Below we outline the details of each of the tasks, describe their implementation and performance measures.

Word Task

The first task we employ is the word task adopted from Eckartz et al. (2012). In this task subjects are presented with a 12 letter-set, e.g. *accehhikllst*. They are then asked to construct as many different words as possible within 5 minutes using the letters in the set.[6] Performance is measured by the length of the constructed words in a convex manner where a one-letter word yields *1* point, a two-letter word gives *1 + 2 = 3* points and so forth. Since the solution space is ex-ante defined, the task is simple to implement and scoring can be directly programmed into the implementation.

[6] As the study was conducted in Konstanz, Germany, we asked them to construct German words. All subjects had excellent command of the language.

Guilford's Unusual Uses Task

Due to its vast utilisation, we use Guilford's *unusual uses task* (UUT) as our second assessment of general creativity. Here subjects have to list as many unusual uses as possible for a provided object within 6 minutes. We follow Bradler et al. (2019) and select either a *sheet of paper*, a *tin can* or a *piece of cord* as the object faced by subjects.[7] Following the literature we use three metrics, *fluency*, *flexibility* and *originality* as performance measures. *Fluency* thereby captures the number of valid uses submitted,[8] while *flexibility* counts the number of distinct categories. *Originality* is a statistical measure of how scarce a use is relative to other submissions. More precisely, each subject receives 1 point for a use named by no other, ½ a point if named by one other and 0 otherwise (see Bradler et al., 2019).

To implement the performance measures, at the beginning of the experiment we assign all subjects either to the role of participant or judge. Participants complete the task as described above, whereas judges classify uses as valid and categorise them. In our implementation, each session consists of 5 judges and 12 participants to allow for a live categorisation of all submitted uses.

Figural Task

As our last task, we employ a figural task that is part of the Torrance Test of Creative Thinking (TTCT Torrance, 1966, 1962). In this task subjects have 5 minutes to construct figures that consist of circles. Subjects are free to create as many figures as they like and have full liberty with regards to the look of the figures and the size of the circles. We then present judges (see section above) with 40 randomly drawn pairs of figure-sets and ask them to select the one they consider more creative.[9] Similar to Charness & Grieco (2019), we ask judges to use their own definition of creativity and ensured that each pairwise comparison was evaluated by at least two judges, allowing us to examine the reliability of ratings.[10] In the end, we use the fraction of won comparisons as our measure of creative performance.

In addition to the three experimental tasks we elicit self-reported measures for creativity in various domains using the Kaufman Domains of Creativity scale (K-DOCS Kaufman, 2012).

Study 1 was conducted in four sessions at the University of Konstanz in May and June, 2017. We recruited subjects using ORSEE (Greiner, 2015) and computerised the experiment with z-Tree (Fischbacher, 2007).[11] To incentivise performance we employed simple pairwise tournament incentives across all tasks.[12] Before each task two subjects were randomly matched

[7] To avoid contamination, we varied the object between sessions and do not find performance differences between sessions.
[8] Valid here means that the use provided is at least vaguely conceivable. Thus, we were rather loose on our restriction as to what was considered valid.
[9] A figure-set is the collection of all figures created by a subject. All of the created figure-sets can be found in Appendix D.3.3.
[10] We find large and significant correlations between the likelihood of winning a comparison across judges.
[11] Instructions can be found in Appendix D.2.1.
[12] Note, as our focus here is not on incentive effects, but on the tasks themselves we do not vary incentive schemes.

and then competed against one another. Performances of both players were compared and the winner received a fixed amount of €8 and the loser nothing.[13] In case of a draw both subjects received €4. In total we recruited 76 subjects across four sessions, 20 of whom acted as judges, categorising and evaluating performances. Each session lasted around 80 minutes and subjects earned on average €18.

2.2 Results

Before addressing the main result for Study 1 we provide an overview of key variables in Table 4.1. Recall that scores in the UUT partly depend on the categorisation by other subjects, which in some cases made revisions necessary to ensure consistency across judges.[14] Similarly, for the figural task, we hired five independent research assistants to re-evaluate the created figures. Both procedures, the re-categorisation as well as the re-evaluation did not significantly differ from the scoring within sessions.[15] For all statistical analyses we use the revised scores that are also shown in Table 4.1.[16] When examining Table 4.1, we can see that there is substan-

Table 4.1: Descriptives - Study 1

	Mean	SD	Min	Max	N
Word task					
Total points	224.64	79.46	0	419	56
Unusual uses task					
Fluency	17.75	7.28	7	40	48
Flexibility	8.02	2.25	3	12	48
Originality	6.00	3.97	0.50	18.00	48
Figural task					
Fraction of wins	0.51	0.21	0.07	0.95	56
Demographics					
Female	0.59	0.50	0	1	56
Age	21.70	2.87	17	30	56

Note. Descriptives for all three creativity tasks. Total points in the *word* task refers to the sum of all points across the correctly identified words. *Fluency* is the number of uses, *flexibility* the distinct categories and *originality* captures statistical rarity. The *figural task* was evaluated by pairwise comparisons and the *fraction of wins* indicates how often a subject won a comparison.

[13] Recall that, in the word task performance is simply the aggregation of all achieved points. In the unusual uses task subjects receive 1 point per use, 1 point per category and (2) 1 point(s) for (very) rare answers. (Very) Rare here means that (no) one other subject in the same session listed the same use. Lastly, in the figural task the fraction of won comparisons is used as a measure of performance.

[14] For the unusual uses task this step appeared to be appropriate as judges in some cases incorrectly specified categories. Reasons for that are twofold: On the one hand it is not always possible to perfectly categorise a use, which led to instances where the same use was allocated differently. On the other hand we intentionally avoided performance incentives for the judges, which allowed judges to wrongly classify uses without encountering financial damage. Revisions were executed by the same researcher in order to ensure consistent allocation to categories throughout all sessions.

[15] The total points before and after re-categorisation are significantly correlated with $r = .90$ (p-value < 0.01). We find similarly strong correlations for the re-evaluations in the figural task with $r = .74$ (p-value < 0.01).

[16] Note that differences in the numbers of observations are due to 8 additional subjects assigned to the role of examiners for the UUT. Contrasting judges, examiners only pre select whether an answer was considered valid or not. This provided an additional layer of evaluations, ensuring the judges only have to classify valid uses.

Table 4.2: Correlations across performance metrics in the UUT

	Fluency	Flexibility	Originality
Fluency	1		
Flexibility	0.727***	1	
Originality	0.882***	0.614***	1

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Within-subject Pearson's correlation coefficients between *fluency*, *flexibility* and *originality* in the unusual uses task (UUT)

tial heterogeneity across tasks and performance metrics. This already gives a first indication that between individuals, different abilities are necessary to perform well in the three tasks.

Since the UUT entails several performance measures (*fluency, flexbility* and *originality*) we can also gain some more insights into the relationship between the different performance metrics. In Table 4.2, we correlate these three metrics with each other and find strong and statistically significant correlations ($r > 0.61$; p-value < 0.01) between them. This suggests that the use of additional criteria does not necessarily provide any more information with respect to creative ability than using a single metric, which is further supported in our subsequent analyses as we throughout find identical results for all three metrics.

Moreover, as subjects had 6 minutes to submit uses in the UUT, we are also able to explore creative output over time. In particular, we can examine how long the task has to be employed to provide stable heterogeneity in performances. Previous evidence on this question has been mixed with some older studies recommending a longer duration of the task (Olczak & Kaplan, 1969; Ward, 1969; Moran III et al., 1983) and others finding stable heterogeneity already after 1 minute (Benedek et al., 2013). To explore this dimension, we correlate the rank an individual has in *fluency* after each minute of the task in Table 4.3. We decided to focus on *fluency* for simplicity but the observed results also hold for *flexibility* and *originality*. We find strong correlations ($\rho > 0.77$; p-value < 0.01) already from minute 2 onwards. This implies, that subjects, who performed better after 2 minutes also had a higher rank at the end of the task. Our result thus supports studies that argue that a relatively short duration of the task is sufficient to identify robust performance differences (Silvia et al., 2008; Benedek et al., 2013).

Table 4.3: Correlation across task duration

	Unusual uses task					
	1 minute	2 minutes	3 minutes	4 minutes	5 minutes	6 minutes
1 minute	1					
2 minutes	0.685***	1				
3 minutes	0.618***	0.904***	1			
4 minutes	0.517***	0.818***	0.918***	1		
5 minutes	0.453***	0.776***	0.878***	0.974***	1	
6 minutes	0.464***	0.775***	0.859***	0.963***	0.979***	1

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Spearman's rank correlation in the UUT for performance in *fluency* across task duration. Each of the times indicate the rank of *fluency* per subject obtained at this point of the task.

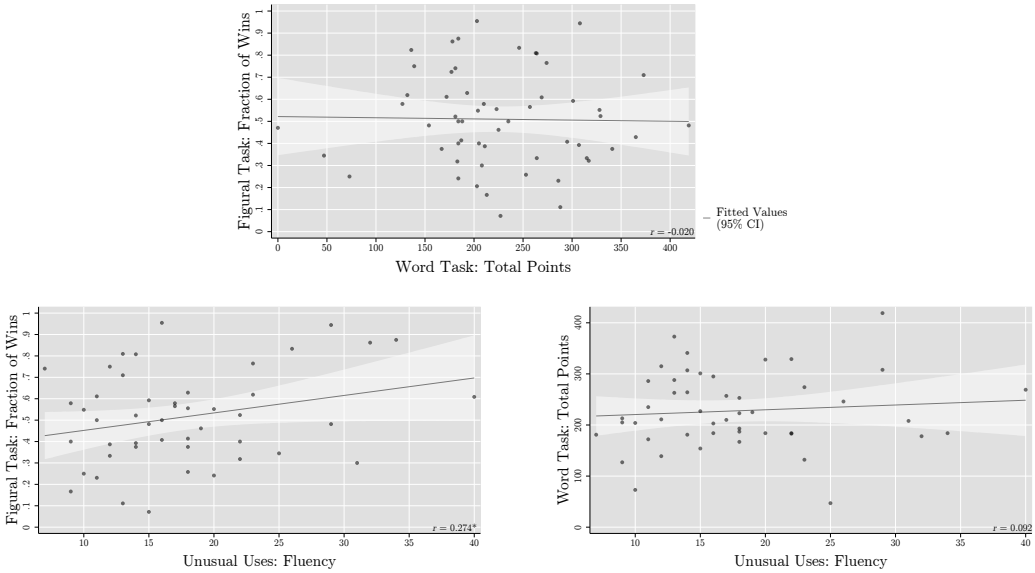
The relationship between *general* creativity tasks

In Study 1, we measure performance in three different tasks aimed to capture *general* creativity. We hypothesise that if the different measurements extract the same creative ability we find a significant relationship in performances between tasks. To explore this hypothesis, Figure 4.1 plots all pairwise relationships between the three tasks.[17] The top graph compares the *word* and *figural* task, the bottom left graph shows the *figural* task and the UUT, and lastly the bottom right graph depicts the relationship between the UUT and the *word* task. Moreover, each graph provides a fitted line based on a simple univariate regression model, as well as Pearson's *r* correlation in the bottom right corner. When examining the three graphs, we can see that there seems to be little to no relationship between performances in the different tasks. This visual impression is also statistically confirmed in Table 4.4 where we provide correlations between and within the three tasks. The only between-task relationship that is marginally significant is between the UUT and the *figural* task for *fluency* and *originality*. However, once controlling for demographics in a regression framework this relationship disappears and becomes insignificant.[18]

Result 1. *We cannot find robust evidence for within-subject correlations in creative performance across the 'word task', 'the unusual uses task' and the 'figural task'.*

Result 1 clearly shows that either the concept of creativity is multifaceted or that it is

Figure 4.1: Relationship between creativity tasks of Study 1



Note. Scatter plots for all pairwise comparisons of the creativity tasks employed in Study 1. *Figural* and *word* task on top, *UUT* and *figural* task on bottom left, *UUT* and *word* task on bottom right. The fitted line corresponds to a simple univariate regression. Pearson's correlation coefficients are provided in the bottom right corners of each graph.

[17] Again using *fluency* as the performance measure for the UUT.
[18] Regression models for all pairwise relationships can be found in Appendix D, Table D.1.2).

Table 4.4: Correlation across creative tasks

	Word task	*Unusual uses task*			*Figural task*
	Total points	Fluency	Flexibility	Originality	Fraction of wins
Word task	1				
Unusual uses					
Fluency	0.0923	1			
Flexibility	0.191	0.727***	1		
Originality	-0.0178	0.882***	0.614***	1	
Figural task	-0.0198	0.274*	0.160	0.261*	1

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Pearson's correlations between all creativity tasks in Study 1. The *word* task is scored by total points achieved, the *figural* task by the fraction of won comparisons. The UUT is presented with all three distinct performance metrics.

a conceptual flaw to consider the three examined tasks to measure a common dimension of creativity. Either way, it is apparent that the tasks employed cannot simply be considered measures of *general* creativity and have to be examined in a more nuanced way. In fact, it is particularly remarkable that we fail to find a relationship between the UUT and the *word* task as both of these tasks could be considered to address the verbal domain of creativity. However, while this is certainly true for the *word* task, the categorisation is not as clear for the UUT. Even though subjects verbally submit answers to the task, the fundamental challenge is much more practical than verbal, by thinking of unusual uses. Moreover, when considering *task openness* we can also see differences between the two tasks. Using the definition of *task openness* by Charness & Grieco (2019), neither task is purely open, but the *word* task is considerably more *closed* than the UUT. Thus, in Study 1, two dimensions, namely domain and openness, varied across the three examined creativity tasks. This categorisation and the lack of relationship between performances in the tasks, ultimately motivated us to conduct Study 2, where we explicitly focus on the domain of *creative problem-solving* whilst varying *task openness*.

3 Study 2 - Assessing creative problem-solving

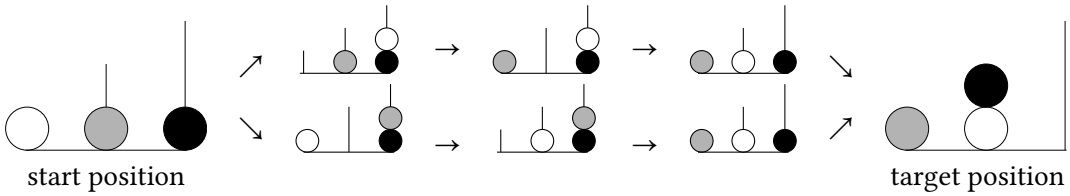
3.1 Design

Like Study 1, Study 2 is also based on a within-subject comparison of creativity measures. However, we now explicitly examine the relationship between open and closed creative problem-solving, which is frequently considered to play a significant role in innovation and is therefore particularly relevant from an economic stand point. We therefore employ two distinct creative problem-solving tasks, which we consider as *closed* and *open* tasks, the Tower of London and a *Hiking task*. Neither of these tasks classify as entirely *open* according to Charness & Grieco (2019), however they substantially differ in their degree of *openess* nonetheless as we will demonstrate below.

Closed Problem-Solving: Tower of London

As our measure of *closed* problem-solving we use the *Tower of London* (ToL) (Krikorian et al., 1994; Debelak et al., 2016). In this task, subjects have to move three objects to three separate rods in order to match a target position, where each column can hold a different number of objects. It is only possible to move one object per move and the time limit is 60 seconds per problem. We instruct subjects to find the shortest possible path between start and target position, which also serves as our performance metric for the task. Figure 4.2 provides an example of the set-up.

Figure 4.2: Example of the Tower of London



All subjects were asked to solve eight distinct ToL problems which vary in complexity, namely the number of minimum moves and the number of possible solution paths when using minimum moves.[19] Lastly, as a robustness check we also include two isomorphic problems that are identical to two others except for the colour coding. A detailed description of all problems can be found in Appendix D.3.4. We consider this task a perfect example for a *closed* creativity task as the goal is ex-ante specified and all possible solution paths are pre-defined. However, despite the lack of ambiguity, subjects still have to identify which path leads them to the target with the fewest moves.

As we found substantial individual level variation in task performances in Study 1, it seems intriguing to further examine whether subjects are able to assess their own creative performance. If this were to be the case, then some tasks could be substituted with carefully crafted survey measures. This question is particularly interesting if we think of the self-assessment as conceptual, defined by domain and openness. In this case it renders the identification of a specific creativity task even redundant. For that matter, in Study 2, we also include a question that specifically addresses *closed* problem-solving. Its exact phrasing is:

> "Consider a problem that has a perfectly specified goal and there are multiple ways to solve it, such as:
> *Come up with a way to protect a raw egg that prevents it from cracking at any height of a fall.*
> How good are you in solving these kind of problems?"

[19] With solution path we mean the order in which the three balls are moved. In the example presented there are two different solutions paths in order to solve the problem with minimum moves.

Open Problem-Solving: Hiking task

To explore *open* problem-solving on the other hand, we employ a task that has been widely used in previous research (see e.g. Myszkowski et al., 2015; Plucker et al., 2014; Carson & Runco, 1999; Chand & Runco, 1993; Okuda et al., 1991). Subjects are confronted with a real-world *hiking* problem and have to come up with as many diverse solutions to the problem as possible within 5 minutes.[20] In some sense, this task is a variant of the previously described UUT now applied to the context of problem-solving. As there is no ex-ante specified solution of the problem, we consider the task as relatively *open*. The problem we utilise in the experiment is:

> *"It's a great day for hiking, and your friend, Jamie, comes to your work and asks you if you want to go hiking. Unfortunately, you have a big project due tomorrow, and it requires a full day to complete. You would rather be hiking. What are you going to do?*
> (Think of as many diverse ideas as you can.)"

As for the UUT we measure performance in this task by *fluency*, *flexibility* and *originality*. *Fluency* is again the total number of answers provided. Similar to the UUT, we categorise all answers and count the number of categories as our measure of *flexibility*. Lastly, for our measure of *originality* we ask all subjects to rate the *originality* of answers from subjects in a previous session on a 7-item Likert scale. Moreover, as for *closed* problem-solving we also additionally craft a survey measure for *open* problem-solving:

> "Consider a problem that has no perfectly specified goal and there are multiple ways to solve it, such as:
> *Come up with a household item that does not yet exist but everyone needs.*
> How good are you in solving these kind of problems?"

In addition to the experimental measures outlined above, we also collect survey data on general creativity, general problem-solving ability and include the 'self-efficacy in problem-solving' questionnaire (Tierney & Farmer, 2002; Beghetto, 2006, 2009; Nazzal, 2015) as well as collect basic demographics.

The study was approved by the Nottingham School of Economics' Research Ethics Committee and conducted online in July 2018 using Prolific as a recruitment platform.[21] We programmed the tasks in LIONESS Lab (Giamattei et al., 2020) and subjects took around 20 minutes and earned a flat payment of £2.50.[22] In total, we recruited 123 subjects across 6 sessions.

[20] We think that since the problem describes a real-world scenario, it is plausible that all subjects are similarly capable to complete this task successfully.
[21] See www.prolific.co for more information on their services.
[22] We decided to pay a flat fee in order to ensure that our experiment does not suffer from unintended incentive effects. Since we were interested in the structure of creativity we did not need to include incentives for meaningful results.

To control for ordering effects we alternated whether the survey questions had to be answered before or after the experimental tasks.[23]

3.2 Results

As for Study 1 we first present an overview of performances in Table 4.5. This time the descriptives are grouped according to *closed* and *open* problem-solving where each group contains information on performances in the respective task as well as responses to the survey measure. Moreover, for each of the problem-solving tasks, Table 4.5 presents multiple performance measures. For the Tower of London these are our main performance metric, the fraction of successful tries with minimum moves, as well as average completion times. With respect to the *hiking* task, we present an overview of *fluency, flexibility* and *originality*. However, here *originality* is a subjective evaluation by other subjects rather than a statistical measure as for the UUT. Nonetheless, similarly to before, we can examine different performance metrics by correlating them with each other. Table 4.6 confirms our findings from above by showing strong and significant correlation coefficients between all 'objective' metrics in both tasks. The negative correlation in the Tower of London is also intuitive, as a better completion time is shorter, whereas a better success rate with minimum moves is higher.

The only weak relationship we can observe between metrics is the 'subjective' *originality* measure in the *Hiking task* correlated with its two 'objective' metrics. Even though we do find marginally significant results, it appears that 'objective' measures such as *fluency* and *flexibility* only weakly relate to 'subjective' evaluations of originality. For all subsequent analysis we are therefore examining *originality* separately when investigating *open* problem-solving.

Table 4.5: Descriptives - Study 2

	Mean	SD	Min	Max	N
Closed problem-solving					
Success (min moves)	0.58	0.22	0	1	123
Completion time	27.87	8.91	13	56	123
Self-report	4.59	1.54	1	7	123
Open problem-solving					
Fluency	6.98	3.29	0	16	123
Flexibility	5.57	2.19	1	14	118
Originality	3.20	0.49	2.00	4.58	118
Self-report	3.63	1.70	1	7	123
Demographics					
Female	0.59	0.49	0	1	123
Age	30.32	9.74	17	64	123

Note. Descriptives for both problem-solving tasks in Study 2. The first two rows for *closed* problem-solving correspond to the Tower of London (ToL). The first is the fraction of successful completions using minimum moves. The second is completion time, where faster is considered better. For the *hiking* task we measure *fluency* as the number of solutions, *flexibility* the covered categories and *originality* is a subjective score by other subjects. The self-reports correspond to the crafted survey measures outlined in the design section.

[23] Instructions can be found in Appendix D.2.2.

Table 4.6: Correlations across performance metrics

Tower of London			
	Success (min moves)	Completion time	
Success (min moves)	1		
Completion time	-0.649***	1	
Hiking task			
	Fluency	Flexibility	Originality
Fluency	1		
Flexibility	0.870***	1	
Originality	0.172*	0.224**	1

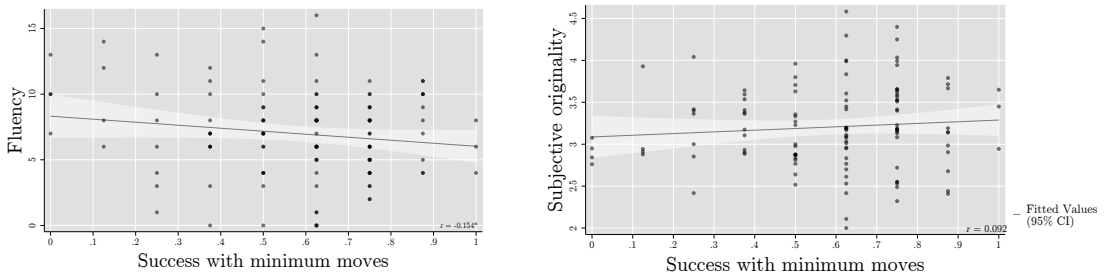
* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Within-subject Pearson's correlations across performance metrics in Study 2. For the Tower of London (ToL), the two metrics are, fraction of successful completions with minimum moves, and completion time in seconds. For the *hiking* task it is the number of solutions (*fluency*), categories covered (*flexibility)*) and a 'subjective' creativity evaluation (*originality*).

Creative problem-solving across levels of *openness*

To finally explore the role of *task openness* whilst holding the domain of creativity constant, we compare performances in the two problem-solving tasks (Figure 4.3). If *domain specificity* holds and task openness does not affect creative performance, we expect to find significant relationships between both tasks. As for Study 1, Figure 4.3 plots the relationship between the ToL (horizontal axis) and the *hiking* task (vertical axis). In the left graph, we use *fluency* as our measure for *open* problem-solving and *originality* in the right. Across graphs we use the percentage of successful completions using minimum moves to measure performance in the ToL. Both graphs also include correlation coefficients in the bottom right corners. By looking at either graph, there is no evidence for a significant relationship between performances. In fact, we even find a marginally significant negative relationship between *fluency* and performance in the ToL (left). As simple correlations do suffer from a lack of controls, we perform additional regression analyses of performance in the ToL on *fluency*, *originality* and demographic controls (see Table 4.7). Our results confirm that there is no significant relationship

Figure 4.3: Relationship of *closed* and *open* problem-solving



Note. Scatter plots for pairwise comparisons of the problem-solving tasks employed in Study 2. Performance in the ToL is on the horizontal and performance in the *hiking* task on the vertical. The left graph plots the 'objective' measure of *fluency*. The right graph captures the 'subjective' metric of originality. The fitted line corresponds to a simple univariate regression. Pearson's correlation coefficients are provided in the bottom right corners of each graph.

Table 4.7: Regression model for *closed* and *open* problem-solving

	Tower of London: Success (minimum moves)			
	(1)	(2)	(3)	(4)
Open PS: Fluency	-0.0106	-0.0132		
	(0.00859)	(0.00889)		
Open PS: Originality			0.0436	0.0319
			(0.0483)	(0.0541)
Questionnaire before	-0.0514	-0.0796	-0.0414	-0.0996
	(0.0790)	(0.0833)	(0.0912)	(0.0935)
Controls	Yes	Yes	Yes	Yes
Self-Reports	No	Yes	No	Yes
Observations	121	121	116	116
R^2	0.343	0.430	0.330	0.411

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Controls are *ethnicity, native language, employment status, student status, education, gender, age*, and a survey measure of *risk*. In addition we also include session fixed effects and the order of the tasks.

across either performance metric in the two tasks.

Result 2. *When holding the domain of creativity constant and varying the level of task openness, we cannot find a significant relationship in individual creative problem-solving ability across the 'Tower of London' and the 'Hiking task'.*

Our results also provide an intuition for the finding by Charness & Grieco (2019), who show that the degree of openness matters for incentive effects. If different creative abilities are required across distinct degrees of openness, incentives are likely to affect them in different ways.

4 The predictive power of self-reported creativity

In addition to our experimental tasks, both studies include self-reported measures of creativity. In fact, self-reports are the most commonly used measure for creativity in psychological studies. It is, however, unclear whether they relate to performance in experimental tasks. To explore this notion across Study 1 and 2, we thus correlate our survey measures of creativity and problem-solving ability with performances in the tasks of both studies. For Study 1, we thereby use a creativity questionnaire that can be decomposed into five separate domains, the 'Domains of Creativity' scale (K-DOCS Kaufman, 2012), while in Study 2 we include four short questions constructed ad-hoc[24] and in addition we also elicit the 'self-efficacy in problem-solving' scale (Tierney & Farmer, 2002; Beghetto, 2006, 2009; Nazzal, 2015). Table 4.8 presents correlations between all survey measures (rows) and corresponding creativity tasks (columns) for both studies. We cannot find a systematic relationship between any of the survey measures

[24]The questions are designed to capture general creativity, general problem-solving ability and *closed* as well as *open* problem-solving.

Table 4.8: Relationship between creative tasks and self-reports

Survey measure	**Creative tasks**				
Study 1					
		Unusual Uses Task			
	Word task	Fluency	Flexibility	Originality	*Figural task*
Domains of Creativity	0.119	0.079	0.018	0.147	0.227*
Self/everyday	0.038	0.023	-0.073	0.076	0.185
Scholarly	-0.131	-0.043	-0.075	0.009	0.215
Performance	0.209	0.184	-0.008	0.206	0.144
Mechanical/scientific	0.073	0.070	0.157	0.149	0.150
Artistic	0.152	-0.002	-0.007	0.004	0.047
Study 2					
			Open problem-solving		
	Tower of London		Fluency	Flexibility	Originality
General creativity	-0.055		-0.076	-0.132	0.056
General problem-solving	0.095		0.036	-0.013	0.049
Closed problem-solving	0.043		0.062	0.014	0.069
Open problem-solving	-0.003		0.166*	0.020	0.076
Self-efficacy in PS	0.037		0.046	-0.034	0.066

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Pearson's correlations between all survey measures and creativity tasks in Study 1 (top) and Study 2 (bottom). The survey measures are in the rows, the creativity tasks in the columns. 'Domains of Creativity' is an aggregate score across all domains of the K-DOC scale. PS = problem-solving.

and a respective creativity task. While the overall picture does not depict a relationship between self reports and tasks, there are two exceptions. In Study 1, we find a weakly significant correlation between the aggregate questionnaire score and the figural task, however once we decompose the K-DOCS, the significance disappears. Especially the domain *artistic* shows a very small correlation, which is surprising as it is an intuitive candidate for relation with the figural task. Overall, it seems that only the domain of *performance* provides somewhat robust and moderate (although insignificant) correlations across the three tasks, which might be worth exploring in subsequent research. In Study 2, we find a marginally significant correlation between *fluency* and our *open* problem-solving survey measure, however the coefficient is relatively small.

All in all, we do not find systematic predictive power of self-reports to explain performances in creativity tasks within either study. This is irrespective of whether we use standardised questionnaires or craft survey measures precisely fitting our selected creativity task. Moreover, the lack of relationship is independent of creative domain or the degree of openness.

Result 3. *Across five tasks and six self-reported measures of creativity, we find that survey measures of creativity do not systematically correlate with performance in experimental creativity tasks.*

While our findings imply that caution is needed when comparing studies that use self-reports with studies employing experimental measures, the lack of correlation between task

performance and survey measures should not be regarded as evidence for the unreliability of creativity questionnaires. Since we cannot argue that experimental tasks are better measurements of creative ability, we can only conclude that one cannot simply substitute a creativity task with a survey measure.

5 Addressing measurement error

One main concern when experimentally investigating creativity tasks is measurement error. Thus, the lack of correlation found above might not be due to different abilities measured by the tasks, but rather a consequence of noise in the elicitation. We can address this concern in both studies in different ways.

First, assuming that noise might be a result of subjective scoring, we can explore an alternative measure for the *figural task*. To do so, we first compute a number of objective measures (number of figures, the total number of circles, the average number of circles per figure, the fraction of moved circles and a measure how much the figures cover the screen) and use these as an instrument for the subjective performance score from our study.[25] When exploring the correlation between the three creative tasks in Study 1, the instrumented scoring, however, does not affect any of our previous findings (Appendix D, Table D.1.3).

We can also address potential noise in Study 2. As we included two isomorphic problems among the eight tasks in the Tower of London, we measure performance in these problems twice for each subject.[26] Thus having repeated measures, we can now utilise the *Obviously Related Instrumental Variable* (ORIV) approach by Gillen et al. (2019) to directly account for measurement error. When having two measures that are obviously related, the ORIV approach leads to a reduction in measurement error by instrumenting each of the measures with the other.[27] After applying the ORIV approach, we can investigate the relationship between the *noise-reduced* isomorphic problems and performance in the *hiking* task (Appendix D, Table D.1.4). We find that using the ORIV approach substantially increases the size of our coefficients but also the associated standard errors, confirming our original result that there is no statistical relationship between performance in both problem-solving tasks. Overall, we do find evidence for the existence of measurement error, accounting for it, however, does not alter our previously obtained findings.

6 Conclusion

Using a within-subject design in two studies, we find no consistent relationship in performance across five different experimental measures of creativity. In a first study, we find that three distinct tasks eliciting *general* creativity do in fact require different abilities. In a follow-up study,

[25] The first stage of the instrumented regression can be found in Appendix D, Table D.1.3.

[26] Isomorphic in this case means identical problems in terms of start and target positions. Problems only differ in their color coding.

[27] Please see the paper by Gillen et al. (2019) for a detailed description, econometric derivation and application examples of the method. One fundamental assumption of the ORIV approach is that the measurement error between measures is independent. This however can be challenged as in our case the Tower of London problems were executed in a rather short time frame and therefore measurement error might not be independent.

when holding the domain of creativity constant but varying the degree of *openness*, we still fail to observe any relationship in creative performance. This suggests that the dimension of task *openness* is crucial when exploring creative ability within an experimental environment. Furthermore, we cannot observe any predictive power of survey measures with respect to performance in our creativity tasks across both studies. This suggests that one cannot simply substitute a task with a self-reported measure of creativity.

Our findings should by no means imply that the utilised tasks are not useful tools to elicit creative abilities. In fact, there seems to be a broad consensus in the literature that the tasks do in fact elicit creative ability. However, despite this consensus our results suggests that within creativity research concepts are still ill defined. Whilst many scholars refer to general creativity and claim to measure it, so far, it looks like the measures are capturing performance in disparate tasks. Overall, the complexity and inconsistencies in our results clearly show that using experimental creativity tasks may be challenging. However, this should not discourage further investigation of creativity and its determinants but rather nudge academics to a more systematic and cautious application of experimental creativity tasks.

Bibliography

Andreasen, N. C. (2005). *The creating brain: The neuroscience of genius.* New York, NY: Dana Press.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76(2), 451–469.

Artes, J., Graves, J., Motika, M., et al. (2019). *Creativity under pressure: Performance payments, task type and productivity.* Technical report.

Attanasi, G., Chessa, M., Ciucani, C., Gallen, S. G., et al. (2020). *Children's grI-creativity: Effects of limited resources in creative drawing.* Technical report.

Attanasi, G., Chessa, M., & Llerena, P. (2021). A survey on experimental elicitation of creativity in economics.

Attanasi, G., Curci, Y., Llerena, P., Pinate, A. C., Ramos-Sosa, M. d. P., & Urso, G. (2019a). *Looking at creativity from east to west: risk taking and intrinsic motivation in socially and culturally diverse countries.* Technical report.

Attanasi, G., Curci, Y., Llerena, P., Urso, G., et al. (2019b). *Intrinsic vs. extrinsic motivators on creative collaboration: The effect of sharing rewards.* Technical report.

Baer, J. (1994a). Generality of creativity across performance domains: A replication. *Perceptual and Motor Skills*, 79(3), 1217–1218.

Baer, J. (1994b). Performance assessments of creativity: Do they have long-term stability? *Roeper Review*, 17(1), 7–11.

Baer, J. (1994c). Why you shouldn't trust creativity tests. *Educational Leadership*, 51(4), 80–83.

Baer, J. & Kaufman, J. C. (2005). Bridging generality and specificity: The amusement park theoretical (APT) model of creativity. *Roeper Review*, 27(3), 158–163.

Beghetto, R. A. (2006). Creative self-efficacy: Correlates in middle and secondary students. *Creativity Research Journal*, 18(4), 447–457.

Beghetto, R. A. (2009). Correlates of intellectual risk taking in elementary school science. *Journal of Research in Science Teaching*, 46(2), 210–223.

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, 7(4), 341.

Bradler, C., Neckermann, S., & Warnke, A. J. (2019). Incentivizing creativity: A large-scale experiment with performance bonuses and gifts. *Journal of Labor Economics*, 37(3), 793–851.

Brown, J. S., Collins, A., & Duguid, P. (1989). Commentary: Debating the situation: A rejoinder to palincsar and wineburg. *Educational Researcher*, 18(4), 10–62.

Carson, D. K. & Runco, M. A. (1999). Creative problem solving and problem finding in young adults: Interconnections with stress, hassles, and coping abilities. *The Journal of Creative Behavior*, 33(3), 167–188.

Chand, I. & Runco, M. A. (1993). Problem finding skills as components in the creative process. *Personality and Individual Differences*, 14(1), 155–162.

Charness, G. & Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17(2), 454–496.

Charness, G. & Grieco, D. (2021). Creativity and corporate culture. *Available at SSRN 3723865*.

Chen, C. X., Williamson, M. G., & Zhou, F. H. (2012). Reward system design and group creativity: An experimental investigation. *The Accounting Review*, 87(6), 1885–1911.

Cropley, A. J. (2000). Defining and measuring creativity: Are creativity tests worth using? *Roeper Review*, 23(2), 72–79.

Cropley, D. H., Kaufman, J. C., & Cropley, A. J. (2011). Measuring creativity for innovation management. *Journal of Technology Management & Innovation*, 6(3), 13–30.

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain*, volume 34. New York, NY: Oxford University Press.

Debelak, R., Egle, J., Köstering, L., & Kaller, C. P. (2016). Assessment of planning ability: Psychometric analyses on the unidimensionality and construct validity of the tower of london task (tol-f). *Neuropsychology*, 30(3), 346.

Eckartz, K., Kirchkamp, O., & Schunk, D. (2012). How do incentives affect creativity? Jena Economic Research Papers, no. 068.

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., & Schudy, S. (2021). The efficacy of tournaments for non-routine team tasks.

Erat, S. & Gneezy, U. (2016). Incentives for creativity. *Experimental Economics*, 19(2), 269–280.

Fimbel, E., Lauzon, S., & Rainville, C. (2009). Performance of humans vs. exploration algorithms on the tower of london test. *PloS one*, 4(9), e7263.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171–178.

Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). Lioness lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1), 95–111.

Gibbs, M., Neckermann, S., & Siemroth, C. (2017). A field experiment in motivating employee ideas. *Review of Economics and Statistics*, 99(4), 577–590.

Gillen, B., Snowberg, E., & Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4), 1826–1863.

Gneezy, U., Laske, K., & Schröder, M. (2021). Teams do not outperform individuals in a simple creative task. *Applied Economics Letters*, (pp. 1–6).

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114–125.

Gross, D. P. (2020). Creativity under fire: The effects of competition on creative production. *Review of Economics and Statistics*, 102(3), 583–599.

Grözinger, N., Irlenbusch, B., Laske, K., & Schröder, M. (2020). Innovation and communication media in virtual teams–an experimental study. *Journal of Economic Behavior & Organization*, 180, 201–218.

Gruber, H. E. & Wallace, D. B. (1999). The case study method and evolving systems approach for understanding unique creative people at work. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 93–115). Cambridge, UK: Cambridge University Press.

Guilford, J. P. (1950). Creativity. *American Psychologist*, 5(9), 444–454.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Hennessey, B. A. & Amabile, T. M. (2010). Creativity. *Annual Review of Psychology*, 61, 569–598.

Kachelmeier, S. J., Reichert, B. E., & Williamson, M. G. (2008). Measuring and motivating quantity, creativity, or both. *Journal of Accounting Research*, 46(2), 341–373.

Kachelmeier, S. J. & Williamson, M. G. (2010). Attracting creativity: The initial and aggregate effects of contract selection on creativity-weighted productivity. *The Accounting Review*, 85(5), 1669–1691.

Kaufman, J. C. (2012). Counting the muses: Development of the Kaufman Domains of Creativity Scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts*, 6(4), 298.

Kaufman, J. C. & Baer, J. (2002). Could steven spielberg manage the yankees?: Creative thinking in different domains. *The Korean Journal of Creative Thinking & Problem Solving*, 12(2), 5–14.

Kaufman, J. C. & Baer, J. (2004). Sure, i'm creative—but not in mathematics!: Self-reported creativity in diverse domains. *Empirical Studies of the Arts*, 22(2), 143–155.

Kaufman, J. C. & Baer, J. (2005). *Creativity across domains: Faces of the muse*. Psychology Press.

Kaufman, J. C., Waterstreet, M. A., Ailabouni, H. S., Whitcomb, H. J., Roe, A. K., & Riggs, M. (2010). Personality and self-perceptions of creativity across domains. *Imagination, Cognition and Personality*, 29(3), 193–209.

Kleine, M. (2021). No eureka! incentives hurt creative breakthrough irrespective of the incentives' frame.

Krikorian, R., Bartok, J., & Gay, N. (1994). Tower of london procedure: a standard method and developmental data. *Journal of clinical and Experimental Neuropsychology*, 16(6), 840–850.

Laske, K. & Schröder, M. (2017). Quantity, quality and originality: The effects of incentives on creativity.

Moran III, J. D., Milgram, R. M., Sawyers, J. K., & Fu, V. R. (1983). Original thinking in preschool children. *Child Development*, (pp. 921–926).

Myszkowski, N., Storme, M., Davila, A., & Lubart, T. (2015). Managerial creative problem solving and the big five personality traits: Distinguishing divergent and convergent abilities. *Journal of Management Development*, 34(6), 674–684.

Nazzal, L. J. (2015). *Engineering creativity: Differences in creative problem solving stages across domains*. PhD thesis, University of Connecticut.

Okuda, S. M., Runco, M. A., & Berger, D. E. (1991). Creativity and the finding and solving of real-world problems. *Journal of Psychoeducational Assessment*, 9(1), 45–53.

Olczak, P. V. & Kaplan, M. F. (1969). Originality and rate of response in association as a function of associative gradient. *The American Journal of Psychology*, 82(2), 157–167.

Plucker, J. A. (1998). Beware of simple conclusions: The case for content generality of creativity. *Creativity Research Journal*, 11(2), 179–182.

Plucker, J. A., Qian, M., & Schmalensee, S. L. (2014). Is what you see what you really get? comparison of scoring techniques in the assessment of real-world divergent thinking. *Creativity Research Journal*, 26(2), 135–143.

Ramm, J., Tjøtta, S., & Torsvik, G. (2013). Incentives and creativity in groups. CESifo Working Paper Series No. 4374.

Runco, M. A. (1994). *Problem finding, problem solving, and creativity*. Greenwood Publishing Group.

Runco, M. A. (2004). Creativity. *Annual Review of Psychology*, 55, 657–687.

Runco, M. A. & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92–96.

Rutström, E. E. & Williams, M. B. (2000). Entitlements and fairness:: an experimental study of distributive preferences. *Journal of Economic Behavior & Organization*, 43(1), 75–89.

Sarooghi, H., Libaers, D., & Burkemper, A. (2015). Examining the relationship between creativity and innovation: A meta-analysis of organizational, cultural, and environmental factors. *Journal of Business Venturing*, 30(5), 714–731.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2(2), 68.

Simonton, D. K. (1997). Histometric studies of creative genius. In M. A. Runco (Ed.), *Creativity Research Handbook* (pp. 3–29). Cresskill, NJ: Hampton.

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–293.

Stein, M. I. (1953). Creativity and culture. *The Journal of Psychology*, 36(2), 311–322.

Tierney, P. & Farmer, S. M. (2002). Creative self-efficacy: Its potential antecedents and relationship to creative performance. *Academy of Management journal*, 45(6), 1137–1148.

Torrance, E. P. (1962). Non-test ways of identifying the creatively gifted. *Gifted Child Quarterly*, 6(3), 71–75.

Torrance, E. P. (1966). *Torrance Test of Creative Thinking: Norms Technical Manual.* Princeton, NJ: Research Edition Personnel Press Inc.

Torrance, E. P. (1974). *The Torrance tests of creative thinking-TTCT Manual and Scoring Guide: Verbal test A, figural test.* Lexington, MA: Ginn.

Toubia, O. (2006). Idea generation, creativity, and incentives. *Marketing Science*, 25(5), 411–425.

Unsworth, K. (2001). Unpacking creativity. *Academy of Management Review*, 26(2), 289–297.

Ward, W. C. (1969). Rate and uniqueness in children's creative responding. *Child Development*, (pp. 869–878).

Measuring creativity: a network exploration task

1 Introduction

In times of automation and exponential improvements in computational processing power, abilities such as creativity that help translate technological advances into useful innovations are becoming more and more important. Thus, identifying a meaningful and robust experimental measure of creativity provides incredible opportunities for numerous scientific and industrial applications. Existing experimental measures of creativity however entail some significant limitations. These limitations can be grouped into two main categories. First, most existing measures treat creativity as a holistic concept rather than domain specific. In Study 1 of Chapter 4, we show that, contrasting this approach, there is no significant relationship between the performance across three *general* creativity tasks. Even when focusing on a single domain of creativity in Chapter 4's Study 2, we were unable to find a relationship in performance between another set of tasks. This suggests that more caution and stringency is required when employing experimental tasks for creativity.

The second critique relates to the question of experimental implementation. As discussed in the previous chapter, a way to classify creative tasks is by their degree of *openness* (Charness & Grieco, 2019). When examining existing methodologies, there appears to be a tension, where on the one hand a certain degree of *openness* is required for subjects to express creative ability. On the other hand, however, an increase in *openness* is likely to make the implementation of a task more cumbersome, subjective and susceptible to potential experimenter demand effects. Addressing both shortcomings in this chapter, we propose a novel method to capture creative ability that focuses on a single domain of creativity, *associative thinking*, and significantly improves on methodological properties. Contrasting other creativity tasks, our instrument includes an ex-ante defined solution space allowing for a clean implementation and objective performance measures, whilst ensuring sufficient *openness* for creative expression. To this end, we construct a semantic network, on which basis individuals perform a series of associative thinking tasks.

In particular in this chapter, we develop two distinct associative thinking tasks, each capturing a specific style of associative thinking, that are easily implementable in an experimental context.[1] As a first step to characterise the properties of our tasks, in a within-subject design, we benchmark each against a well-established creativity task that entails similar properties. We find evidence that performance in our proposed tasks correlates significantly with performance in their corresponding task, whilst improving experimental properties. In addition to relationships in performance, we also find that task behaviour, as identified by decision

This chapter is based on joint work with Urs Fischbacher, Chris Starmer & Fabio Tufano. I thank Werner Güth, Sandro, Sandro Ambühl and all CeDEx members and the participants of the ASFEE Conference 2019, ESA European Meeting 2019, CCC Conference 2019, and NIBS Workshop 2019 for valuable inputs. I am also grateful for funding by the Network for Integrated Behavioural Sciences (NIBS).

[1] In addition, the two tasks also address distinct levels of *openness*.

times, is identical in the benchmarking pairs of tasks but distinct between the two associative thinking tasks. Despite the differences in decision time behaviour, we do however find that performance in our two associative thinking tasks is significantly correlated. This suggests that both proposed tasks elicit a common facet of creative ability. Finally, we also vary incentive structures between subjects, finding incentives to only have an effect on one type of associative thinking.

Before describing our proposed method and design in more detail, let us first describe what *associative thinking* entails and why we think it plays a crucial role in innovation. Associative thinking is defined as "*the creative thinking process as the forming of associative elements into new combinations which [...] meet specified requirements [...]. The more mutually remote the elements of the new combination, the more creative the process or solution*" (Mednick, 1962, p.221). Therefore, associative thinking can be seen as the creative ability to combine knowledge and information in order to synthesise something new. This makes it a pillar of innovation. An illustrating example for the relevance of associative thinking is the so-called 'Medici-Effect' (Johansson, 2004), which takes its name from the Italian Medici dynasty in Florence in the 14^{th} century. Due to the wealth of the family, the Republic of Florence attracted various artist, painters, architects, scientists, philosophers and philanthropists. The accumulation and interaction of these individuals is considered to have driven years of ground breaking innovations and ultimately fostered the emergence of *The Renaissance*, thus playing a crucial role in the development of modernity. The underlying idea behind the 'Medici-Effect' is that being exposed to other scientific or artistic fields directly translates into innovation by association. The ability to associate knowledge and construct something novel thus becomes a key driver of innovation as well as entrepreneurial and scientific progress.

While the example above is intended to demonstrate the relevance of associative ability for innovation, we can also think of associative thinking as a key feature of creativity. In fact, Steve Jobs once described creativity as "*...just connecting things. When you asked creative people how they did something, they feel a little guilty because they didn't really do it, they saw something [...]. That's because they were able to connect experiences they've had and synthesize new things.*" (Jobs, 1996) This quote shows that for Jobs associative thinking is not only an ingredient of creativity but creativity itself. On a similar note Albert Einstein once stated that "*...combinatory play seems to be the essential feature in productive thought*". Einstein considered combinatory play as engaging with seemingly unrelated subjects in order to create new scientific discoveries. We therefore see the ability to connect dots and form novel associations not only crucial for innovation, but as the *bridge* between creativity and innovation. This is also supported by non-empirical studies that explicitly highlight the role of associative thinking to foster firm level innovation (Scott & Bruce, 1994; O'Connor & McDermott, 2004; Hittmár et al., 2014).

Due to this relevance, developing a robust experimental method that measures individual associative ability seems key in order to study and identify its role in a variety of contexts. In addition, we also improve on existing creativity tasks with respect to the experimental properties of our measure. As mentioned above, a main concern for experimental creativity tasks is a loosely defined solution space. Whilst this allows for creative expression of subjects, it

does result in methodological issues when evaluating and scoring creative performance. Scoring is usually subjective, cumbersome and based on various definitions and understandings of creativity. Moreover, for an implementation in economic experiments including incentives, there is a need to recruit other subjects or researchers as evaluators. This complicates the implementation of these tasks when using incentives and hinders further scientific progress in this area. To address these issues in our proposed method, we ex-ante create a semantic network which serves as the underlying skeleton of our proposed tasks. The network consists of English nouns which are connected by meaning.[2] We then define a valid association as a connection between two words in the network and ask subjects to perform tasks to discover these connections. Utilising the underlying network structure ensures a perfectly defined solution space and allows for automated scoring, thus simplifying the use of incentives.

Our proposed method and its application in this study contribute to the psychological literature on associative thinking and its relationship to creativity (Mednick, 1962; Bowden et al., 2005; Batey & Furnham, 2006; Benedek et al., 2012; Beaty & Silvia, 2012; Lee & Therriault, 2013; Beaty et al., 2014; Verhaeghen et al., 2017), with most studies finding associative thinking to play a significant part in creative performance. Furthermore, as we are exploring semantic associations, we are directly building on work by Gough (1976), who first introduce *word association tasks* as a measure of creativity. Subsequent research has extended his approach by examining distance of associations Acar & Runco (2014) or neural involvement in associative thinking Whitman et al. (2010).

We are however most closely contributing to the literature on experimental creativity tasks used in economics as outlined in Chapter 4. Addressing various identified shortcomings of other tasks, we aim to provide the scientific community with an easily implementable tool to conduct more research on associative thinking. Moreover, as we also explore incentive effects within our study we add to the literature on the role of incentives for creativity.

This chapter is structured as follows. In Section 2, we outline the construction of our semantic network. In Section 3, we describe our proposed tasks and provide the experimental design to benchmark them against existing creativity measures. In Section 4, we present results of the benchmarking. Lastly, in Section 5, we discuss our findings and provide concluding remarks.

2 The semantic network

As outlined in the introduction, at the heart of our proposed method lies a semantic network, where words are connected based on their meaning. Thus, before presenting the details of our proposed tasks, we first describe the construction and properties of this network.

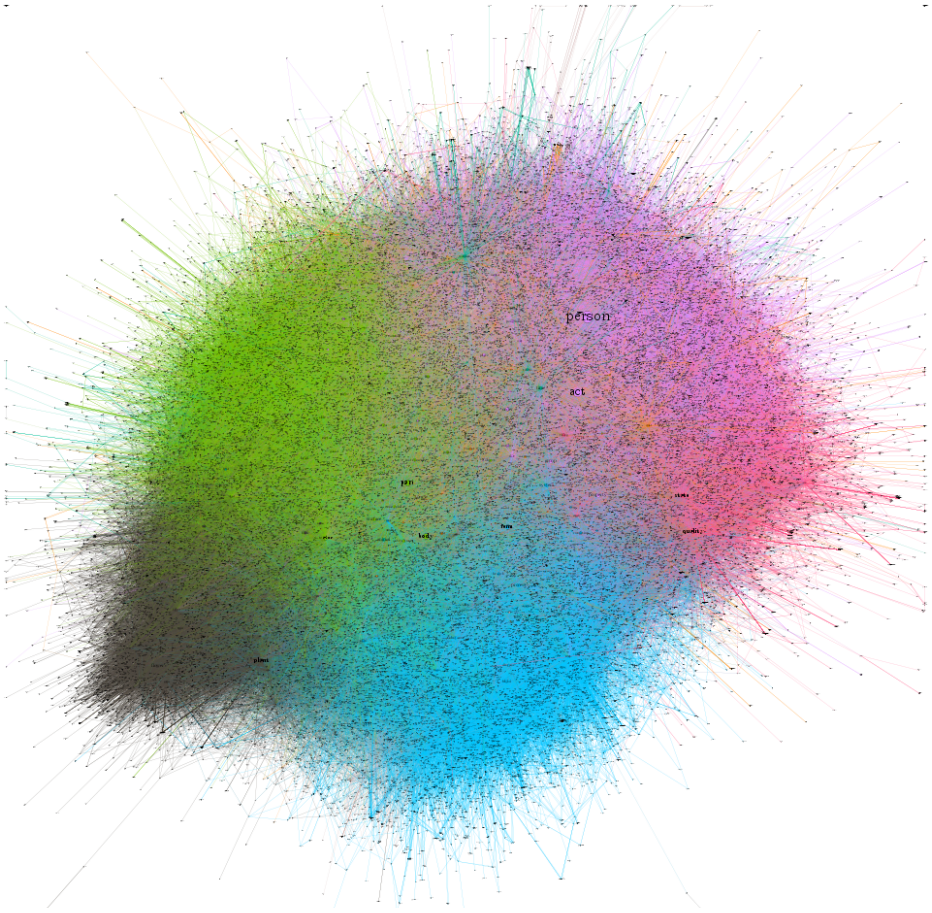
In order to associate words by meaning, we base our entire network on dictionary definitions. By construction all words within the same definition share some relationship as they all refer to a unique underlying theme (i.e. the word explained). However, not all words in definitions entail meaning that is useful for our purposes. For example the definition of pro-

[2]Section 2 provides details regarding the construction of the network.

crastinator is "*someone who postpones work especially out of laziness or habitual carelessness*". When examining the definition, it is apparent that in particular *nouns* share meaningful information, *procrastinator*, *work*, *laziness*, and *carelessness*, whereas verbs, adjectives or adverbs entail considerably less related meaning. Therefore, within our semantic network, we say there is a link between two words if they are nouns that occur in the same definition. As the same noun is likely to occur in several definitions we can then construct a network of all nouns in a dictionary. To do so we use the open source *WordSet* dictionary entailing around 177,000 entries.[3] Using a *natural language processing tool* (Stanford Parser),[4] we extract each noun in the dictionary and identify the associated nouns in its definition.

Following our outlined definition of a link, we construct a network spanning across 41,234

Figure 5.1: The semantic network



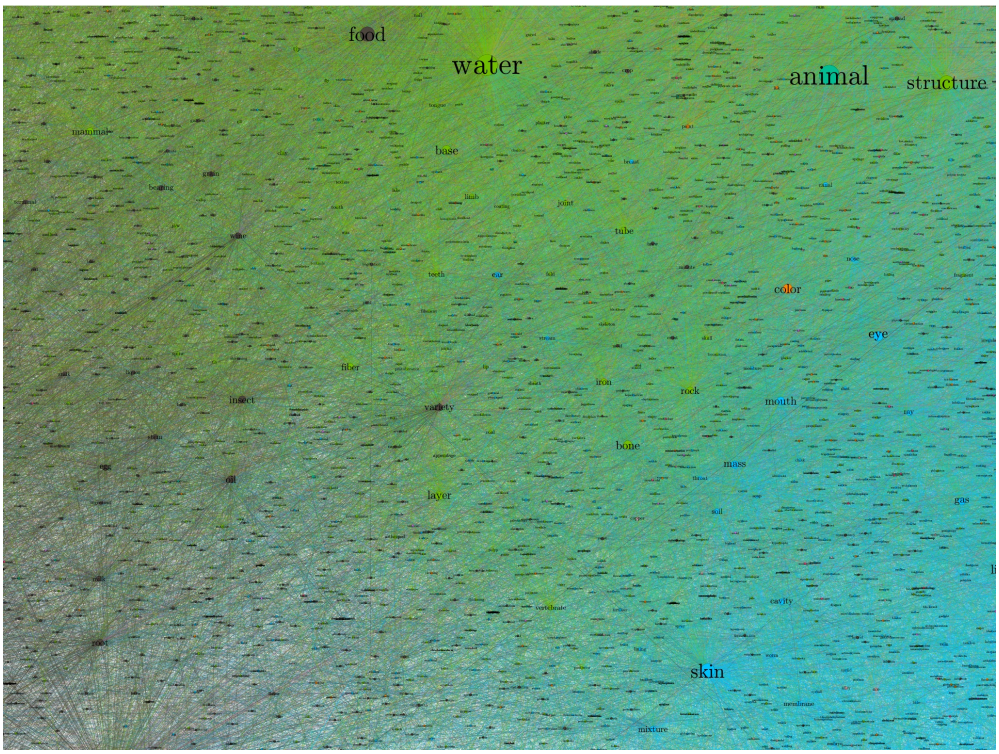
Note. This network is constructed as outlined in the text. The colours correspond to the 'modularity class', which is assigned by an algorithm identifying more connected words. Moreover, larger words are due to more associations for this word.

[3]The dictionary is a collaborative project initially based on the WordNet project of the University of Princeton.
[4]Developed by the Natural Language Processing (NLP) Group of the University of Stanford. See https://nlp.stanford.edu/software/lex-parser.shtml.

words with a total of 275,247 links between them.[5] Provided the completeness of the underlying dictionary, the network encompasses all one-word nouns in the English language and their relationship with each other. To provide additional details regarding the network, each word is on average connected to 13.44 other words and the longest possible path between two words is of length 8, showing a considerable distance between words.[6] Moreover, there is substantial heterogeneity within the network. The word with the most associations (*person*) has 4,296 other nouns connected to it, while there are 4,226 words (10.3%) that only have a single association.

For illustration purposes, Figure 5.1 depicts the entire network, where components that are more densely connected are colour coded and words with more associations are larger.[7] Moreover, the colouration of the network is solely based on links and does not rely on any information about the meaning of words. When examining a snapshot of the network in Figure 5.2, we can see that words that are intuitively considered as associated are also identically coloured in the network. For instance, if we look at the word 'skin' on the bottom right of the figure, we can find words such as 'membrane', 'lining', 'mouth', 'eye' to have the same colour and are allocated close in the network. This shows that our initial approach of defining nouns present within the same definition as 'linked by meaning' actually produces intuitive

Figure 5.2: A snapshot of the semantic network



[5] All details and files to recreate and utilise the network can be obtained upon request.
[6] For reference, according to the idea of 'six degrees of separation' by Milgram (1967), every pair of individuals living in the world is connected by a path of length 6.
[7] We use the open-source software *Gephi* to illustrate networks, which includes a feature to compute a 'modularity class' to identify connected components of networks.

associations between words.

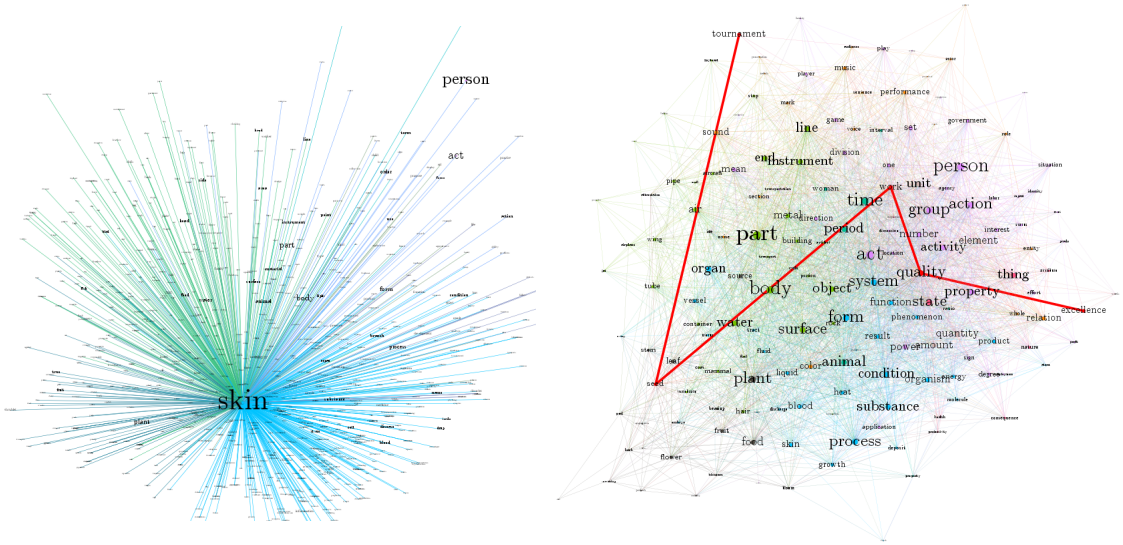
3 Developing and benchmarking our associative thinking tasks

Having constructed this network, we now use it as the underlying structure for two distinct associative thinking tasks, which we label *Local Search* and *Depth Search*.

In the *Local Search* task, subjects are presented with a single word from the network and are instructed to list as many direct associations as possible to the presented word as they can find. Their performance is then evaluated based on the number of successfully identified links. The task thus challenges and assesses the ability to use associative thinking for divergent creativity, meaning the ability to generate diverse ideas by association from a single provided starting point. While the task allows for a relatively free and open form of creative thinking, it is simple to evaluate performance by comparing submitted words to associations in the network. Our semantic network thus provides a perfectly defined solution space, ensuring that the task can be automatically scored without any additional evaluations by researchers or other subjects. The left picture in Figure 5.3 illustrates this feature of the task, as it depicts all valid associations to the word 'skin'.

Contrasting that, our second associative thinking task requires subjects to travel from one word to another in the shortest time possible across the pre-defined network. We label this task *Depth Search*, as subjective have to consider deeper associations within the network, including some that are multiple steps away from their current position. In contrast to *Local Search*, this task entails a *closed* solution to a problem by means of a perfectly defined target. Subjects are thus asked to identify the most relevant associations to reach this target. An example of *Depth Search* is provided in Figure 5.3 (right-hand picture). It depicts a possible shortest path from

Figure 5.3: Proposed associative thinking tasks



Note. Examples of our associative thinking tasks. *Local Search* on the left, and *Depth Search* on the right. All links in the figure correspond to existing links in the semantic network.

'excellence' (right) to 'tournament' (top). In this case a possible path would be via the words 'quality', 'work' and 'seed'. However, there are of course many more possible paths and also some other *shortest* paths for the same start and target words.

Moreover, after having described both of our tasks, we can also use the categorisation of *task openness* as examined in Chapter 4, to further illustrate differences between the two. First, contrasting the study by Charness & Grieco (2019), both of our tasks include a delineated goal and should therefore not be considered as entirely *open*. However, as *openness* has to be considered as a continuous measure, we can still classify *Local Search* as substantially more open than *Depth Search*. The reason is that despite subjects being instructed to find *valid* associations, what is considered as *valid* is not precisely known when completing the task. *Depth Search*, on the other hand, includes very little ambiguity, which is why we classify it as *closed*.

We now explore their characteristics by relating them to commonly used creativity tasks. Despite shortcomings of some established creativity tasks, we think their utilisation has shown a consensus in what is considered creative ability but their methodological properties imply challenges for use in experimental economics. Thus, showing that our proposed tasks elicit identical creative ability as established methodologies within a single structural framework whilst improving methodological properties provides considerable support for our tasks. The two tasks we chose as benchmarks are the *Unusual Uses Task* (UUT) and the *Tower of London* (ToL) (both utilised in Chapter 4). The reason why we think those tasks are well suited is that they entail similar features as the *Local Search* and *Depth Search* task and we think they require comparable types of creative ability.

In the next section, we will thus explain the implementation of our tasks and their relationship to the established creativity tasks in more detail.

3.1 Local Search and Unusual Uses Task

For the *Local Search* task, subjects have 2 minutes to generate as many different word associations for a given word as possible. In line with our network structure, subjects are only allowed to submit singular English nouns and each submitted word is scored according to the pre-defined network, where for each valid word subjects receive one point. To ensure comparability to the creativity task, subjects did not receive any feedback regarding validity of their submissions during the experiment.

Before outlining the matched creativity task, we first describe how we selected the words subjects faced during our experiment. Since any of the 41,234 words could serve as a base word for *Local Search*, we follow a systematic procedure to randomly choose words for the task conditional on clearly defined criteria. First, we exclude all words that have fewer than 500 possible answers ensuring a large enough solution space. This leaves us with a possible set of 72 words. In a second step, we exclude any word that has a centrality score smaller than 0.30. In this instance we use 'eigenvalue centrality', which is a common network measure to evaluate *importance* of a node. Using it as an exclusion criteria consequently ensures that the remaining words are prominent and provide good starting points for subjects, leaving us

Table 5.1: Problem overview of the *Local Search* task and the *Unusual Uses Task*

Local Search	# of possible associations
Metal	836
Head	979
Skin	831
Instrument	833
Unusual Uses Task	# of possible associations
Cardboard	N/A
Tin can	N/A
Paperclip	N/A
Brick	N/A

Note. A list of all possible solutions for the *Local Search* task can be found in Appendix E.3.1. The unusual uses task does not include a number for associations due to its undefined solution space.

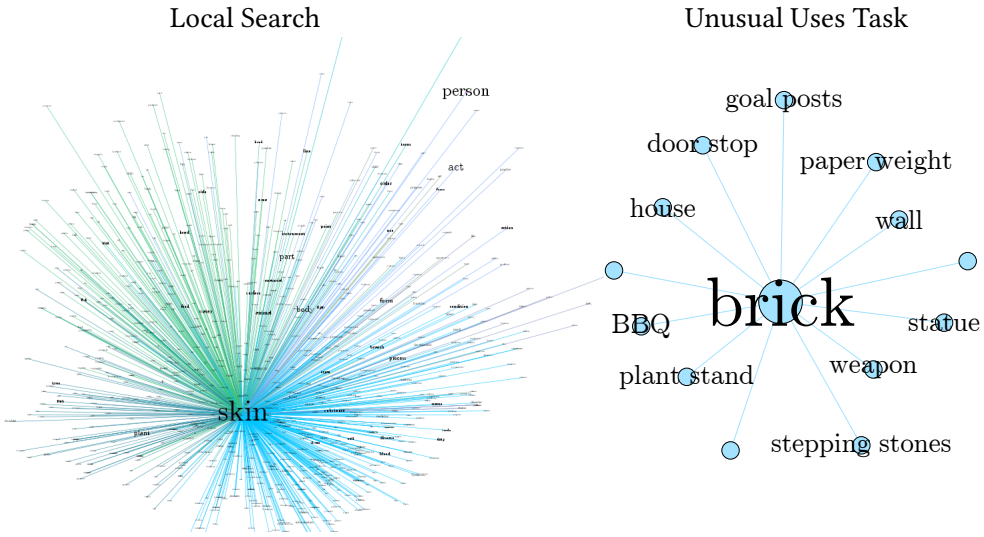
with 32 possible words. We then exclude all words that are abstract nouns such as 'activity' or 'term'. Lastly, for the remaining 10 words we explore the overlap of solutions spaces and exclude words with more than 20% overlap. After applying these criteria we are only left with six words from which we randomly select 4. The final words for the *Local Search* task used in our experiment are *metal*, *instrument*, *skin*, and *head*. Table 5.1 shows the number of associations for each word.

As mentioned above, to benchmark our *Local Search* task we compare it to the *UUT*, one of the most commonly utilised creativity task in the experimental literature. During the task, subjects are presented with a specific object and have to name as many *uses* for it as possible. Mimicking, the *Local Search* task, in our set-up, subjects have 2 minutes to come up with uses for a given object.[8] Upon completion, submissions are scored and examined for validity. Table 5.1 shows which objects we used for the *UUT* in the experiment. Moreover, the table also shows how the performance measure differs between tasks. While *Local Search* has a clearly defined solution space provided by the network, the *UUT* does not provide ex-ante solutions (*N/A* in Table 5.1).

In addition to its popularity, we think the *UUT* is a good benchmark for the *Local Search* task, as both tasks ultimately address divergent thinking abilities. In other words, both focus on the ability of subjects to generate many different solutions. Either task challenges subjects to form associations from a pre-defined starting point and generate associated answers. Performance is once evaluated by the number of *valid* links (as defined in the network) and once by the number of *feasible* uses for an object. To better illustrate this relationship we provide a schematic overview of both tasks in Figure 5.4. The picture on the left shows again the example of 'skin' and all possible associations based on our network. Similarly, on the right we provide example uses of a 'brick'. However, in this case it is not possible to represent all solutions, as there is a degree of subjectivity in what use is considered *feasible*, represented by

[8]Our findings from Chapter 4 suggest that a task duration of 2 minutes is sufficient to provide stable variation between individuals.

Figure 5.4: Visual comparison of *Local Search* and *UUT*



Note. Examples of *Local Search* (left) and the *Unusual Uses Task* (right). The missing labels in the right graph indicate the undefined solution space in the *UUT*.

the unlabelled associations in the right picture.

Given the *open* solution space of the *UUT*, it is not trivial to decide which performance measure should be used to compare the two tasks. As discussed in Chapter 4, typically the *UUT* is assessed by three distinct metrics: the total number of uses submitted (*fluency*), the distinct categories covered (*flexibility*) and statistical scarcity (*originality*). Despite these dimensions being seemingly intuitive, the latter two are challenging to implement as they require external evaluators and ultimately rely substantially on which categories and uses are considered as distinct.[9] In Chapter 4 we found substantial correlations ($r > 0.61$; p-value < 0.01) between the three performance measures, which is why for this study we only use the 'objective' dimension of *fluency* to evaluate performance in the *UUT*.

3.2 Depth Search and Tower of London

Let us now turn to our second associative thinking task. As outlined above, during the *Depth Search* task we present subjects with a specific start and target word and ask them to travel from start to target as fast as possible. Moreover, to ensure that we disentangle *Local* and *Depth Search* as much as possible we provide subjects with all direct next steps once they reach a specific word. For example, starting at the word 'screwdriver', subjects can observe all words linked to that word, for example 'handle', 'screw' or 'metal'. Once a subject chooses to continue selecting, for instance, the word 'handle', they can then observe all directly linked words for 'handle' and so on.[10] This way, subjects do not have to generate their own associations but

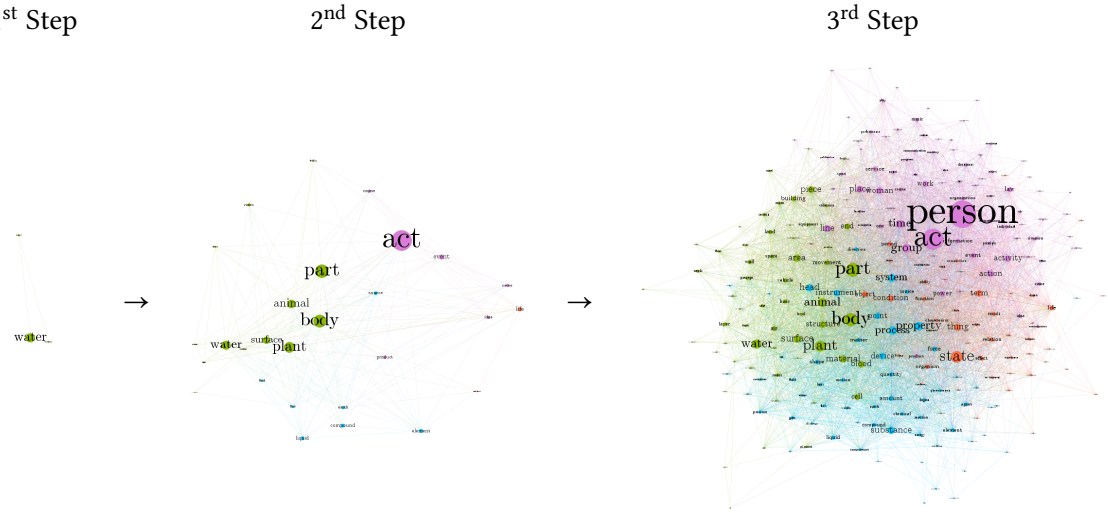
[9] Suppose the object provided is a *brick*. When providing answers, one subject could give uses as "building a house, building a school, building a church", whereas another subject might just say 'building a building". Evaluators then have to decide whether the first subject submitted three uses, or a single use like the second subject. The *UUT* thus leaves substantial room for subjective interpretations and scoring, which complicates its implementation.

[10] See screenshots of the task in Appendix E.2.

they only have to identify words that are most likely to move them closer to the target word. Within our experiment, subjects have 2 minutes to find a path between the start and target word, we then use *completion time* as our measure of performance. Note that, contrasting to the implementation in Chapter 4, we decided to use completion time and not the number of moves as a metric, as this allows subjects to experiment during the task and move back after identifying a wrongly taken step. As we consider experimentation a key part of creative ability we did not want to penalise a wrongly taken step.[11]

Choosing the specific paths that are utilised in our experiment was again not trivial. The complete network entails words that are directly connected to more than 1,500 other words (e.g. *body*, *state* or *plant*). As it is impossible to visually present all these links to subjects we did not use the entire network for *Depth Search* but rather a smaller sub-network. For its construction, we start at a specific word, which we label the *seed* of the sub-network. We then identify all nouns that are in the *seed's* definition. Following that, we inspect the definitions of all these nouns to identify their nouns and iterate this step once more. This provides us with a set of words that all stem from the same *seed*.[12] To connect these, we then use the links from our complete network. Figure 5.5 visually represents the iterative process for the seed 'river'. We can see that by only taking nouns that are in the definition of 'river' and then add nouns that are in the definitions of those, we already generate a dense sub-network by step 3. The big advantage of these smaller sub-networks is that they allow us to visually present subjects

Figure 5.5: Sub-network creation for *Depth Search* (seed: *River*)



Note. The 3-step iterative process of constructing a sub-network for the example of 'river'. In Step 1, we find the definition of 'river', in Step 2, the definitions of the nouns in the definition of Step 1. Step 3 follows the same procedure based on all definitions identified in Step 2.

[11] Of course, moving back and forth through the network will need more time than identifying the desired path straight away, however including moves as an additional performance measure would imply an added penalty for experimentation.

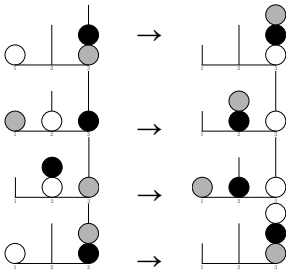
[12] Note that it does not have to be the case that these words are in fact linked as in our complete network if 2 words during the iterative process never appear in the same definition but they do in some other definition not examined.

with all direct associations for the *Depth Search* task.[13]

To identify which sub-networks to use in our experiment, we follow a similar process of exclusion as for the *Local Search* task. After having constructed all possible 41,234 sub-networks,[14] we first exclude any sub-networks with fewer than 150 and more than 300 words.[15] This leaves us with 11,036 possible networks, out of which we randomly select 15 as candidates for our final tasks. To ensure that there is sufficient distance between possible start and target words, we moreover only include sub-networks with a longest path larger or equal to 4. This step narrows our number of sub-networks to 6, which we examine for suitable paths for our *Depth Search* task.

The final problems for the experiment are then chosen as follows. First, in line with our previous step, in each sub-network we only include paths of at least length 4. We then exclude paths where either start or target word have fewer than three connections, to provide subjects with sufficient possibilities and choices when starting a path or converging towards the target.[16] This leaves us with 49 possible paths across the six networks, of which we randomly select one per network. Lastly, to match the number of problems in the *Local Search* task we

Table 5.2: Problem overview of the *Depth Search* task and the *Tower of London*

Depth Search		
Sub-network seed	Path	Path Length
Rudd	*dereliction* → *ounce*	4
Dhal	*excellence* → *tournament*	4
Crinoid	*conduit* → *outburst*	4
Syringe	*cold* → *schematic*	4
Tower of London		
	Path	Path Length
		6 7 6 6

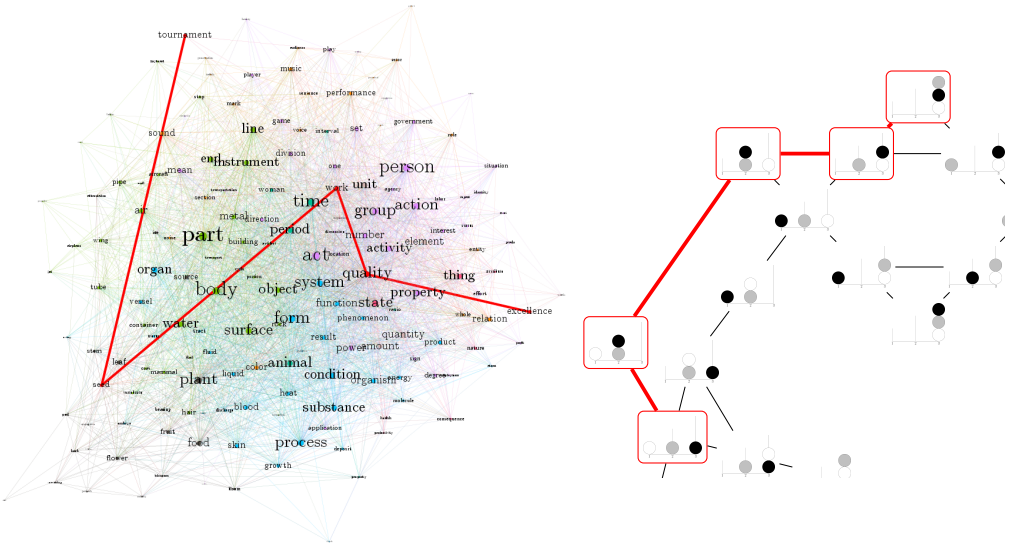
Note. Depth Search problems (top) include the seeds of the sub-networks in column 1. All *ToL* problems are within the same network. The networks for both tasks can be found in Appendix E.3.2.

[13] In fact, if subjects cannot observe all possible next steps at ease, they automatically engage in *Local Search*, which is what we explicitly aim to avoid, hoping to disentangle the two associative thinking styles as much as possible.
[14] 41,234 is the number of words in the network and therefore the number of possible seeds for sub-networks.
[15] We arguably have to strike a balance between not making the sub-network too convoluted to compromise the implementability of the task and not making it too small so that the solution is trivial. Experience from piloting lead to the conclusion that 150-300 words provides a good compromise.
[16] Suppose, a word is only connected to one other word, in this case there is no reasoning required at all to move in a certain direction, rendering the first step redundant.

again randomly choose four out of the remaining six paths. The final sub-network seeds and path details are presented in Table 5.2, for an illustration of the final four sub-networks, please see Appendix E, Figure E.3.1.

To benchmark the *Depth Search* task, we chose another established, and frequently used, creativity task: the *Tower of London* (ToL). As we already utilised this task in the previous chapter, we will only briefly summarise its key elements. In this task subjects have 1 minute to move three objects that are positioned on three rods as fast as possible from a start to a target position.[17] Like in the *Depth Search* task, we use *completion time* as our measure of performance.[18] The three rods fit differently many objects, from one to three. Thus, subjects have to plan ahead and identify best moves in order to reach the target position. Moreover, as there is a finite number of moves, the solution space of the *ToL* can be displayed in a network, where solving a problem is equivalent to finding a path in that network.[19] For this reason, we think that the *ToL* requires subjects to engage in similar projective reasoning to *Depth Search*. Another key similarity between both tasks is that subjects are always aware of all possible next moves in both environments. Figure 5.6 again compares an example of both tasks to further illustrate their similarities. The left picture shows the *Depth Search* task by illustrating the shortest path between 'excellence' and 'tournament' in a sub-network.[20] The right picture

Figure 5.6: Visual comparison of *Depth Search* and *ToL*



Note. Example for *Depth Search* in the left graph for the path 'excellence; to 'tournament'. An example problem of the *ToL* on the right.

[17] In a pilot we found that a ratio of 2:1 in total time provides similar levels of success rates. Thus, *ToL* problems could be considered half as complex as *Depth Search* problems.

[18] Note that in Chapter 4 we used successful completions with minimum moves as a measure for performance. However, we here want to avoid to penalise experimentation, which is why completion time seems to be the more appropriate measure. Moreover, we found strong correlations between both metrics in the previous chapter.

[19] See Appendix E.3.4 for a graphical representation of the network.

[20] Which has the seed 'dhal' and was utilised in the final experiment.

on the other hand shows a section of the *ToL* solution network and depicts a shortest path between two of its positions. Figure 5.6 highlights the similarities between the tasks as both rely on a pre-defined network and challenge subjects to identify a path within it.

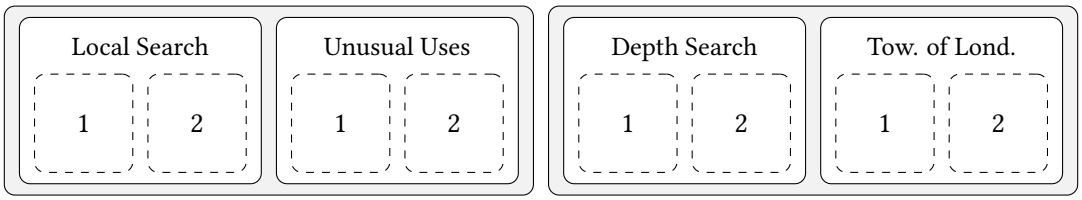
To select the problems for the *Tower of London* we only examine paths of length 6 or longer such that, similar to *Depth Search*, we have enough complexity to measure associative ability. After this exclusion criteria we then randomly choose four out of the 213 possible paths, ensuring they are not isomorphic.[21] Table 5.2 lists all four final problems of the *ToL*, as well as *Depth Search* including the starting and target positions/words. Note that while *Local Search* has the clear advantage of a pre-defined solution space over the *UUT*, both *ToL* and *Depth Search* allow for an objective, well defined, measure of performance. The reason for developing the *Depth Search* task is to provide a *closed* associative thinking task based on the same underlying structure as an *open* one (*Local Search*). This increases our possibility to compare associative thinking ability across levels of *openness*.

3.3 Procedure and incentives

To execute this benchmarking, we employed a within-subject design where each subject completed a total of four distinct tasks. Both of the two network tasks, *Local* and *Depth Search*, and the two creativity tasks, the *Unusual Uses Task* (UUT) and the *Tower of London* (ToL). Moreover, we asked subjects to complete each task twice to further investigate performance robustness in all tasks. The exact problems were drawn randomly for each individual. Figure 5.7 provides an overview of our experimental design. We can think of the study as consisting of two main blocks: *Local Search* and *UUT* as well as *Depth Search* and *ToL*. To control for potential learning or order effects, we randomised the order of all four tasks and within each task the chosen problems.

After subjects completed the main experiment, we elicited several individual characteristics in an ex-post survey. As our proposed associative thinking tasks are inherently based on language, we measured linguistic proficiency using the *LexTale* vocabulary task.[22] During the *LexTale*, subjects were presented with 60 strings of letters and had to decide for each string whether it is an existing English word or not. Moreover, to further control for language profi-

Figure 5.7: Experimental overview



Note. Each subject completes all parts outlined in the figure. The order of all tasks is randomised. The 2 problems for each task are randomly drawn from a set of 4.

[21] As the *ToL* includes objects with three different colors, each problem can be represented in 6 different ways.
[22] see http://www.lextale.com/index.html, Numerous studies have shown that the task successfully captures language abilities on individuals.

ciency, we only recruited subjects with English as their first language. Secondly, in line with Chapter 4 we also elicited two self-reported associative ability and creativity scales, the 'associative and bisociative problem-solving ability scale' (Jabri, 1991) and the 'Ideational Behavior Scale' (Runco et al., 2001) as well as basic demographics.

A crucial advantage of our proposed network tasks is that the ex-ante defined solution space means that both can be easily incentivised. To contribute to the emerging literature regarding incentives for creativity (Attanasi et al., 2021), we thus also explored incentive effects within our study. We thereby varied incentives between-subjects, with each subject either receiving a flat payment for completing all tasks, or piece-rate incentives. More precisely, subjects could receive £0.50 for each valid word and use submitted in the *Local Search* task and *UUT* respectively. In the *Depth Search* task subjects could receive £0.10 for each second that remained from the 2 minutes after reaching the target word.[23] Since subjects had 2 minutes to complete a *Depth Search* problem but only 1 minute for the *Tower of London*, they received a bonus of £0.20 for each second left in the latter. At the end of the study, we selected one of the eight problems at random for payment.

Overall, using Prolific we recruited 400 subjects, half of which participated in the incentivised treatment.[24] All data was collected online in April and July 2020. The experiment was approved by Nottingham School of Economics' Research Ethics Committee and programmed using LIONESS Lab (Giamattei et al., 2020). On average subjects took around 30 minutes to complete all 8 tasks and questionnaires and earned on average around £12.50/hr across both treatments. Note that earnings differed between incentivised and non-incentivised conditions. Subjects in the incentivised condition earned on average a bonus of £6.80, resulting in a mean payment of around £18/hr including the participation fee, while participants in the non-incentivised condition received the flat fee of £7.20/hr. Our incentives are thus quite substantial, ensuring that in case incentives matter for performance we should be able to identify this effect with our set-up.[25]

3.4 Hypotheses

Before moving on to our results, we briefly state hypotheses with regards to the benchmarking exercise described above.

1. Performance in the *Local Search* task and the *Unusual Uses Task* is significantly correlated. Subjects require a similar ability to perform well in both tasks.

2. We also expect a significant relationship in performance between the *Depth Search* task and the *Tower of London*. Both tasks elicit a similar associative thinking ability.

3. Since both pairs of tasks vary in their level of *task openness*, we do not expect a correlation between the two associative thinking or creativity tasks.

[23] If it took a subject 60 seconds to complete the task, they would thus earn for instance (120-60)*£0.10 = £6.
[24] See www.prolific.co for more information on their services.
[25] Instructions for all tasks can be found in Appendix E.2. We also provide a link to go through the experiment interactively.

4. In line with Chapter 4, we expect no correlation between survey measures and task performance across all four tasks.

5. We expect positive incentive effects across tasks, but more so in the *Depth Search* task and the *Tower of London* as these tasks are characterised by lower levels *openness* (see Charness & Grieco, 2019)

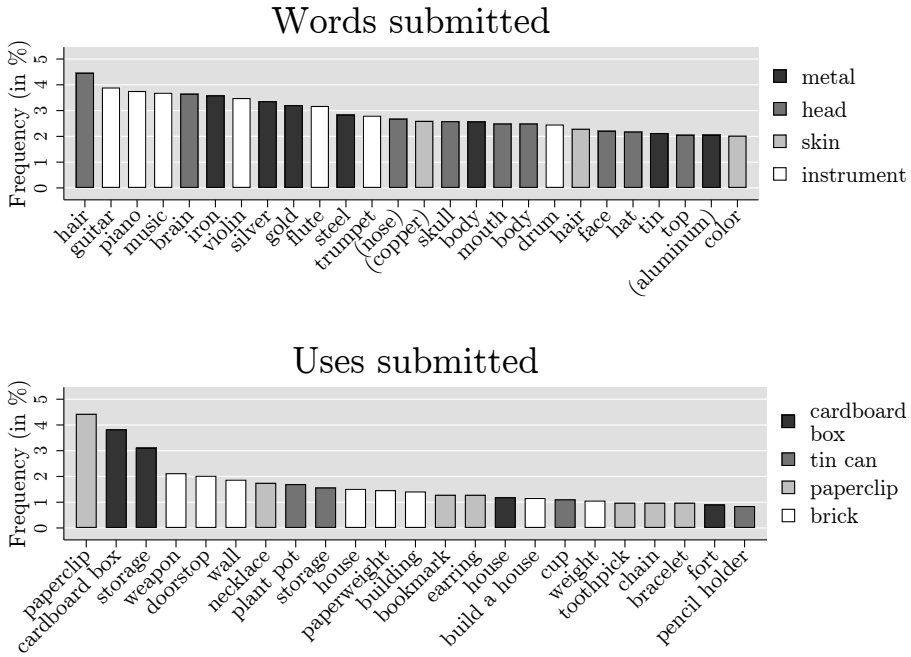
4 Results

This section is structured as follows. For each benchmarking pair of tasks, we first provide descriptive overviews before analysing the relationship between the tasks. We then move on to a comparison across all four tasks. Lastly, we explore possible incentive effects.

4.1 Local Search and Unusual Uses Task

Before diving into the comparison of the *Local Search* task and the *Unusual Uses Task*, let us examine behaviour in both tasks. Figure 5.8 demonstrates the most common words/uses submitted across problems in the respective tasks, where the *Local Search* task is in the top graph and the *UUT* on the bottom. In each graph the different underlying words/objects are represented in different colours. We can see for instance that the most common association for the word 'head' was 'hair', while the most common use mentioned for a 'paperclip' was using it as a 'paperclip' itself. The words in brackets for the *Local Search* task indicate incorrect submissions. Despite some frequently submitted words being incorrect, one can see that the

Figure 5.8: Most common submissions in the *Local Search* task and the *UUT*

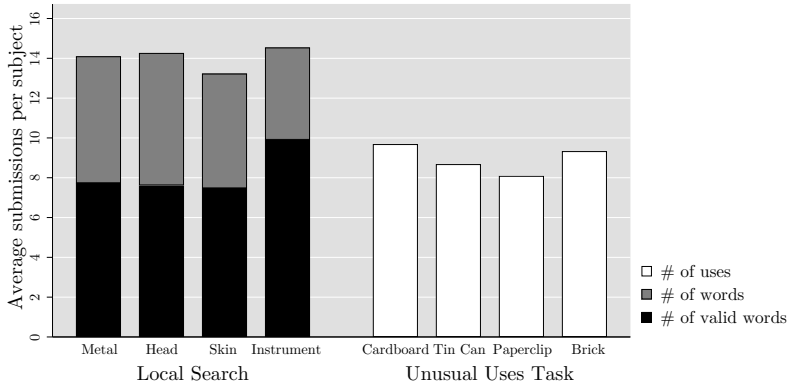


Note. All words/uses submitted across all problems in the *Local Search* task and the *UUT*. The vertical axis shows the frequency of occurrence across all subjects in percent.

most commonly named words are also linked within our semantic network. Overall, we find that on average 58.22% of submitted words were in fact linked in our network. This suggests that our network does capture what subjects intuitively consider associated words.

When examining submissions in the *UUT*, we can see that common submissions decline sharply after the first uses, which is likely due to the fact that some uses in Figure 5.8 actually have the same meaning (e.g. 'building' and 'building a house'). This indicates a previously mentioned weakness specific to the *UUT*, where some subjects submit similar uses as distinct ones, whereas others consider them as identical.[26]

Figure 5.9: *Fluency* scores across problems and tasks



Note. The bar chart plots all submitted words and uses. For the *Local Search* task on the left, the bar is split between all submitted words (grey) and words that are valid according to the underlying network (black).

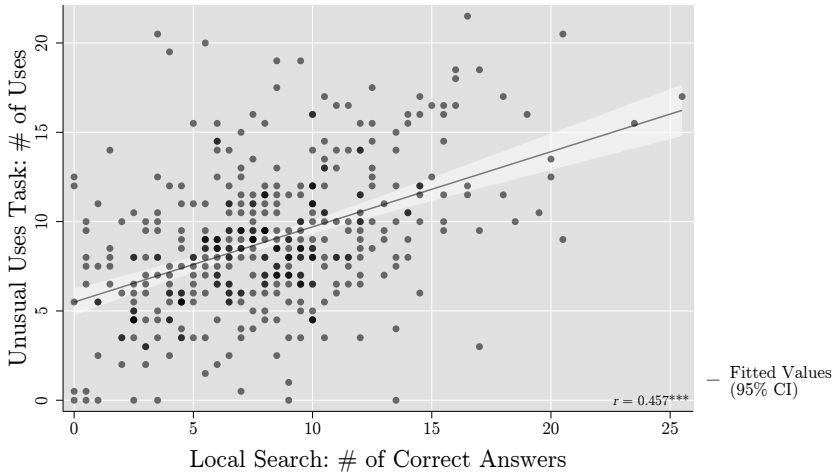
We next turn to the analysis of *fluency*, our measure of performance in both tasks. Recall that by *fluency* we mean the number of *valid* submissions in each task. Figure 5.9 shows the total number of submissions in the *Local Search* task and the *UUT*, including the separation into *valid* words for the former (black).[27] The figure therefore allows us to compare the number of valid words in the *Local Search* task to the submitted uses in the *UUT*. We find surprising similarities in average valid submission numbers between both tasks (8.17 words and 9.27 uses respectively), implying a comparable degree of difficulty in both tasks. However, similar numbers of submissions, do not imply that subjects' performance correlates between tasks.

To investigate this question we regress performance in the *Local Search* task on performance in the *UUT*. Figure 5.10 visualises the *Local Search* as being a significant predictor for performance in the *UUT* task, showing a significant and substantial correlation between both tasks ($r = 0.457$; p-value < 0.01). This suggests that in fact subjects who perform well in the *UUT*, also perform well in our proposed *Local Search* task. Therefore, in line with Hypothesis 1, our evidence suggests that our proposed method does measure the same associative ability

[26] The only option to resolve this issue is to evaluate all submissions individually and classify them as distinct or identical, which within our data corresponds to 5,243 evaluations and re-classifications. Evaluating submissions in *Local Search* by contract can be directly implemented within the study, as only words linked in the semantic network are considered correct.

[27] The number of invalid uses in the *UUT* is negligible and therefore not represented in the Figure 5.9.

Figure 5.10: Relationship between the *Local Search* task and the *Unusual Uses Task*



Note. The fitted line corresponds to a univariate regression between the *Local Search* (horizontal) task and the *UUT* (vertical). Pearson's correlation with significance is provided in the bottom right corner of the graph.

that is required for the *UUT*, whilst substantially improving methodological properties. This result is robust to the inclusion of controls into a regression (Appendix able E.1.1). In addition to performance in the *UUT*, we find that younger subjects as well as those with a higher measure of linguistic proficiency perform better in the *Local Search* task. The best predictor of performance in *Local Search*, however remains performance in the *UUT*.

Result 1. *Performance in the Local Search task and the Unusual Uses Task is significantly correlated. This is robust to the inclusion of additional controls.*

We can interpret Result 1 as evidence that both of the explored tasks measure a 'common associative root' related to creative ability. Finally, when examining within-task robustness, we find strong correlations between the two rounds that subjects completed of the same task. This robustness in performance supports the notion that our tasks do in fact measure a stable individual ability. In fact, we can utilise this specific design aspect to again apply the *Obviously Related Instrumental Variable* (ORIV) approach as in the previous chapter (Gillen et al., 2019). By instrumenting both measures with each other the we can account for potential measurement error to provide *cleaned* correlations. As we have two measures of the *Local Search* task and the *UUT*, we can reduce noise in both sides of the correlation. Using the *ORIV* method, we find an increased correlation coefficient ($r = 0.59$; p-value < 0.01), indicating that the relationship between the two tasks might be even stronger than in the raw correlation. Nonetheless, further supporting our findings, all observed relationships also hold when examining the rounds independently (Appendix E, Table E.1.2).

4.2 Depth Search and Tower of London

Having found a significant relationship between the *UUT* and the *Local Search* task, we now again turn to the relationship between the *Depth Search* task and the *Tower of London*. To first get an impression of behaviour in the respective tasks, Table 5.3 provides descriptives of a

Table 5.3: Descriptives - Depth Search and Tower of London

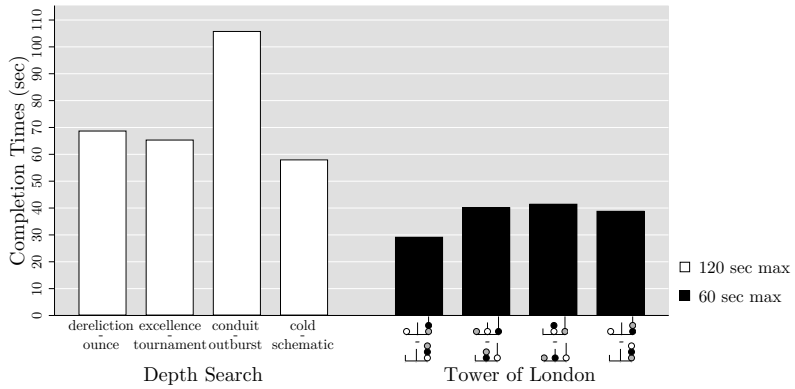
	Depth Search			Tower of London			
	Mean	Min	Max	Mean	Min	Max	N
Completion time	74.88	12	120	37.64	8	60	800
Success	0.70	0	1	0.76	0	1	800
Shortest path	0.18	0	1	0.24	0	1	800
# of Moves	6.97	4	39	9.04	6	23	609

Note. The *Depth Search* task is presented in columns 2-4. The *ToL* is in columns 5-7. The shortest path is conditional on successfully completing a problem. All of the *Depth Search* paths had a minimum of four moves. For the Tower of London they had a minimum path length of 6 or 7 (see Table 5.2).

variety of performance metrics for each. In addition to *completion time* as our main measure for performance (top), we can also examine success rates and moves required. As Table 5.3 shows, on average, 70% (76%) of subjects successfully travelled from the start to the target word (position) within the given time limits for the *Depth Search* (*ToL*) task. It thus appears that by allowing subjects twice as much time for *Depth Search*, we achieved similar levels of difficulty between the two tasks. While the majority of subjects thus managed to complete the tasks, only few (18% and 24%) were able to find a shortest path.[28] This is also reflected in the average number of moves that are above the minimum possible (4 and 6 moves). Overall, we can see that again both tasks show very similar characteristics.

Let us now compare *completion time* as our key performance measure across tasks. In order to avoid losing a sizeable number of observations we score all subjects the maximum time if they did not succeed within the given time limit. As we are confident that all subjects could eventually complete the problems, this scoring is likely to simply provide a lower bound for

Figure 5.11: Completion times across problems in the *Depth Search* task and the *ToL*



Note. The bar chart plots all average completion times across problems. The maximum time for the *Depth Search* task was 2 minutes, for the *ToL* 1 minute.

[28] Note that this is significantly lower than what we observed in Chapter 4 (52%). This difference might be driven by changes in instructions. While, in Chapter 4, we asked subjects to solve the problems with fewest moves, we now explicitly encourage them to finish problems as fast as possible.

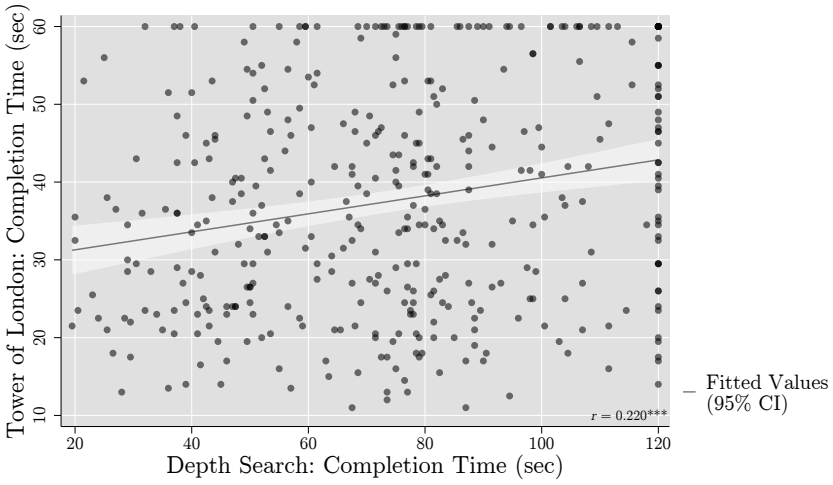
subjects that failed to do so.[29] Figure 5.11 demonstrates average completion times in seconds separately for each of the four problems. Overall, we see that performance is very consistent in the different problems, except for the path from 'conduit' to 'outburst'. Again the figure shows that on average subjects needed twice as long to complete the *Depth Search* task, justifying our 2:1 ratio for the time limits.

To formally test whether performance is correlated between tasks beyond these similarities, we again regress completion time in the *Depth Search* task on completion time in the *ToL*. Figure 5.12 visualises this relationship by plotting the *ToL* on the vertical and the *Depth Search* task on the horizontal axis and providing the Pearson's correlation coefficient in the bottom right corner. We find a smaller but still significant relationship between performances in both tasks ($r = 0.220$; p-value < 0.01). Thus, even though weaker than for the previous set of tasks, we find support that subjects performing well in *Depth Search* do indeed perform better in the *ToL*.[30] Again, this result is robust to the inclusion of additional controls (Appendix E, Table E.1.1), where, despite a reduction in effect sizes, the relationship remains marginally significant (p-value < 0.1). Moreover, we again find that younger subjects, and those with higher language proficiency perform better in our task.

Result 2. *There is a significant relationship in performance between Depth Search and the Tower of London.*

As above, Result 2 can be interpreted to support that both investigated tasks capture the same kind of associative thinking. Finally, when examining within-task robustness, we again

Figure 5.12: Relationship between the *Depth Search* task and the *Tower of London*



Note. The fitted line corresponds to a univariate regression between the *Depth Search* (horizontal) task and the *ToL* (vertical). Subjects that did not complete a problem are assigned the maximum time of 120 seconds for the *Depth Search* task and 60 seconds for the *ToL*. Pearson's correlation with significance is provided in the bottom right corner of the graph.

[29] Except stated otherwise, all reported results also hold for restricting the sample to subjects that finished the problem.

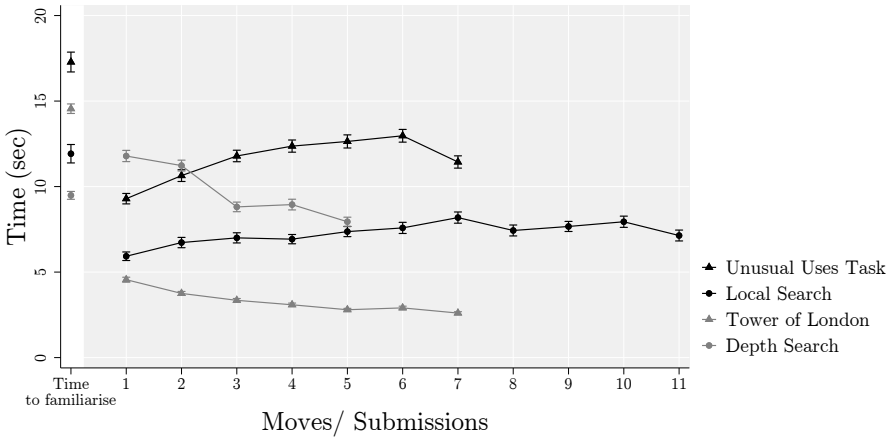
[30] When excluding unsuccessful attempts, the correlation drops to $r = 0.147$; p-value < 0.01.

find significant correlations across rounds for both tasks. As in the previous section, we can also apply the *ORIV* approach in this case to account for potential measurement error (Gillen et al., 2019). Once addressing noise in measurement, we find substantial increases in the correlation between the *Depth Search* task and the *ToL* ($r = 0.55$; p-value $= 0.026$). In fact, the *ORIV* coefficient is 2.5-times as large as the *raw* estimate, suggesting considerable measurement error in the individual elicitations. Moreover, this result also indicates that multiple elicitations of our *Depth Search* task can provide significantly more information than a single measure. Altogether, the ORIV approach therefore further strengthens our Result 2. In addition, all previous findings also hold, nonetheless, when examining each round independently (Appendix E, Table E.1.3).

4.3 Correlation across pairs of tasks

After exploring the pairs of tasks separately, we now compare behaviour across all tasks. An interesting difference between the pairs of the tasks are thereby move and submission timings. Moving away from task performance, we can compare behaviour in all four tasks by examining decision times. In particular, in all four tasks we measure the time for each individual to submit a new word/use in the *Local Search* task and the *UUT*, as well as the time it takes them to decide on a move in the *Depth Search* task and the *ToL*. Figure 5.13 illustrates how submissions and moves change over time in all four tasks. We plot the submission and move count on the horizontal axis, thus an observation at 2 captures the average time between the second and third move or submission of word/use. When examining task behaviour through the lens of decision timings, we can see that the times between submissions in the *Local Search* task and the *UUT* (black) increase over time. This seems intuitive, as it gets increasingly harder to think of new answers. In the *Depth Search* task and the *ToL* (grey), by contrast, we see a

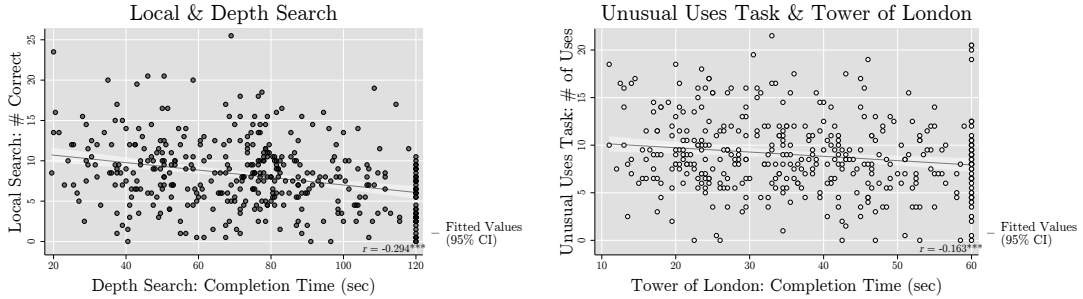
Figure 5.13: Submission time patterns across tasks



Note. Line graph of submission patterns for all tasks. Our set of tasks is represented by colour (black and grey), the two network and creativity tasks are indicated by marker (circle and triangle). The first observation is disconnected as subjects familiarised themselves with the interface. Observation numbers drop from left to right, as subjects provided fewer uses/words or reached the target. At the final point of each line there are around 65% of the observations left.

different pattern. The closer subjects get to the target word/position, the faster they make decisions about their next move. This evidence extends the analysis of performances from above and highlights similarities in behaviour within and differences between the pairs of tasks, strengthening the argument that they do capture distinct types of associative thinking.

Figure 5.14: Relationship within associative thinking and creativity tasks



Note. Scatter plots across sets of tasks. Network tasks on the left (*Local Search* vertical and *Depth Search* horizontal). Creativity tasks on the right (*UUT* vertical and *ToL* horizontal). Fitted line correspond to a univariate regression. Pearson's correlations are in the bottom right corners.

To test this more formally, we explore whether there is a significant association between the tasks. In particular, we regress performance in the *Depth Search* task on performance in the *Local Search* task and do the same for the *ToL* and *UUT*. Figure 5.14 depicts both relationships visually, with the comparison of our 'network tasks' in the left graph and the 'creativity tasks' on the right. Moreover, we again provide correlation coefficients in the bottom right corner of each graph. As we can see, in contrast to Hypothesis 3, we do find statistically significant correlations in both sets of tasks. Note that the negative sign is in line with intuition, as a better completion time is shorter, whereas a better fluency score is higher. The correlation is $r = -0.294$ (p-value < 0.01) for the network tasks (left) and $r = -0.16$ (p-value < 0.01) for the *UUT* and and *ToL* (right), indicating that the relationship is stronger in the case of our network tasks. This finding suggests that all tasks explored require some common underlying ability. Taken together with our findings in Chapter 4, the significant relationship between the 'creativity tasks' is the first time that we find a robust, even though relatively small, within-subject correlation in performance between two distinct creativity tasks taken 'off the shelf'. This underlines the need for further research exploring the inherent properties of experimental creativity tasks to develop a better understanding of their similarities and differences.

Result 3. *Performance in both associative thinking as well as creativity tasks correlates with each other. The relationship is stronger between the two network tasks.*

Lastly, as in the previous chapter we also explore the relationship between experimental tasks and several survey measures of creativity. In line with our previous findings, we find no statistically significant relationship between self-reports and any of the four tasks (Appendix able E.1.4).

Despite the promising findings with respect to our proposed network tasks, one potential aspect that could explain correlation across all tasks in addition to underlying creative ability

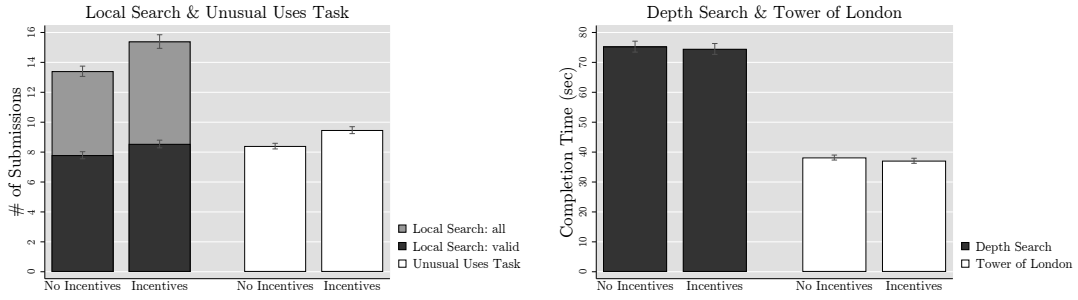
is motivation. If certain subjects have higher degrees of intrinsic motivation to complete the tasks and this motivation translates into performance, it could drive the relationships across all tasks. By varying incentives between subjects we cannot only address this concern, but also speak more broadly to the literature of incentives for creativity.

4.4 The role of incentives

Even though the main purpose of our benchmarking exercise is to validate our two proposed network tasks, we also investigate incentive effects across them. Recall that our tasks vary in terms of their level of *openness* and previous research has argued that this dimension plays a crucial role for the effectiveness of incentives. We therefore expect to see difference across tasks. In particular, with the *Depth Search* task and the *ToL* being relatively more *closed*, we expect stronger incentive effects for these compared to the other set of tasks.

Figure 5.15 shows average performances in all four tasks across incentivised and non-incentivised conditions. Contrasting our expectations, we can see that incentives only have a positive effect on performance in the *Local Search* task and the *UUT*. In both, subjects submit significantly more answers when being incentivised. This also translates into an increased number of *valid* words found in the *Local Search* task (black). The positive incentive effect in the *UUT* is also in line with previous findings by Bradler et al. (2019), who show that incentives matter for the *UUT*. With respect to the *Depth Search* task and the *ToL*, by contrast, we fail to observe any significant effect of incentives, even though it appears that there is a very marginal reduction in completion times in both tasks in the presence of incentives. This lack of incentive effects is in line with Eckartz et al. (2012), who find no effect of incentives in the very *closed* word task (utilised in Chapter 4). Our findings suggest that incentive effects across different levels of *openness* might be non-linear. While, Charness & Grieco (2019) find that *open* creativity is not enhanced by incentives, also very *closed* tasks as *Depth Search*, the *ToL* or the *word task* do not seem to respond to extrinsic incentives. It could thus be that only intermediate levels of *openness* allow for incentivisation. This is however only a possible explanation as it could also be that other features of the tasks, not considered by us, matter

Figure 5.15: Incentive effects across tasks



Note. Average performance (fluency left, completion time right) across all tasks. The incentivised and not-incentivised conditions are next to each other in each case. The *Local Search* task (left) is presented with all submissions (grey) and valid words (black). The error bars indicate standard errors.

for incentive effects. Thus, further research is required to explore the relationship between *openness* and incentives in more detail.

Result 4. *Incentives positively affect performance in Local Search and the Unusual Uses Task, but have no effect on Depth Search and the Tower of London.*

To revert back to the question of motivation as a possible confound in our previous analyses, we can replicate or results 1-3 also when solely examining the incentivised condition (Appendix E, Table E.1.5). We therefore, conclude that motivation is not the underlying reason for the significant relationships found above. Finally, our finding that incentives work (or fail to work) in the same way within the benchmarking pairs of tasks further supports that they indeed share similar characteristics.

5 Conclusion

In this chapter we develop and characterise a novel method to measure associative thinking ability. We are convinced that associative thinking plays a key role in creativity, especially as a direct link to innovation. Our method relies on an ex-ante defined solution space in form of a semantic network. This improves on existing tasks by removing any ambiguity in terms of evaluating or scoring performances. Within this network we propose two distinct associative thinking tasks: *Local Search* and *Depth Search*. While the former captures the ability to generate as many *valid* associations as possible, the latter requires the identification of *most relevant* associations in order to reach a pre-defined target. We then benchmark both tasks against established measures of creativity: the *Unusual Uses Task* and the *Tower of London*. We find significant relationships between our proposed tasks and each of the benchmarking tasks in a number of ways. First, performance in each pair of tasks is significantly correlated. Secondly, we show that behaviour as identified through decision times is identical in each pair of tasks, but different between the pairs of tasks. Lastly, we also find that incentives affect each pair of tasks identically, supporting their similarities even further. The accumulation of these factors supports our hypothesis that the same ability is required to perform well in the *Local Search* task and the *UUT* one the one hand, and in the *Depth Search* task and the *ToL* on the other. We therefore offer an experimental method that captures features of established creativity tasks whilst significantly improving on their experimental properties, providing researchers with a tool to explore creative associative thinking in various contexts and environments. Lastly, when exploring incentive effects across tasks our results suggest that a certain degree of task *closeness* might be required for positive incentive effects while incentives do not affect tasks that are very *closed*.

Throughout Part II of this thesis we explored the relationship between different experimental tasks that are designed to measure creativity. After identifying a lack of correlation across different domains and degrees of openness we limited our focus to a specific type of creative ability, associative thinking. For this specific domain we then developed and tested two new tasks that capture the same underlying abilities as two very established creativity tasks while improving significantly on their experimental properties. The next step in the exploration of our developed method is to study the role of individual associative thinking ability

in a variety of economic and non-economic environments. The following section therefore provides a brief overview of three examples of how I plan to utilise our instrument in future research. The examples are not intended as independent research proposals but instead provide a high-level overview of the adaptability of our experimental tool.

Bibliography

Acar, S. & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal*, 26(2), 229–238.

Attanasi, G., Chessa, M., & Llerena, P. (2021). A survey on experimental elicitation of creativity in economics.

Batey, M. & Furnham, A. (2006). Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, Social, and General Psychology Monographs*, 132(4), 355–429.

Beaty, R. E. & Silvia, P. J. (2012). Why do ideas get more creative across time? an executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, 6(4), 309.

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition*, 42(7), 1186–1197.

Benedek, M., Könen, T., & Neubauer, A. C. (2012). Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3), 273.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7), 322–328.

Bradler, C., Neckermann, S., & Warnke, A. J. (2019). Incentivizing creativity: A large-scale experiment with performance bonuses and gifts. *Journal of Labor Economics*, 37(3), 793–851.

Charness, G. & Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17(2), 454–496.

Eckartz, K., Kirchkamp, O., & Schunk, D. (2012). How do incentives affect creativity? Jena Economic Research Papers, no. 068.

Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). Lioness lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1), 95–111.

Gillen, B., Snowberg, E., & Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4), 1826–1863.

Gough, H. G. (1976). Studying creativity by means of word association tests. *Journal of Applied Psychology*, 61(3), 348.

Hittmár, Š., Varmus, M., & Lendel, V. (2014). Proposal of model for effective implementation of innovation strategy to business. *Procedia-Social and Behavioral Sciences*, 109, 1194–1198.

Jabri, M. M. (1991). The development of conceptually independent subscales in the measurement of modes of problem solving. *Educational and Psychological Measurement*, 51(4), 975–983.

Jobs, S. (1996). Steve jobs: The next insanely great thing. Wired. Retrieved from https://www.wired.com/1996/02/jobs-2/.

Johansson, F. (2004). *The medici effect.* Penerbit Serambi.

Lee, C. S. & Therriault, D. J. (2013). The cognitive underpinnings of creative thought: A latent variable analysis exploring the roles of intelligence and working memory in three creative thinking processes. *Intelligence*, 41(5), 306–320.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69(3), 220.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1), 60–67.

O'Connor, G. C. & McDermott, C. M. (2004). The human side of radical innovation. *Journal of Engineering and Technology Management*, 21(1-2), 11–30.

Runco, M. A., Plucker, J. A., & Lim, W. (2001). Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal*, 13(3-4), 393–400.

Scott, S. G. & Bruce, R. A. (1994). Determinants of innovative behavior: A path model of individual innovation in the workplace. *Academy of Management Journal*, 37(3), 580–607.

Verhaeghen, P., Trani, A. N., & Aikman, S. N. (2017). On being found: How habitual patterns of thought influence creative interest, behavior, and ability. *Creativity Research Journal*, 29(1), 1–9.

Whitman, R. D., Holcomb, E., & Zanes, J. (2010). Hemispheric collaboration in creative subjects: Cross-hemisphere priming in a lexical decision task. *Creativity Research Journal*, 22(2), 109–118.

Applications

In the previous two chapters, I presented methodological contributions to experimental creativity research. More precisely, after identifying shortcomings of established creativity tasks in terms of conceptual approaches and methodological properties in Chapter 4, I describe the development a novel instrument to elicit creative associative thinking in Chapter 5 finding substantive evidence that the developed tasks capture abilities necessary for established creativity tasks, whilst improving on their implementation and performance evaluation.

An obvious next step to consider is how to apply the developed instrument for further research. On the one hand the network tasks provide a creative environment that can be used to study, for instance, the effect of different incentives. On the other hand, I consider the method a diagnostic instrument. This implies that it is a tool to elicit individual associative thinking ability, which can then be related to other individual characteristics and behaviour in a variety of different contexts. In the next paragraphs, I will provide a description of three possible applications of the instrument demonstrating the variability and validity of the developed tool. The examples outlined below are a non-exhaustive list of planned applications in my subsequent research. Each example serves as a high-level overview of a research project and is not intended as an outline of a self-contained research proposal.

Firstly, I intend to utilise the proposed network task as a creative environment in itself. With small adjustments, our network task allows to combine the *Local Search* and *Depth Search* task into a unique setting. This can be achieved by asking, subjects to find a path from a start to a target word as in the *Depth Search* task, but now without providing them with all directly associated words. Subjects thus have to engage in the *Local Search* task at every word to identify a next relevant step with the ultimate goal to reach the target word. When combining the tasks as described, the environment closely resembles innovation, where each step towards the target resembles a step in an innovation process.[1] Once thinking of our environment as an innovative process, I can then test economic aspects considered as crucial for innovation. More precisely, I can ask subjects to engage in an innovation competition by asking multiple subjects to simultaneously complete the same path in a network. The network structure then enables me to explore the role of *patenting* by allowing subjects to patent certain words and obtaining exclusive access of these, generating obstacles for other subjects to identify a path to the target word.[2] The task thus lends itself very naturally for experimental investigations of innovations and in particular for group or competitive settings.

Contrasting an application where the network environment provides the creative setting, the developed method can also be employed as a diagnostic tool to elicit individual level creative ability. The following two applications thus utilise the tool to explore the impact of individual level associative thinking ability on other dimensions: creative collaboration and strategic sophistication.

[1] See Verworn & Herstatt (2002) for an overview of innovation process models.
[2] See Boldrin & Levine (2013) for a review of the impact of patents on innovation.

With respect to collaborative collaboration, I combine two topics from this thesis by focusing in particular on the role of individual associative thinking ability (Chapter 5) and team cohesion (Chapter 2) with the goal to explore team composition in collaborative creativity. While high individual creative ability could be of particular importance for collaborative creativity, the latter might need to be complemented by other team members with good social or project management skills to provide optimal output. Exploring specific individual characteristics that impact successful collaborative creativity can significantly enhance our understanding of which dimensions are essential for building such a team. Moreover, on top of team composition, cohesion is also likely to be a key ingredient to explain successful creative ideation. Previous research has shown that the ability of group members to better coordinate amongst each other directly translates into improved team performance in non-creative tasks, which I also expect to translate to creative collaboration. (Gächter et al., 2019). However, while team cohesion may lower team coordination costs, due to inherent homophily in social networks (e.g. McPherson et al., 2001), it could reduce team diversity and, thus, the team's creativity (e.g. Uzzi et al., 2013; Schimmelpfennig et al., 2022). In my view, studying team composition as a function of individual and team characteristics and relate it to collaborative creativity can thus provide substantive insights into the assembly of creative teams. Moreover, the developed associative thinking tasks thereby serve as ideal tools to elicit individual creative ability.

I consider the exploration of collaborative creativity a more extensive research agenda that can transition from the laboratory to the field via three steps. First, in line with the development of our network instrument, I plan to examine collaborative creativity within a laboratory setting. Following this, a potential step into a more complex creative environment is to examine collaborative creativity in escape room games as in Englmaier et al. (2021). Escape rooms provide complex creative environments that challenge teams not only in their creative problem-solving ability, but also their ability to find systematic approaches and work together as a team. They thus provide a semi-controlled environment to study how creative ability, as elicited with the developed network tasks affects performances. Ultimately, the final step is to explore whether the findings with ad-hoc teams in the laboratory and escape rooms translate to already established creativity teams. I therefore plan to relate individual creative ability as elicited with the developed instrument to creative collaboration of 'research and development' teams in firms. Using firms collected information on the performance of teams and relating these to identifiable individual and team characteristics can provide hands-on insights into the selection of creativity teams.

Lastly, our network instrument can also be applied to a question that has been at the heart of creativity research. Within psychology there is an on-going debate regarding the relationship between creativity and intelligence. I plan to contribute to this debate by testing whether creative ability is related to strategic sophistication, which has been shown to be linked to intelligence (Carpenter et al., 2013). This in turn could contribute to our understanding of heterogeneity in strategic sophistication between individuals. In particular, I want to explore whether associative thinking ability as identified in the *Depth Search* task predicts equilibrium behaviour in strategic games. The reason I expect that the two abilities could be related is that the *Depth Search* task challenges subjects to project and identify optimal next steps to

reach their target. A possible strategy to solve this task is to consider the target word and work backwards by considering possible associations close to the target word. This strategy appears similar to backward induction as required in strategic games, suggesting that individuals who perform better in the *Depth Search* task are also more likely to exhibit equilibrium behaviour in environments requiring strategic sophistication.

The three applications outlined above provide examples of application how I plan to utilise the developed instrument in order to integrate creativity with established economic research questions. In addition, the examples highlight the diversity of the method and demonstrate how it can be applied and modified for different research questions. Ultimately, I consider the development of the network instrument a door opener for a larger research agenda providing experimental scientists with the opportunity to study creativity in a variety of economic and non-economic settings.

Bibliography

Boldrin, M. & Levine, D. K. (2013). The case against patents. *Journal of Economic Perspectives*, 27(1), 3–22.

Carpenter, J., Graham, M., & Wolf, J. (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior*, 80, 115–130.

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., & Schudy, S. (2021). The efficacy of tournaments for non-routine team tasks.

Gächter, S., Starmer, C., & Tufano, F. (2019). *The surprising capacity of the company you keep: revealing group cohesion as a powerful factor of team production.* Technical report, CeDEx Discussion Paper Series.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444.

Schimmelpfennig, R., Razek, L., Schnell, E., & Muthukrishna, M. (2022). Paradox of diversity in the collective brain. *Philosophical Transactions of the Royal Society B*, 377(1843), 20200316.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472.

Verworn, B. & Herstatt, C. (2002). *The innovation process: an introduction to process models.* Technical report, Working paper.

Appendix for Chapter 1

A.1 - Additional analyses

[**All subsequent analyses are referenced in the document. They are structured by section within each chapter. Brief descriptions are provided above every analysis.**]

A.1.1 - Finding an optimal design

Figure A.1.1 plots the *mean* euclidean distance of simulations across different sub-sample sizes. We can see high levels of inaccuracies for ρ estimations. The are due to outliers as described in the text.

Figure A.1.1: Mean euclidean distance for different subsample sizes

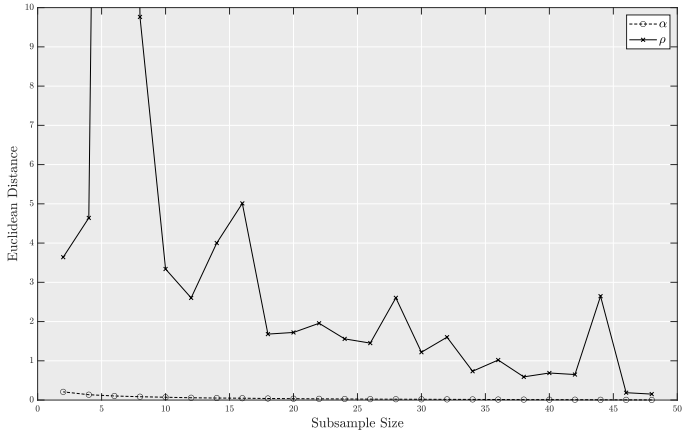


Figure A.1.2 plots the *mean* euclidean distance of simulations across different sub-sample sizes excluding values $\rho = 1$ and $\rho < -1$. The outliers from above disappear and we observe much higher levels of accuracy.

Figure A.1.2: Mean euclidean distance for different subsample sizes ($-1 \leq \rho < 1$)

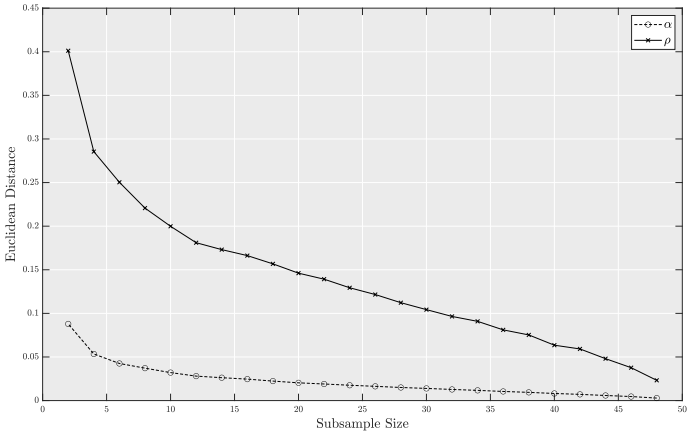
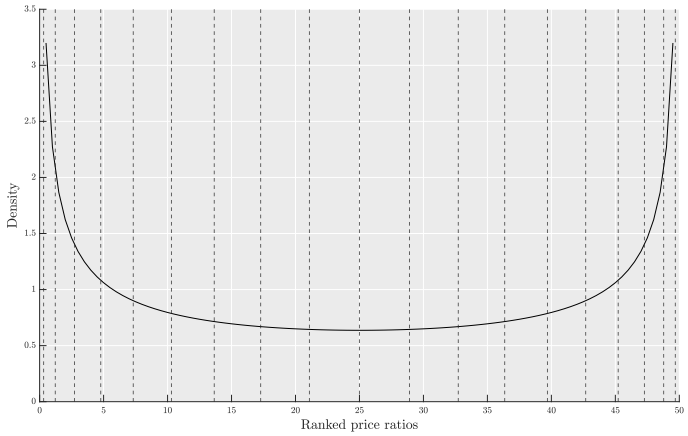


Figure A.1.3 shows how to implement the distributed bimodal draw. The horizontal describes all price ratios from 0.1 to 10. The dotted lines correspond to 5% of a β-distribution with $\alpha = \beta = 0.5$. For the implementation we pick one random price ratio from each of the 20 bins in the figure.

Figure A.1.3: β-distribution ($\alpha = 0.5$,$\beta = 0.5$), 20 bins



The individual bins are [0.1000,0.1110], [0.1110,0.1440], [0.1440,0.1980], [0.1980,0.2718], [0.2718,0.3635], [0.3635,0.4709], [0.4709,0.5913], [0.5913,0.7218], [0.7218,0.8591], [0.8591,1.0000], [1.0000,1.1640], [1.1640,1.3854], [1.3854,1.6912],[1.6912,2.1236], [2.1236,2.7510], [2.7510,3.6792], [3.6792,5.0505], [5.0505,6.9444], [6.9444,9.0090], [9.0090,10.0000].

A.1.2 - Validating with the *ALP* data

Figure A.1.4 plots mean euclidean distance across sub-sample sizes for the ALP data. As in Figure A.1.1, we find substantial outliers for estimations of ρ.

Figure A.1.4: Mean euclidean distance for different subsample sizes - ALP Data

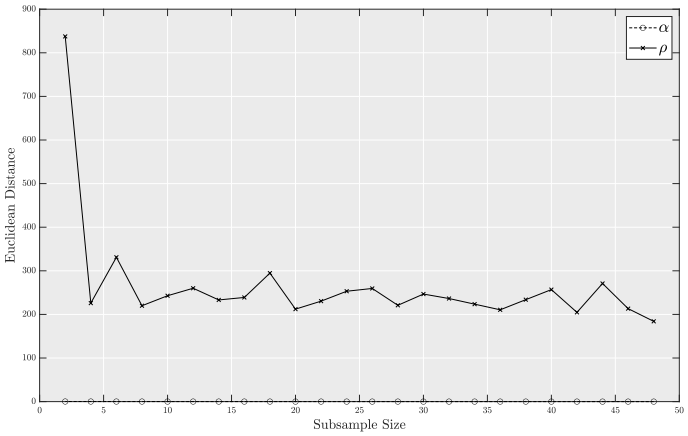


Table A.1.1 presents part of the regression models for demographic variables explaining estimated parameters. For brevity we do not provide the entire regression models (63 coefficients). All omitted variables do not have predictive power and are not commented on in the text.

Table A.1.1: Regression of demographics (ALP)

	α		ρ	
	(1) Original	(2) Condensed	(3) Original	(4) Condensed
age	0.00108	0.000551	0.0000508	0.000607
	(0.000776)	(0.000621)	(0.00175)	(0.00188)
Female	-0.0535***	-0.0388**	0.0110	0.0257
	(0.0202)	(0.0171)	(0.0504)	(0.0516)
$5,000 to $7,499	0.175*	0.164**	0.126	-0.0313
	(0.100)	(0.0813)	(0.317)	(0.384)
$7,500 to $9,999	0.0551	0.0550	0.354**	0.186
	(0.0721)	(0.0592)	(0.170)	(0.154)
$10,000 to $12,499	0.139*	0.139**	0.113	-0.00832
	(0.0718)	(0.0598)	(0.174)	(0.152)
$12,500 to $14,999	0.176**	0.146***	-0.105	-0.110
	(0.0696)	(0.0551)	(0.210)	(0.177)
$15,000 to $19,999	0.0575	0.0815	-0.177	-0.200
	(0.0674)	(0.0507)	(0.179)	(0.167)
$20,000 to $24,999	0.0600	0.0987**	0.0400	-0.0670
	(0.0635)	(0.0449)	(0.166)	(0.151)
$25,000 to $29,999	0.0328	0.0712	0.0549	0.0589
	(0.0672)	(0.0496)	(0.176)	(0.144)
$30,000 to $34,999	0.0824	0.0996**	-0.0731	-0.246*
	(0.0587)	(0.0433)	(0.164)	(0.145)
$35,000 to $39,999	0.0481	0.0699	-0.172	-0.271*
	(0.0711)	(0.0536)	(0.185)	(0.164)
$40,000 to $49,999	0.0984*	0.116***	-0.0247	-0.112
	(0.0587)	(0.0438)	(0.166)	(0.143)
$50,000 to $59,999	0.0938	0.119***	-0.147	-0.196
	(0.0596)	(0.0425)	(0.165)	(0.148)
$60,000 to $74,999	0.143**	0.152***	-0.138	-0.215
	(0.0599)	(0.0446)	(0.176)	(0.155)
$75,000-$99,999	0.0727	0.0946**	0.0525	-0.00710
	(0.0574)	(0.0416)	(0.164)	(0.138)
$100,000-$124,999	0.0362	0.0544	0.0760	0.0201
	(0.0614)	(0.0467)	(0.172)	(0.149)
$125,000-$199,999	0.0753	0.0978**	-0.0141	-0.200
	(0.0636)	(0.0477)	(0.172)	(0.149)
$200,000 or more	0.0692	0.128**	0.253	0.132
	(0.0835)	(0.0632)	(0.228)	(0.210)
Observations	786	786	544	533
R^2	0.075	0.077	0.158	0.149

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. We exclude most demographics from the table as they do not provide meaningful results. Demographic information excluded is *born in the US, current living situation, work status, type of work, education level*, and *ethnicity*. To obtain the full regression model, please contact the author.

A.1.3 - Own data collection

Table A.1.2 shows Kolmogorov-Smirnov tests to indicate no significant difference between our *Control* and *Treatment 20 + 30* when pooling the data across collections. The distributions are identical for the robust range of $-1 \leq \rho < 1$.

Table A.1.2: Pooled - Overview of Treatment comparisons

Comparison	α	ρ	$-1 \leq \rho < 1$
Control vs. Treatment 20 + 30	0.383	0.017^{**}	0.270
Control vs. Treatment 20 + 30	0.383	0.013^{**}	0.195

Note. All values reported are p-values of Kolmogorov-Smirnov tests. Once restricting our sample, there is no difference in treatments.

A.2 - Experimental instructions

[Instructions were encountered by all subjects except when stated otherwise. Two dotted lines represent a new screen in the programme. All programmes are available upon request. Comments for the reader are included as [...] and were not seen by the subjects.]

Welcome to this HIT!

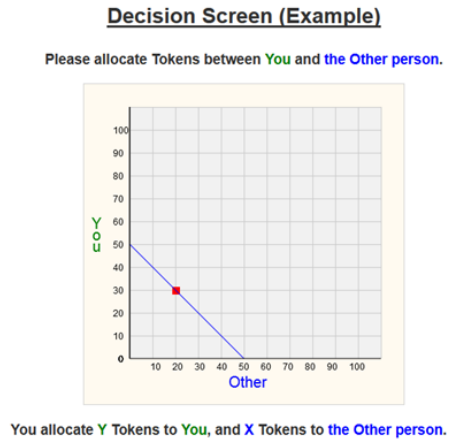
In this study you will be paid **$1.00 and a bonus** for completing this study. Please read all instructions carefully, answer the comprehension questions and complete the associated tasks.

Please click 'Continue' to proceed

Instructions

Please read the instructions carefully.

In this experiment, you will participate in 50 independent decision situations that share a common form. Each decision situation will be presented on a screen like the **example screen** below.



In each decision situation, you will be asked to allocate Tokens between You and Other. Other is another participant, randomly chosen in each decision situation. Each choice will involve choosing a point on a line representing possible Token allocations.

To choose an allocation, use the mouse: Click or drag to move the pointer on the computer screen to the allocation that you wish to choose. Once you use the mouse, you can also use the arrows on the keyboard to move the pointer. The computer will only allow you to choose allocations that are on the line.

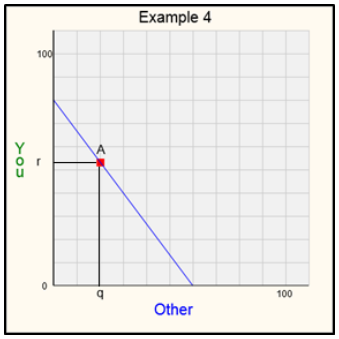
After that, confirm your decision by clicking on the Submit button. Once you have clicked the Submit button, your decision cannot be revised.

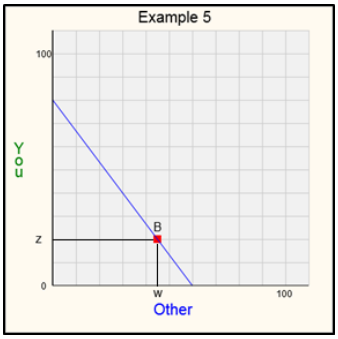
Next, you will be asked to make a decision in another independent situation. This process will be repeated until all 50 **[OR 20]** decision situations are completed.

Each decision situation will start by having the computer select a line randomly from the set of lines that (i) intersect with at least one of the axes at 50 or more Tokens and (ii) have no intercept above 100 Tokens or below 0 Tokens. Examples of lines you might face are shown in Examples 2 and 3.

If the line is **relatively flat** (Example 2), Other's Token increase by **more than 1 Token** as you decrease Your Token by 1 Token. If the line is **relatively steep** (Example 3), Other's Token increase by **less than 1 Token** as you decrease Your Token by 1 Token.

In each decision, you may choose any point that is on the line. For example, as illustrated in Example 4, selecting allocation A represents a decision to allocate r Tokens to You and q Tokens to Other. Similarly, in Example 5 selecting allocation B represents a decision to allocate z Tokens to You and w Tokens to Other.





Your earnings

At the end of the study, you will be paid $1 as a participation fee. In addition, you will be paid a bonus based on decisions made during the experiment.

For your bonus, **one out of the 50 [OR 20] decision situations will be randomly selected for each participant**. Your bonus will depend on the Tokens allocated in the selected decision situation. The Tokens will be converted to cash at the rate of 25 Tokens = $1.

For this randomly chosen decision situation, you will receive the Tokens you allocated to You. Another person who was matched with you in your decision situation will receive the Tokens you allocated to Other. In addition, there will be another participant in this experiment who was matched with you for the purpose of their own allocation decision. You will receive the Tokens this person allocated to Other. Please note this person is not the same person that you allocated Tokens to.

You will also be able to **practice the task** and familiarise yourself with the decision screen.

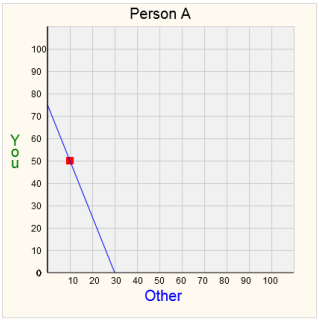
Please **click continue to answer the comprehension questions.**

Comprehension Questions

Please answer all questions.

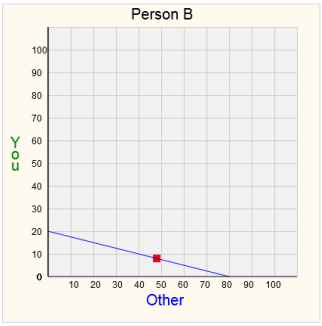
They serve to check your understanding of the Task. To review the instructions please click the botton on the right

---



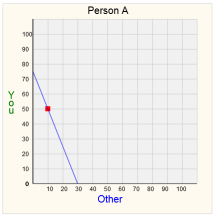
You allocate 50.0 Tokens to You, and 10.0 Tokens to the Other person.

- Q1a) How many Tokens has Person A allocated **to him/herself**? [**Numeric entry**]
- Q1b) How many Tokens has Person A allocated **to Other**? [**Numeric entry**]

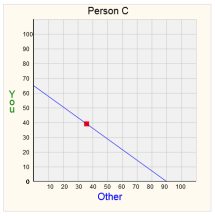


You allocate 8.0 Tokens to You, and 48.0 Tokens to the Other person.

- Q2a) How many Tokens has Person B allocated **to him/herself**? [**Numeric entry**]
- Q2b) How many Tokens has Person B allocated **to Other**? [**Numeric entry**]



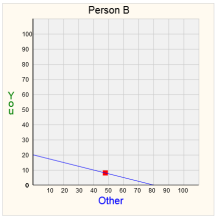
You allocate 50.0 Tokens to You, and 10.0 Tokens to the Other person.



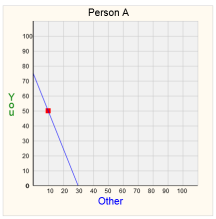
You allocate 39.0 Tokens to You, and 36.0 Tokens to the Other person.

Q3) Suppose that the number of Tokens Person C allocated to Other will be given Person A. How many Tokens **will Person A earn from her own and Person C's allocation?** (Multiple choice)

- 49 Tokens
- 89 Tokens
- 86 Tokens



You allocate 8.0 Tokens to You, and 48.0 Tokens to the Other person.



You allocate 50.0 Tokens to You, and 10.0 Tokens to the Other person.

Q4) Suppose that the number of Tokens Person A allocated to Other will be given Person B. How many Tokens **will Person B earn from her own and Person A's allocation?** (Multiple choice)

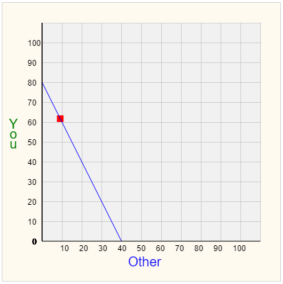
- 18 Tokens
- 58 Tokens
- 98 Tokens

On the next page you can practice the task. Please click continue to move on.

Decision Screen (Practice)

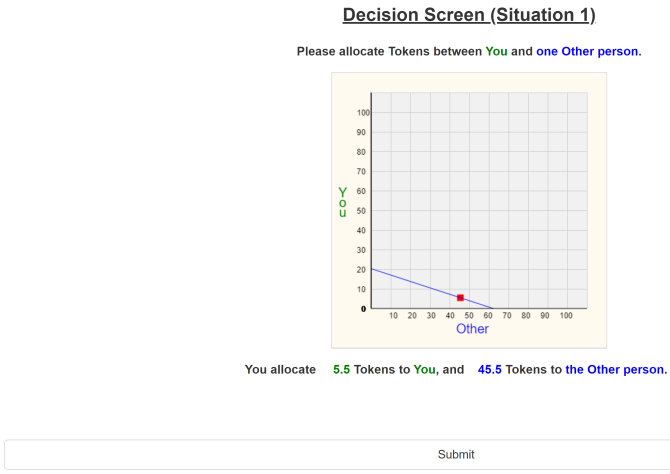
Below is the practice to help you familiarise with Task. Note that your decision in this practice **will not affect** your earnings.

Please allocate Tokens between You and the Other person.



You allocate 61.7 Tokens to You, and 9.1 Tokens to the Other person.

Submit



You completed this task. Please click continue to move on.

Just a couple of short questions to wrap up.

- What is your age? [**Numeric input**]
- What is your gender?
  - Female
  - Male
  - Other
  - Prefer not to say
- Which of the following describes you best?
  - Asian or Pacific Islander
  - Black or African American
  - Hispanic or Latino
  - Native American or Alaskan Native
  - White or Caucasian
  - Multiracial or Biracial
  - A race/ethnicity not listed here
  - Prefer not to say
- What is your highest level of education?
  - No formal education
  - High school diploma
  - College degree
  - Vocational training

  - Bachelor's degree
  - Master's degree
  - Professional degree
  - Doctorate degree
  - Other
  - Prefer not to say
- What is your work status?
  - Employed
  - Self-Employed/Freelance
  - Interning
  - Part-time
  - Full-time
  - Unemployed - Looking for work
  - Unemployed - Not looking for work
  - Homemaker
  - Studying
  - Military/Forces
  - Retired
  - Not able to work
  - Other
  - Prefer not to say
- Some people talk about 'left', 'right' and 'centre' to describe parties and politicians. With this in mind, where would you place yourself? (7-item Likert)
- Any feedback from your side regarding this study is highly appreciated. Please leave comments below. (Text input)

Your earnings

Your total earnings are **$1** + **a bonus**.

The bonus depends on one of your choices and a choice by one other participant.

To receive your earnings, please enter this code into MTurk

Random MTurk ID

After you have done that, you can close this window. We thank you for participating in our study.

Appendix for Chapter 2

B.1 - Experimental instructions

[**Instructions were encountered by all subjects except when stated otherwise. Here, we report the full set of questions we used. The questionnaire was implemented using the survey software Qualtrics (www.qualtrics.com) In the implementation, except for the introductory/background questions, we randomised the order of the other blocks that measure relationship closeness (IOS, WE, SCI, PAM, RCI, Love, and Like). [X] represents the named initial in all questions. All programmes are available upon request.**]

The questions relate to several scales, for an overview:

- Q1-Q6 and Q155-Q156 are introductory/background questions;
- Q7 refers to the "Inclusion of the Other in the Self" (IOS) scale (Aron et al., 1992) **OR** our extended IOS scale
- Q8 refers to the "We" scale (Cialdini et al., 1997)
- Q9 and Q10 refer to the "Subjective Closeness Index" (SCI) (Berscheid et al., 1989)
- Q11-Q25 refer to the "Personal Acquaintance Measure" (PAM) (Starzyk et al., 2006)
- Q26-Q64 refer to the "Relationship Closeness Inventory" (RCI) (Berscheid et al., 1989)
- Q65-Q77 and Q78-Q90 refer respectively to Loving and Liking scales (Rubin, 1970)
- Q91-Q130 refer to the "Balanced Inventory of Desirable Reporting" (BIDR) (Paulhus & Reid, 1991)
- Q131-Q150 refer to the "Mini-IPIP scales: Measure of the Big 5" (Mini IPIP) (Donnellan et al., 2006)
- Q151-Q152 refer to "Self-reported altruism" (Falk et al., 2018)
- Q153 refers to the "Inclusion of the Other in the Self" (IOS) scale (Aron et al., 1992) **OR** our extended IOS scale *for a stranger*
- Q154 refers to the "We" scale (Cialdini et al., 1997) *for a stranger*

Thank you for participating in our Study!

In this HIT we will ask you to respond to a questionnaire on the nature of interpersonal relationships.

Our interest is entirely scientific. All answers will be treated confidentially and will only be reported in aggregated statistical form.

There are no right or wrong answers in this survey, we are only interested in your honest assessment. If you feel uncomfortable answering some questions you will have opportunities to select "prefer not to answer" as an answer.

Please enter your Prolific ID here: [**Text input**]

We are currently investigating the nature of interpersonal relationships. As part of this study, we would like you to answer the following questions about your relationship with another person.

[*Depending on the 'relationship level' treatment subjects saw one of the following*]

Specifically, we would like you to choose the one person with whom you have the closest, deepest, most involved, and most intimate relationship, and answer the following questions with regard to this particular person. For some of you, this person may be a dating partner or someone with whom you have a romantic relationship. For others of you, this person may be a close, personal friend, family member, or companion. It makes no difference exactly who this person is as long as she or he is the one person with whom you have the closest, deepest, most involved, and most intimate relationship.

OR

Specifically, we would like you to choose a person with whom you have a good friendship, who is more than an acquaintance, but not your closest, or most intimate relationship, and answer the following questions with regard to this particular person. For some of you, this person may be a personal friend. For others of you, this person may be a family member, or companion. It makes no difference exactly who this person is as long as she or he is a good friend, who is more than an acquaintance, but not your closest, or most intimate relationship.

OR

Specifically, we would like you to choose a person whom you consider an acquaintance, but no more than an acquaintance, and answer the following questions with regard to this particular person. For some of you, this person may be a colleague at work. For others of you, this person may be a neighbour, or member of your wider social network. It makes no difference exactly who this person is as long as she or he is a person who you consider an acquaintance, but no more than an acquaintance.

Please select this person carefully since this decision will affect the rest of this study. With this person in mind, please respond to the following questions.

[Q1] Who is this person? Please give the initial of the first name only. [**Text entry**]

This person will be referred to as [X] in all questions that follow.

.....

.....

[Q2] What is your gender?

- Female
- Male
- Prefer not to say

[Q3] What is [X]'s gender?

- Female

- Male
- Prefer not to say

[Q4] What is your age? [**Numeric input**]

[Q5] How long have you known [X]? Please indicate the number of years and months.

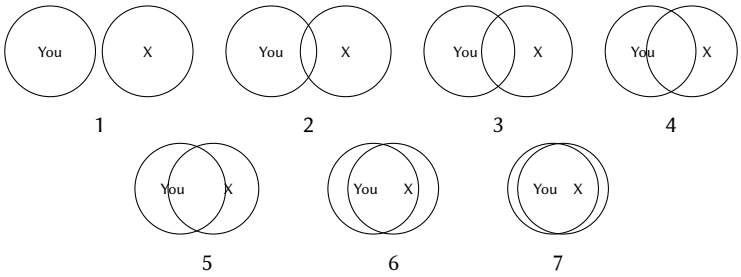
- Years [**Numeric input**]
- Months [**Numeric input**]

[Q6]Which of the following best describes your relationship with [X]? (Check only one)

- WORK: co-worker
- WORK: your boss/supervisor
- WORK: your subordinate
- FAMILY: aunt/uncle
- FAMILY: sister/brother
- FAMILY: parent
- FAMILY: cousin
- ROMANTIC: married
- ROMANTIC: engaged
- ROMANTIC: living together
- ROMANTIC: dating only this person
- ROMANTIC: dating this person and others
- FRIEND: close friend (non-romantic)
- FRIEND: casual friend
- ACQUAINTANCE (please give short description)
- OTHER (please give short description)
- prefer not to say

[*Depending on the treatment allocation subjects saw <u>one</u> of the following two IOS scales*]

[Q7a] In the following figure we ask you to consider which of these pairs of circles best describes your relationship with this person (referred to as [X] in all questions that follow). In the figure "X" serves as a placeholder for X, that is, you should think of "X" being [X]. By selecting the appropriate number please indicate to what extent you and [X] are connected. [**Numeric input**]



OR

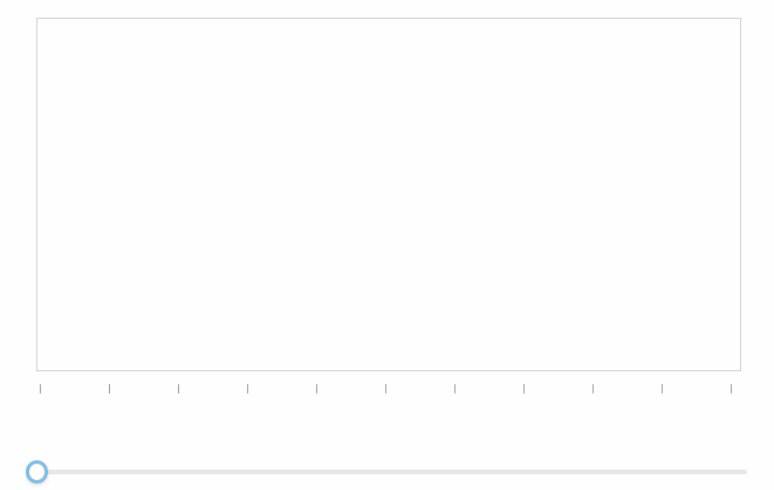
[Q7b] Once you move the slider below, a pair of circles will appear in the box. The position of the slider will determine the extent to which the circles overlap. When the slider is all the way to the left, the circles will look like this 

. When the slider is near the middle, the circles look like this 

. With it all the way to the right the circles look like this 

. You should interpret the degree of overlap as representing the relationship between you and [X]

Please position the slider so that the circles indicate to what extent you and [X] are connected.



.....

.....

[Q8] Please, select the appropriate number below to indicate to what extent you would use the term "WE" to characterize you and [X]. [**8-item Likert scale; 1 - Not at all; 7 - very much so; 8 - prefer not to answer**]

.....

.....

[Q9] Relative to all your other relationships (both same and opposite sex) how would you characterize your relationship with [X]? [**8-item Likert scale; 1 - Not close at all; 7 - very close; 8 - prefer not to answer**]

[Q10] Relative to what you know about other people's close relationships, how would you characterize your relationship with [X]? [**8-item Likert scale; 1 - Not close at all; 7 - very close; 8 - prefer not to answer**]

Please think of [X], read each statement carefully, and click the answer that best corresponds to your agreement or disagreement with each statement. [**For Q11 - Q25: 6-item Likert scale; 1 - Strongly Disagree; 2 - Disagree; 3 - Neither Agree nor Disagree; 4 - Agree; 5 - Strongly Agree; 6 - Prefer Not to Answer**]

[Q11] I have known [X] for many years.

[Q12] I have known [X] for a long time.

[Q13] I have gone to parties (social events) with [X].

[Q14] [X] often hides his/her true feelings from me.

[Q15] Seeing [X] is part of my weekly routine.

[Q16] I know what [X]'s goals are.

[Q17] [X] hides his/her true feelings from me.

[Q18] [X] has told me about his/her interests.

[Q19] I have spent time with [X] and his/her friends.

[Q20] [X] avoids showing his/her true feelings around me.

[Q21] [X] and I go way back.

[Q22] I am familiar with [X]'s friends.

[Q23] I see [X] a lot.

[Q24] Seeing [X] is part of my daily routine.

[Q25] [X] has told me what his/her goals are.

We would like you to estimate the amount of time you typically spend alone with [X] during the day. We would like you to make these time estimates by breaking the day into morning, afternoon, and evening, although you should interpret each of these time periods in terms of your own typical daily schedule. (For example, if you work a night shift, "morning" may actually reflect time in the afternoon, but is nevertheless time immediately after waking.) Think back over the past week and select the average amount of time, per day, that you spent alone with [X], with no one else around, during each time period. If you did not spend any time with [X] in some time periods, select 0 hour(s) and 0 minutes.

[Q26] DURING THE PAST WEEK, what is the average amount of time, per day, that you spent alone with [X] in the MORNING (e.g., between the time you wake and 12 noon)?

- Hours [**Numeric input**]
- Minutes [**Numeric input**]

[Q27] DURING THE PAST WEEK, what is the average amount of time, per day, that you spent alone with [X] in the AFTERNOON (e.g., between 12 noon and 6pm)?

- Hours [**Numeric input**]
- Minutes [**Numeric input**]

[Q28] DURING THE PAST WEEK, what is the average amount of time, per day, that you spent alone with [X] in the EVENING (e.g., between 6pm and bedtime)?

- Hours [**Numeric input**]
- Minutes [**Numeric input**]

[Q29] Compared with the "normal" amount you usually spend alone with [X], how typical was the past week?

- typical
- not typical

[Q30] The following is a list of different activities that people may engage in over the course of one week. For each of the activities listed, please check all of those that you have engaged in alone with [X] in the past week. Check only those activities that were done alone with [X] and not done with [X] in the presence of others.

In the past week, I did the following activities alone with [X] (Check all that apply)

- did laundry
- prepared a meal
- watched TV
- went to an auction/antique show
- attended a non-class lecture or presentation
- went to a restaurant
- went to a grocery store
- went for a walk/drive

- discussed things of a personal nature
- went to a museum/art show
- planned a party/social event
- attended class
- went on a trip (e.g., vacation or weekend)
- cleaned house/apartment
- went to church/religious function
- worked on homework
- spent time together on the internet (e.g., Skype, FaceTime, surfing together, etc)
- discussed things of a non-personal nature
- went to a clothing store
- talked on the phone
- went to a movie
- ate a meal
- participated in a sporting activity
- outdoor recreation (e.g., sailing)
- went to a play
- went to a bar
- visited family
- visited friends
- went to a department, book, hardware store, etc.
- played cards/board game
- attended a sporting event
- exercise (e.g., jogging, aerobics)
- went on an outing (e.g., picnic, beach, zoo, winter carnival)
- wilderness activity (e.g., hunting, hiking, fishing)
- went to a concert
- went dancing
- went to a party
- played music/sang
- other (please describe briefly)

.....

.....

The following questions concern the amount of influence [X] has on your thoughts, feelings, and behavior. Using the 7-point scale below, please indicate the extent to which you agree or disagree (from 1 - I strongly disagree to 7 - I strongly agree) [**For Q31 - Q57: 8-item Likert scale; 1 - I strongly disagree; 7 - I strongly agree; 8 - prefer not to answer**]

[Q31] [X] will influence my future financial security.

[Q32] [X] does not influence everyday things in my life.

[Q33] [X] influences important things in my life.

[Q34] [X] influences which parties and other social events I attend.

[Q35] [X] influences the extent to which I accept responsibilities in our relationship.

[Q36] [X] does not influence how much time I spend doing household work.

[Q37] [X] does not influence how I choose to spend my money.

[Q38] [X] influences the way I feel about myself.

[Q39] [X] does not influence my moods.

[Q40] [X] influences the basic values that I hold.

[Q41] [X] does not influence the opinions that I have of other important people in my life.

[Q42] [X] does not influence when I see, and the amount of time I spend with, my family.

[Q43] [X] influences when I see, and the amount of time I spend with, my friends.

[Q44] [X] does not influence which of my friends I see.

[Q45] [X] does not influence the type of career I have.

[Q46] [X] influences or will influence how much time I devote to my career.

[Q47] [X] does not influence my chance of getting a good job in the future.

[Q48] [X] influences the way I feel about the future.

[Q49] [X] does not have the capacity to influence how I act in various situations.

[Q50] [X] influences and contributes to my overall happiness.

[Q51] [X] does not influence my present financial security.

[Q52] [X] influences how I spend my free time.

[Q53] [X] influences when I see [X] and the amount of time the two of us spend together.

[Q54] [X] does not influence how I dress.

[Q55] [X] influences how I decorate my home (e.g., apartment, house, dorm room, ...).

[Q56] [X] does not influence where I live.

[Q57] [X] influences what I watch on TV.

Now we would like you to tell us how much [X] affects your future plans and goals. Using the 7-point scale below, please indicate the degree to which your future plans and goals are affected by [X] by clicking the appropriate scale. If an area does not apply to you (e.g., because you have no plans or goals in that area), click "1 - not at all". [**For Q58 - Q64: 8-item Likert scale; 1 - not at all; 7 - a great extent; 8 - prefer not to answer**]

[Q58] [X] affects my vacation plans.

[Q59] [X] affects my marriage plans.

[Q60] [X] affects my plans to have children.

[Q61] [X] affects my plans to make major investments (house, car, etc.).

[Q62] [X] affects my plans to join a club, social organization, church, etc.

[Q63] [X] affects my school-related plans.

[Q64] [X] affects my plans for achieving a particular financial standard of living.

Please, read each statement carefully, and click the answer that best corresponds to your agreement or disagreement with each statement.[**For Q65 - Q77: 10-item Likert scale; 1 - Not at all true; disagree completely; 5**

- Neither agree nor disagree; 9 - Definitely true; agree completely; 10 - Prefer not to answer]

[Q65] If [X] were feeling badly, my first duty would be to cheer him/her up.

[Q66] I feel that I can confide in [X] about virtually everything.

[Q67] I find it easy to ignore [X]'s faults.

[Q68] I would do almost anything for [X].

[Q69] I feel very possessive toward [X].

[Q70] If I could never be with [X] I would feel miserable.

[Q71] If I were lonely my first thought would be to seek [X] out.

[Q72] One of my primary concerns is [X]'s welfare.

[Q73] I would forgive [X] for practically anything.

[Q74] I feel responsible for [X]'s well-being.

[Q75] When I am with [X] I spend a good deal of time just looking at him (her).

[Q76] I would greatly enjoy being confided in by [X].

[Q77] I would be hard for me to get along without [X].

Please, read each statement carefully, and click the answer that best corresponds to your agreement or disagreement with each statement.[**For Q78 - Q90: 10-item Likert scale; 1 - Not at all true; disagree completely; 5 - Neither agree nor disagree; 9 - Definitely true; agree completely; 10 - Prefer not to answer**]

[Q78] When I am with [X] we are almost always in the same mood.

[Q79] I think that [X] is unusually well adjusted.

[Q80] I would highly recommend [X] for a responsible job.

[Q81] In my opinion, [X] is an exceptionally mature person.

[Q82] I have great confidence in [X]'s good judgment.

[Q83] Most people would react very favorably to [X] after a brief acquaintance.

[Q84] I think that [X] and I are quite similar to each other.

[Q85] I would vote for [X] in a group election.

[Q86] I think that [X] is one of those people who quickly wins respect.

[Q87] I feel that [X] is an extremely intelligent person.

[Q88] [X] is one of the most likable people I know.

[Q89] [X] is the sort of person whom I myself would like to be.

[Q90] It seems to me that it is very easy for [X] to gain admiration.

Using the scale as a guide, select a number beside each statement to indicate how much you agree with it. [**For Q91 - Q130: 8-item Likert scale; 1 - not true; 4 - somewhat true; 7 - very true; 8 - prefer not to answer**]

[Q91] My first impressions of people usually turn out to be right.

[Q92] It would be hard for me to break any of my bad habits.

[Q93] I don't care to know what other people really think of me.

[Q94] I have not always been honest with myself.

[Q95] I always know why I like things.

[Q96] When my emotions are aroused, it biases my thinking.

[Q97] Once I've made up my mind, other people can seldom change my opinion.

[Q98] I am not a safe driver when I exceed the speed limit.

[Q99] I am fully in control of my own fate.

[Q100] It's hard for me to shut off a disturbing thought.

[Q101] I never regret my decisions.

[Q102] I sometimes lose out on things because I can't make up my mind soon enough.

[Q103] The reason I vote is because my vote can make a difference.

[Q104] My parents were not always fair when they punished me.

[Q105] I am a completely rational person.

[Q106] I rarely appreciate criticism.

[Q107] I am very confident of my judgments.

[Q108] I have sometimes doubted my ability as a lover.

[Q109] It's all right with me if some people happen to dislike me.

[Q110] I don't always know the reasons why I do the things I do.

[Q111] I sometimes tell lies if I have to.

[Q112] I never cover up my mistakes.

[Q113] There have been occasions when I have taken advantage of someone.

[Q114] I never swear.

[Q115] I sometimes try to get even rather than forgive and forget.

[Q116] I always obey laws, even if I'm unlikely to get caught.

[Q117] I have said something bad about a friend behind his or her back.

[Q118] When I hear people talking privately, I avoid listening.

[Q119] I have received too much change from a salesperson without telling him or her.

[Q120] I always declare everything at customs.

[Q121] When I was young I sometimes stole things.

[Q122] I have never dropped litter on the street.

[Q123] I sometimes drive faster than the speed limit.

[Q124] I never read sexy books or magazines.

[Q125] I have done things that I don't tell other people about.

[Q126] I never take things that don't belong to me.

[Q127] I have taken sick-leave from work or school even though I wasn't really sick.

[Q128] I have never damaged a library book or store merchandise without reporting it.

[Q129] I have some pretty awful habits.

[Q130] I don't gossip about other people's business.

Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Indicate for each statement whether it is 1. Very Inaccurate, 2. Moderately Inaccurate, 3. Neither Accurate Nor Inaccurate, 4. Moderately Accurate, or 5. Very Accurate as a description of you.[**For Q131 - Q150: 6-item Likert scale; 1 - very inaccurate; 2 - moderately inaccurate; 3 - neither accurate nor inaccurate; 4 - moderately accurate; 5 - very accurate; 6 - prefer not to answer**]

[Q131] Am the life of the party.

[Q132] Sympathize with others' feelings.

[Q133] Get chores done right away.

[Q134] Have frequent mood swings.

[Q135] Have a vivid imagination.

[Q136] Don't talk a lot.

[Q137] Am not interested in other people's problems.

[Q138] Often forget to put things back in their proper place.

[Q139] Am relaxed most of the time.

[Q140] Am not interested in abstract ideas.

[Q141] Talk to a lot of different people at parties.

[Q142] Feel others' emotions.

[Q143] Like order.

[Q144] Get upset easily.

[Q145] Have difficulty understanding abstract ideas.

[Q146] Keep in the background.

[Q147] Am not really interested in others.

[Q148] Make a mess of things.

[Q149] Seldom feel blue.

[Q150] Do not have a good imagination.

.....

.....

[Q151] We now ask you for your willingness to act in a certain way. Please again indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so," and a 10 means "very willing to do so."

How willing are you to give to good causes without expecting anything in return? [**Likert scale**]

[Q152] Imagine the following situation: Today you unexpectedly received 1,600 pounds. How much of this amount would you donate to a good cause? [**Numeric input**]

.....

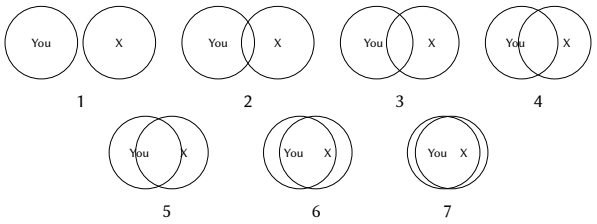
.....

[*Depending on the treatment allocation subjects saw <u>one</u> of the following two IOS scales for a stranger*]

Please note: All questions below refer to a stranger.

[Q153a] In the following figure we ask you to consider which of these pairs of circles best represents your relationship with a stranger. In the figure "X" serves as a placeholder for a stranger, that is, you should think of "X" being the stranger. By selecting the appropriate number please indicate to what extent you and a stranger are

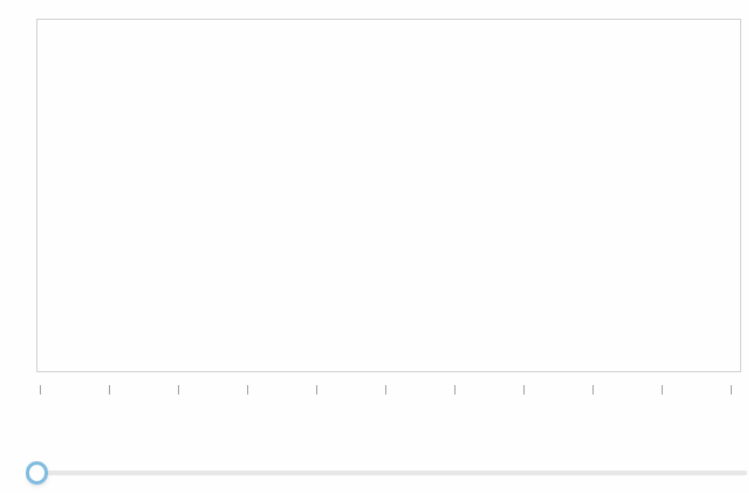
connected. [**Numeric input**]



OR

[Q153b] Once you move the slider below, a pair of circles will appear in the box. The position of the slider will determine the extent to which the circles overlap. When the slider is all the way to the left, the circles will look like this  . When the slider is near the middle, the circles look like this  . With it all the way to the right the circles look like this  . You should interpret the degree of overlap as representing the relationship between you and a stranger. In the figure "X" serves as a placeholder for a stranger, that is, you should think of "X" being the stranger.

Please position the slider so that the circles indicate to what extent you and a stranger are connected.



[Q154] Please, select the appropriate number below to indicate to what extent you would use the term “WE” to characterize you and a stranger. [**8-item Likert scale; 1 - Not at all; 7 - very much so; 8 - prefer not to answer**]

.....

.....

Thank you!

You’re almost done, just answer these two questions and the study is done.

[Q155] To what extent have you participated in other studies involving similar questionnaires on Prolific before taking this study? Take a guess if you are not sure. [**5-item Likert scale; 1 - never; 2 - 1-10; 3 - 11-20; 4 - 21-50; 5 - more than 50**]

[Q156] What is your nationality?

- UK
- Other [**Text input**]

Appendix for Chapter 3

C.1 - Additional analyses

[**All subsequent analyses are referenced in the document. Brief descriptions are provided above every analysis.**]

C.1.1 - Results - Social cohesion and altruistic giving

Table C.1.1 presents results of Wilcoxon Rank-Sum text to test altruistic giving across social distances. It confirms visual evidence from the text, we find significant differences between SD = 1 and the other two levels of social distance.

Table C.1.1: Wilcoxon rank-sum tests - Giving across social distances

Comparison	z statistic	p-value
SD = 1 vs. SD = 2	3.456	0.001
SD = 1 vs. Nameless	5.234	0.001
SD = 2 vs. Nameless	1.710	0.087

Note. Results from Wilcoxon Rank-Sum tests.

Table C.1.2 includes regression models examining IOS scores on Tokens passed. Models (2), (4) and (6) include controls. We cannot find a significant impact of IOS scores on tokens passed.

Table C.1.2: Altruistic giving by IOS Score

	Pooled		Anonymous		Non-Anonymous	
	(1)	(2)	(3)	(4)	(5)	(6)
IOS Score	0.269	0.157	0.283	0.135	0.259	0.190
	(0.198)	(0.222)	(0.258)	(0.285)	(0.210)	(0.246)
Exchange Rate (Reference: 3:1)						
1:1		-0.318		-0.508		-0.0964
		(0.789)		(0.921)		(1.050)
1:3		-0.0589		-0.635		0.624
		(1.517)		(1.636)		(2.008)
Controls	No	Yes	No	Yes	No	Yes
Observations	356	356	192	192	164	164
R^2	0.008	0.057	0.008	0.095	0.008	0.049

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Controls are *age, gender, ethnicity, study year, pro-social survey measures and IOS score for stranger.* Standard errors in all models are clustered at the individual level.

Table C.1.3 replicates the analysis from above. Now with fewer observations and IOS scores, we find a small but significant effect of IOS on altruistic giving.

Table C.1.3: Altruistic giving by IOS Score for IOS > 1

	Pooled		Anonymous		Non-Anonymous	
	(1)	(2)	(3)	(4)	(5)	(6)
IOS Score	0.548**	0.276	0.727**	0.411	0.372	0.190
	(0.232)	(0.238)	(0.305)	(0.330)	(0.271)	(0.261)
Exchange Rate (Reference: 3:1)						
1:1		-0.247		-0.328		-0.154
		(0.937)		(1.155)		(1.206)
1:3		-0.196		-0.730		0.437
		(1.789)		(1.936)		(2.360)
Controls	No	Yes	No	Yes	No	Yes
Observations	257	257	138	138	119	119
R^2	0.027	0.115	0.043	0.161	0.014	0.105

* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Controls are *age, gender, ethnicity, study year, pro-social survey measures and IOS score for stranger.* Standard errors in all models are clustered at the individual level.

C.1.2 - Results - Social cohesion and distributional preferences

Table C.1.4 tests the parameter distributions across social distances. The table reports p-values of Kolmogorov-Smirnov tests. We see that as for altruistic giving, only α is significantly related to SD = 1. We find no effects for ρ or the other two social distances.

Table C.1.4: Estimated parameters across social distances

Comparison	α	ρ	$-1 \leq \rho < 1$
SD = 1 vs. SD = 2	0.001	0.486	0.543
SD = 1 vs. Nameless	0.001	0.642	0.501
SD = 2 vs. Nameless	0.332	0.455	0.767

Note. All values reported are p-values of Kolmogorov-Smirnov tests.

Table C.1.5 replicates the regression analysis from the text only for IOS > 1. Contrasting the results for altruistic giving, we still find no relationship between IOS and distributional preferences. However, estimated coefficients do increase.

Table C.1.5: Preference parameters by IOS Score for IOS > 1

	α			ρ		
	(1) Pooled	(2) Anony.	(3) Non-Anony.	(4) Pooled	(5) Anony.	(6) Non-Anony.
IOS Score	-0.0122	-0.0104	-0.0146	0.0170	0.0221	0.0337
	(0.0105)	(0.0124)	(0.0138)	(0.0350)	(0.0324)	(0.0662)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	76	42	34	61	36	25
R^2	0.183	0.194	0.413	0.430	0.378	0.565

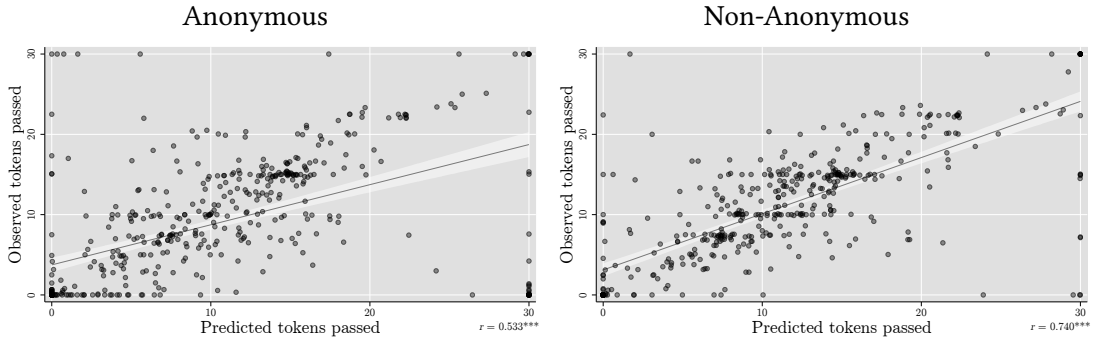
* p < 0.1, ** p < 0.05, *** p < 0.01.

Note. Controls are *age, gender, ethnicity, study year, pro-social survey measures and IOS score for stranger.* Standard errors in all models are clustered at the individual level.

C.1.3 - Results - Distributional preferences and altruistic giving

Figure C.1.1 compares *predicted* and *observed* tokens passed for both anonymous levels. In both levels of anonymity we find strong and significant correlations between our predicted and observed tokens passed.

Figure C.1.1: Predicted vs observed tokens passed by anonymity



C.2 - Experimental instructions

[Instructions were encountered by all subjects except when stated otherwise. Two dotted lines represent a new screen in the programme. All programmes are available upon request. Comments for the reader are included as [...] and were not seen by the subjects.]

[Network Elicitation]

Welcome!

If you are a second or third year undergraduate you are invited to participate in our online study. The entire study consists of 3 different parts. The first part is a 20 minute study today, followed by two additional parts in the next few weeks.

Please click "Continue" to proceed.

Earn Money and have a Chance to Win valuable Prizes!

You have been invited to participate in a study on economic decision making. The study consists of three separate Parts **in this autumn term**. You can only participate once in each part. You have just started the first part which is expected to take a maximum of 20 minutes.

Your base reward for participating in this part is **£3**. You can also win additional money from decisions you make during this part ranging **from £0 to £6. We will invite you again in 2- and 3 weeks to participate in Parts 2 and 3 of the study.** We hope that you will complete all three parts and encourage you to do so. All earnings will be paid out to you together with your participation rewards after the end of the study by December 6, 2021 via PayPal.

On top of this, you will also be automatically enrolled in a lottery where you can win cash prizes. After each part (1, 2, and 3) you finish you will be given tickets for the lottery. The more tickets you have, the greater your odds of winning increase. You can only complete a new part if you have finished the previous one. This implies for example that you can only complete Part 3 if you have previously finished Part 2. For each part of the experiment, you will receive the following number of tickets when completing each part

- **Part 1: 1 Ticket**
- **Part 2: 5 Tickets**
- **Part 3: 15 Tickets**

So, if you complete all three parts of the experiment, you **will receive a total of 21 tickets** for the lottery. Once Part 3 is finished, five tickets will be picked at random to determine the winners of the cash prizes. Note that each participant can only win one cash prize. **Each cash prize consists of £100 on top of the experimental earnings** transferred to your PayPal account by 6 December, 2021.

At the end of today's part you will be **provided with a unique ID, which you need to login to complete Parts 2 and 3** of this experimental series. It is therefore important that you save your ID securely. We will also email you your ID again after you finished Part 1.

.....

.....

Informed Consent

CeDEx researchers, Simon Gaechter, Chris Starmer, Fabio Tufano and Malte Baader, from the School of Economics at University of Nottingham, are conducting this study into how people make decisions. **You may participate in this study if you are a University of Nottingham undergraduate student.** In line with our standard practices, no part of this online study will involve any form of deception.

Who can I contact? If you wish to contact any of the researchers regarding the study please write an email to: **malte.baader@nottingham.ac.uk**

Study information. The study will be conducted online and in three parts. **You can only participate once in each part.** Part 1 will start right after you review this consent form and is expected to take about 20 minutes. In all parts, we will ask you and other participants to make a series of decisions and provide some demographic information. We will contact you again for Parts 2 and 3 in **2- and 3-week time** which are expected **to take around 30 minutes each.** Please note that we will share some of the information you provide to us with other participants in the study. However, no information will be shared with individuals that do not participate; all data will be stored anonymously; and all published results will only be reported at group level.

How much can I earn across all 3 parts?

In order to earn money for the study you must register with the CeDEx participant database and provide a PayPal account for payment. Otherwise, you cannot participate.

1. Today, you will be paid £3 for participating. Additionally, you can win a bonus of up to £6 depending on your and other participants decisions.
2. For each of the other parts you will receive £3 for participating and a bonus depending on your decisions.
3. For each of the three parts you complete, you enter a lottery to win one of five £100 cash prizes. For every part you complete you will receive an increasing number of additional tickets for the lottery, so your chances increase.

How long will it take? The part today will take around 20 minutes to complete. The follow-up parts will take around 30 minutes each.

Do I have to participate? No, participation is entirely voluntary and you are under no obligation to take part. You are free to withdraw at any point before or during parts of the study. All data collected will be kept confidential and used for research purposes only. Data collected will be stored in compliance with the Data Protection Act.

For participating in this first part of the study you will receive the following payments:

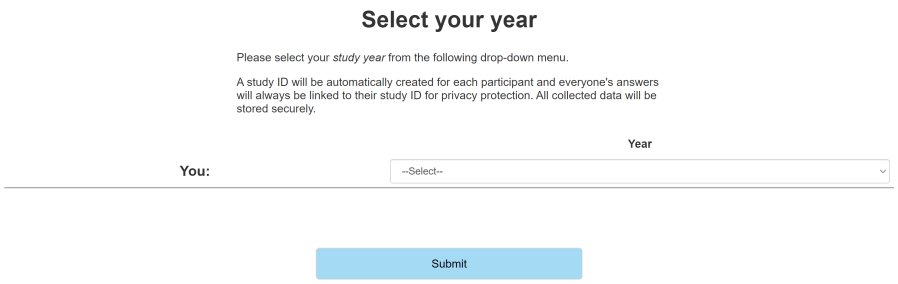
1. For completing Part 1 you will receive a **participation fee of £3** which will be transferred to your PayPal account by December 6, 2021. This is the minimum compensation you will receive for participating in the part.
2. You can earn additional money ranging from £0 to £6 from the decisions you make during the first part of our study. Those winnings will be added will also be transferred to your PayPal account by December 6, 2021.
3. Upon completion, today you will also **receive 1 ticket for a lottery** to win £100.

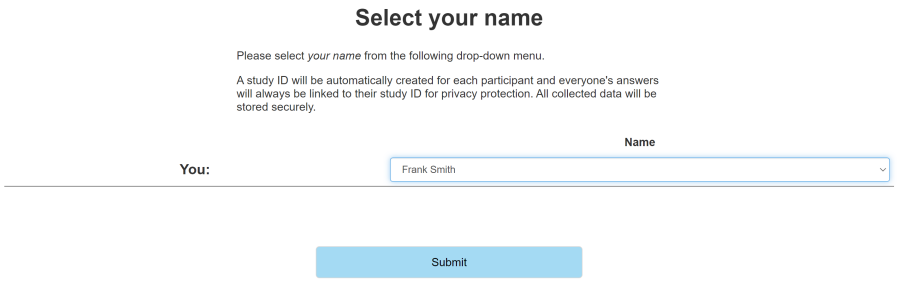
By clicking the 'I consent' button, **you agree to the following for all three parts of the study:**

1. I have read and understood the information about the study above.
2. I understand that some of the information I provide will be shared with participants in this study. However, no information will be shared with individuals that do not participate in the study.
3. I confirm that I have been given enough information about this study, and I voluntarily agree to take part.

If **you do not agree**, please click the 'I do not consent' button.

Thank you for your time.





Hi, [FirstName]

[Depending on their registration status subjects saw one of the following]

We could not find a record of you in our CeDEx (Centre for Decision Research and Experimental Economics) participant database. In order **to transfer you your money** for this study **please register with CeDEx** and add a **PayPal account** using the link below. [**Button to register**]

OR

We found a record of you in our CeDEx (Centre for Decision Research and Experimental Economics) participant database. However we could **not find a PayPal account** in order to transfer **you your money** for this study. [**Text input**]

OR

We found a record of you in our CeDEx (Centre for Decision Research and Experimental Economics) participant database.

The money you earn in this study will be transferred to **the PayPal account** registered with your CeDEx account. If you want to change the associated PayPal account please email "malte.baader@nottingham.ac.uk" for amendments.

.....

.....

Naming Task

We now ask you to complete a task that involves thinking about other people who might be participants in the study. Please follow the instructions and you can earn a bonus in addition to your participation fee.

.....

.....

Naming Task

On the next screen, please select the names of **10 people you know in your study year and programme**. All other participants in this study will also be asked to name 10 people.

Ideally, you **should name a total of 10 people, beginning with the ones you consider closest to you**. If you cannot identify as many as 10 people you know in your study year and programme, we will randomly select additional students from your programme and year for the remaining names.

To name a person, please click the "Add name" button and select their name from the dropdown menu. Everyone is listed in alphabetical order by first name. Note that you can only add another name if you have selected a name in all displayed dropdown menus. To remove the last name in the list simply click the "Remove name" button.

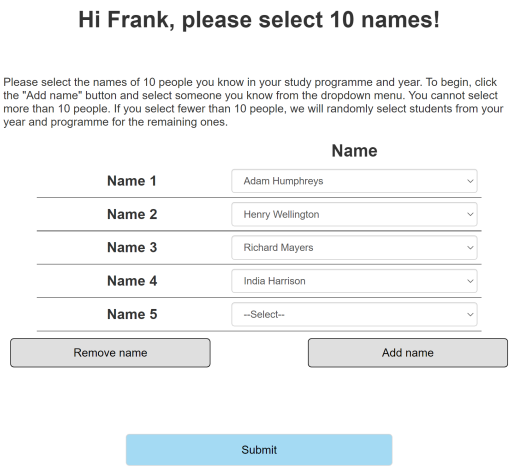
Bonus Earnings for the Naming task

The bonus for the Naming Task is designed to give you an incentive to name people who you think are most likely to know you. If you list a person who also completes the study and names you as well, with 50 percent probability you will receive a prize of £0.50 and £0.00 otherwise. If you name a person who does not name you, you will receive no bonus from choosing that name. Since you are allowed to name up to 10 individuals, you have 10 independent chances to win £0.50, meaning your bonus will be between £0 and £5.

In order to maximise your bonus, think carefully about the names you enter on the list.

.....

.....


Hi Frank, please select 10 names!
Please select the names of 10 people you know in your study programme and year. To begin, click the "Add name" button and select someone you know from the dropdown menu. You cannot select more than 10 people. If you select fewer than 10 people, we will randomly select students from your year and programme for the remaining ones.
Name
Name 1
Adam Humphreys
Name 2
Henry Wellington
Name 3
Richard Mayers
Name 4
India Harrison
Name 5
--Select--
Remove name
Add name
Submit

Circles Task

For each person you selected (or that was randomly selected) in the Naming Task on the previous screen you are now asked to complete a "Circle Task" to indicate the relationship you have with this person. To indicate your relationship, you will see the screen displayed below (in red).

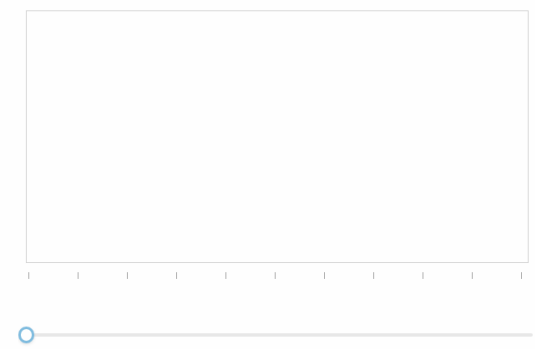
Once you move the slider below, a pair of circles will appear in the box. The position of the slider will determine the extent to which the circles overlap. When the slider is all the way to the left, the circles will look like this 

. When the slider is near the middle, the circles look like this 

. With it all the way to the right the circles look like this 

. You should interpret the degree of overlap as representing the relationship between you and the person.

Please position the slider so that the circles indicate to what extent you and the respective other person are connected.



*Please note that for the sake of the instructions the screen above is **not** interactive and the slider **cannot** be moved.*

For each person you selected (or that was randomly selected) in the Naming Task on the previous screen you are now asked to complete a "Circle Task" to indicate the relationship you have with this person. To indicate your relationship, you will see the screen displayed on the right of the name of the person.

Once you move the slider below, a pair of circles will appear in the box. The position of the slider will determine the extent to which the circles overlap. When the slider is all the way to the left, the circles will look like this 

. When the slider is near the middle, the circles look like this 

. With it all the way to the right the circles look like this 

. You should interpret the degree of overlap as representing the relationship between you and the person.

Please position the slider so that the circles indicate to what extent you and the respective other person are connected.

Just a couple of short questions to wrap up.

- What is your age? [**Numeric input**]
- What is your gender?
  - Female
  - Male
  - Other
  - Prefer not to say
- Which of the following describes you best?
  - Asian or Pacific Islander
  - Black or African American
  - Hispanic or Latino
  - Native American or Alaskan Native
  - White or Caucasian
  - Multiracial or Biracial
  - A race/ethnicity not listed here
  - Prefer not to say
- Imagine the following situation: you won 1,000 Pounds in a lottery. Considering your current situation, how much would you donate to charity? [**Numeric input**]
- How do you assess your willingness to share with others without expecting anything in return? *Please use a scale from 0 to 10 where 0 means "completely unwilling to share" and a 10 means you are "very willing to share". You can also use the values in-between to indicate where you fall on the scale.* [**11-point Likert scale**]
- Any feedback from your side regarding this study is highly appreciated. Please leave comments below. [**Text input**]

Your personal ID

As a log-in for the next two parts in 2- and 3- week time, please make a note of your personal ID.

Please note, to avoid confusion there can never be a 0 ("zero") or 1 ("one") in your ID and all letters are upper case.

Make sure you save this ID and have access to it for the all remaining parts.

ID: [**UniqueID**]

Thank you

This concludes Part 1 of the study. We will invite you for Part 2 in 2 weeks where you can again earn money and increase your chances of a cash prize of £100.

Thank you very much for your participation. Once you have recorded the ID you can close this window.

Wave 2 & 3

Welcome to this study!

This is part 2 of the experimental series.

In this part you will again be paid **£3 for participating and an additional bonus** depending on your and other participant's choices.

You will also receive **5 additional lottery tickets** that increase your chances of **winning one of five £100 cash prizes.**

Please read all instructions carefully, answer the comprehension questions and complete the associated tasks.

Please click 'Continue' to proceed

Please enter your ID

At the end of the first part of this series of experiments we provided you with a unique ID. Please enter this ID in the box below. We also sent you a reminder with your personal ID to your University email address using the subject 'Online Experiment: Invitation Part 2'.

[**Text input**]

Welcome [First name]

If the name above is your actual name, please click 'Continue' to move on. Otherwise, click the 'Back'-button in order to insert your correct ID.

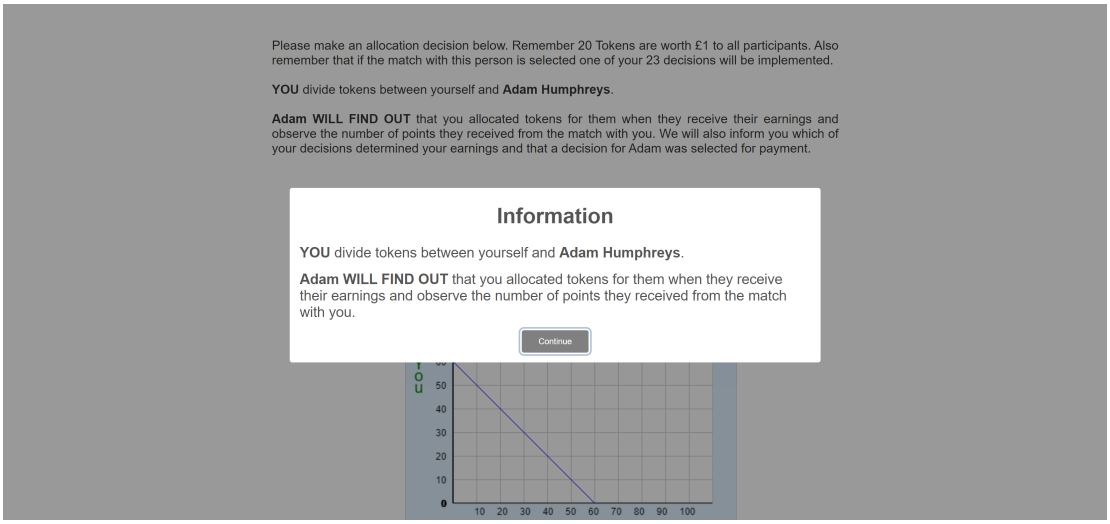
Success!

Today's experiment consists of 1 Task. The instructions will be provided step by step as you progress through today's part. Please click 'Continue' to start the instructions.

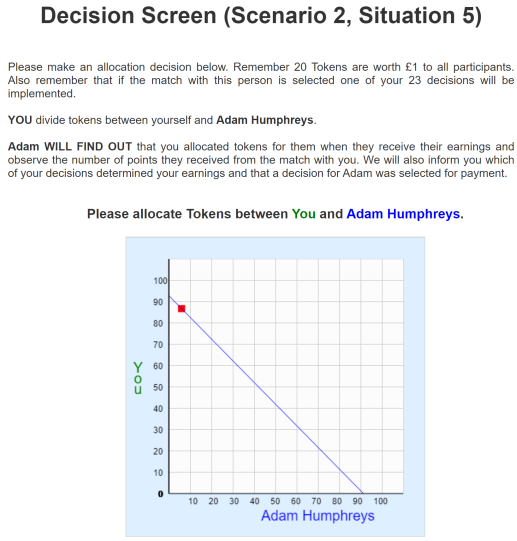
[**Instructions and comprehension questions exactly as in condensed AM/FKM (Appendix A.2.**]

[**Subjects perform the task below 23 rounds for 3 partners: SD = 1, SD = 2 and Nameless**]

[**When subjects load the screen they receive this message**]



[**This is the decision screen. Subjects play 23 rounds of this**]



.....

.....

Thank you for participating in this part of the study!

Please remember, we will send you another email to complete the third and last part of this experimental series next week.

For part 3 you will receive another 15 lottery tickets to win one of the five £100 cash prizes on top of earnings from your decisions.

You can now close this window.

As usual, any feedback from your side regarding this part of the study is highly appreciated. Please leave comments below. [**Text input**]

.....

.....

Appendix for Chapter 4

D.1 - Additional analyses

Table D.1.1: Overview of studies investigating incentives for creativity

#	Authors (Year)	Creativity task(s)	Incentives	Effects found	
Individual creativity					
1	Rutström & Williams (2000)	Tower of Hanoi	NA	NA	
2	Ariely et al. (2009)	Packing quarters	Threshold	medium incentives:	+
				high incentives:	-
3	Eckartz et al. (2012)	Word task	Piece-rate & tournament	Piece-rate:	0
				Tournament:	0
4	Artes et al. (2019)	Matchstick puzzles	Piece-rate, tournament & time pressure, stakes	High stakes:	-
				Others:	0
5	Kleine (2021)	Candle problem	Piece-rates	Gain framing:	-
				Loss framing:	-
6	Bradler et al. (2019)	Unusual Uses Task	Tournament & gift-exchange	Tournament:	+
				Gift-exchange:	0
7	Artes et al. (2019)	Unusual Uses Task	Piece-rate, tournament time pressure, stakes	Tournament:	+
				Others:	0
8	Laske & Schröder (2017)	Word illustration task	Piece-rate	Quantity:	+
				Quality:	+
				Originality:	0
9	Toubia (2006)	Idea generation	Piece-rate	Own ideas:	+
				Impact on others:	++
10	Charness & Grieco (2019)	Verbal & math task	Tournament & Ranking	Ranking:	+
				Tournament:	++
11	Attanasi et al. (2019a)	Draw, verbal & math task	NA	NA	
12	Attanasi et al. (2020)	Drawing task	Choice of tools	Restricted choice:	+
				Free choice:	+
13	Kachelmeier et al. (2008)	Design rebus puzzles	Piece-rate	Piece-rate:	-
14	Kachelmeier & Williamson (2010)	Design rebus puzzles	Piece-rate	Piece-rate:	+/-
15	Erat & Gneezy (2016)	Design rebus puzzles	Piece-rate & Tournament	Piece-rate:	0
				Tournament:	-
16	Charness & Grieco (2019)	Invention & society description task	Tournament & Ranking	Ranking:	+
				Tournament:	+

Task openness →

Table D.1.1: Overview of studies investigating incentives for creativity

#	Authors (Year)	Creativity task(s)	Treatments	Effects found	
Group creativity					
17	Ramm et al. (2013)	Candle problem	Group v individual	Group (tournament):	0
			No v tournament	Group (No):	+
18	Grözinger et al. (2020)	Word illustration	Group only	Video:	-
			F2F v video v chat	Chat:	- -
19	Attanasi et al. (2019b)	Draw, verbal & math task	Group v individual	Group (ind):	+
			Ind v coop v comp	Group (coop):	0
				Group (comp):	+
20	Charness & Grieco (2021)	Draw, verbal & math task	Group only	Rank (comp):	+
			No v rank v tournament	Rank (coop):	+
				Tournament (comp):	+
				Tournament (coop):	++
21	Chen et al. (2012)	Use of an abandoned house	Group v individual	Group (piece-rate):	-
			Piece-rate vs tournament	Group (tournament):	+
22	Gneezy et al. (2021)	Create video title	Group v individual	Group:	0
			No v tournament	Tournament:	0
Field studies and observational data					
23	Englmaier et al. (2021)	Solve escape rooms	Group only	Tournament:	+
			Tournament		
24	Gross (2020)	Logo innovation competition	Tournament	Tournament:	+/-
25	Gibbs et al. (2017)	Bonus scheme idea	Piece-rate	Bonus scheme:	+

Task openness →

[**All subsequent analyses are referenced in the document. Brief descriptions are provided above every analysis.**]

D.1.1 - Study 1 - Comparing *general* creativity tasks

Table D.1.2 presents the regression results for all pairwise comparisons in Study 1. With the inclusion of controls, all marginal effects found previously are insignificant.

Table D.1.2: Pairwise comparisons with controls across all three tasks (Study 1)

	Word Task (1)	UUT (2)	Figural Task (3)
Figural: Fraction of Wins	-7.606		
	(48.71)		
Domains of creativity	19.00	1.470	0.0248
	(42.78)	(4.394)	(0.0926)
Female	20.04	-0.478	-0.00952
	(23.09)	(2.287)	(0.0846)
Age	2.969	0.0446	-0.00914
	(5.303)	(0.457)	(0.0144)
Study Subject (Reference Category: Business & Economics)			
Law	-41.66	-5.808	-0.0497
	(40.41)	(3.626)	(0.143)
Natural Sciences	22.76	-2.418	0.0715
	(40.16)	(3.807)	(0.115)
Social Sciences	98.26**	-8.905*	-0.0922
	(43.12)	(4.576)	(0.138)
Psychology	74.01	1.526	0.109
	(52.64)	(5.607)	(0.175)
Humanities	51.64	-7.807	-0.256*
	(61.81)	(4.663)	(0.146)
Politics	20.95	-6.684	-0.0744
	(44.39)	(3.981)	(0.142)
Word: Total Points		-0.00269	-0.0000568
		(0.0181)	(0.000487)
Unusual Uses: Correct Uses			0.00529
			(0.00610)
Observations	56	48	48
R^2	0.313	0.351	0.302

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Note. dfgdf

D.1.2 - Addressing measurement error

Table D.1.3 presents the first and second stage of our noise reduced figural task. The first model presents evidence on the first stage regression, relating subjective performance to a number of objective metrics. In models (2) and (3), we relate the instrumented score of the figural task to the other creativity tasks, without finding an effect.

Table D.1.3: Addressing measurement error in the figural task

	First Stage	Second Stage	
	(1) Min Success	(2) Word points	(3) Fluency
# of figures	-0.00826 (0.00773)		
# of circles	0.00459** (0.00225)		
Circles per figure	0.00131 (0.00426)		
Fraction Moved	-0.0339 (0.106)		
Area covered	0.532*** (0.156)		
Figural task (instr.)		-69.95 (91.91)	14.30 (9.938)
Controls	Yes	Yes	Yes
Other Self-Reports	Yes	Yes	Yes
Observations	56	56	48
R^2	0.397	0.323	0.398

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Note. dfgdf

Table D.1.4 demonstrates the results of using the ORIV approach. We show regression outputs including controls for each of the isomorphic problems (A.1, A.2 and B.1, B.2) as well as the combined result, as instrumented. Accounting for measurement error does not affect the results.

Table D.1.4: Addressing Measurement Error - ORIV Approach

	Huiking task: Fluency		
	(1) OLS	(2) OLS	(3) ORIV
ToL Problem A.1	-2.546*** (0.893)		
ToL Problem A.2		-0.919 (1.277)	
Instrumented			4.578 (17.04)
Observations	121	121	242
ToL Problem B.1	-1.134 (0.888)		
ToL Problem B.2		0.544 (0.941)	
Instrumented			0.785 (2.728)
Observations	121	121	242

* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors in parentheses.

Note. Success with the fewest moves is the measure of performance in the ToL. ToL Problem A.1 and A.2 are one isomorphic pair, and B.1 and B.2 are the other. All models include controls *ethnicity, native language, employment status, student status, education, gender, age,* and a *survey measure of risk.* In addition we also include a treatment dummy and session fixed effects.

D.2 - Experimental instructions

[**Instructions were encountered by all subjects except when stated otherwise. Two dotted lines represent a new screen in the programme. For Study 1, the instructions were printed out for the subjects. The z-Tree programme did not include any additional instructions. All programmes are available upon request. Comments for the reader are included as [...] and were not seen by the subjects.**]

.....

.....

D.2.1 - Study 1

[**We utilised distinct instructions for the roles of regular participant, examiner and judge. Here we present only the instructions for regular participants as they are most important for our study. All files and instructions can be provided upon request.**]

- Regular participant -

Instructions [**translated from German**]

Welcome to this economic experiment. Please be silent at your computer, do not communicate with other participants throughout the experiment and switch off your mobile phone.

In case you have any questions, please raise your arm and wait until one of the experimenters comes to you.

For the participation of todays experiment you will receive your earnings in cash. Your earnings will be computed in points. **The experiment consists of 2 parts. Part 1 consists of 3 stages and part 2 of 2 stages**. The points that you earned will be converted into Euros according to the following exchange rate:

Part 1 - Stages 1, 2 & 3:	**1 points = 1 Euro.**
Part 2 - Stages 1:	**50 points = 1 Euro.**
Part 2 - Stages 2:	**1 points = 1 Euro.**

Your final income is composed of your **total earnings from part 1, one randomly selected stage from part 2 and 3 Euros for completing the questionnaire.** During the experiment neither you nor any other participant will receive feedback about the performance.

Please read the instructions carefully. Afterwards you will make your decisions in the experiment. All decisions will be dealt with anonymously. In the following the exact procedure of the experiment is explained.

Part 1

In this part you will be matched with a randomly selected participant into a group at the beginning of every stage.

Stage 1 – Uses Task

This task consists of 1 round. In this task all participants will be working **6 minutes** on a task that requires creativity.

Example of the task

Please list as many, as different and as unusual uses for a rubber tire as you can think of. Do not restrict yourself to a specific size of a tire. You can also list uses that require several tires. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible.

Please confirm **every entered use** by clicking OK.

Evaluation of the task

In this experiment 5 participants act as judges and assign your answers to a category and the appropriate use. Moreover, 2 participants act as examiners and assess whether your answer is valid. Answers are valid when they are practicable and when their realization is at least vaguely conceivable. Please describe the possible use in a few words if necessary (Using the example of the rubber tire: “sled” or “flower box” are clear answers, whereas “target” would require further explanation such as “ball game with tire as target”.).

Throughout this stage you act as a participant, hence not as judge or examiner.

You receive **1 point** per valid answer and **1 point** per mentioned category. Using the example of the rubber tire: “car tire” and “bicycle tire” yield 2 points and belong to the category “tires as wheels”, thus result in 1 extra point. The answer “swing seat” yields 1 point and is a different category (category “toys”), thus results in 1 additional point.

Moreover you receive **0.5 points** for original (rare) answers and **1 point** for very original (very rare) answers. An answer is considered (very) original if only (very) few people think of it. To this end, the answers are compared to all mentioned answers in this experiment. An answer that is only mentioned by one other participant is considered original and an answer only mentioned by you is very original.

Your achieved points will then be compared with your group member and the winner receives 8 points, the loser 0 points. In case of a draw both receive 4 points.

Stage 2 – Word Task

This task consists of 1 round. At the beginning of this task you will again be matched with a randomly chosen participant into a group. You have **5 minutes** in order to construct English words using a letter-set consisting of 12 letters. You can construct short and long words. In the process, not only nouns, adjectives and verbs count, but also articles or prepositions. Please confirm every entry with the enter-key in order to receive feedback whether your answer is valid or invalid.

Example of the task

Suppose you have to construct words from the following letter-set: **"abcdeabdeh"**.

- the word “bad” is valid.
- the word "babe" is valid, as the letter “b” appears twice in the set
- the word “dach” is NOT valid, as it is a German word
- the word “day” is NOT valid, as there is no “y” in the set

Evaluation of the task

You only receive points if you construct existing English words. The longer the constructed words, the more points you receive. (Words consisting of 2 letters result in $1 + 2 = 3$ points, words consisting of 3 letters in $1 + 2 + 3 = 6$ points, and so on).

After every valid or invalid answer you receive feedback whether the word is admissible and how many points you receive for the entered word.

As in the previous task, your achieved points will be compared with the points of your group member and the member with more points is the winner. The winner receives again 8 points, the loser 0 points. In case of a draw both receive 4 points.

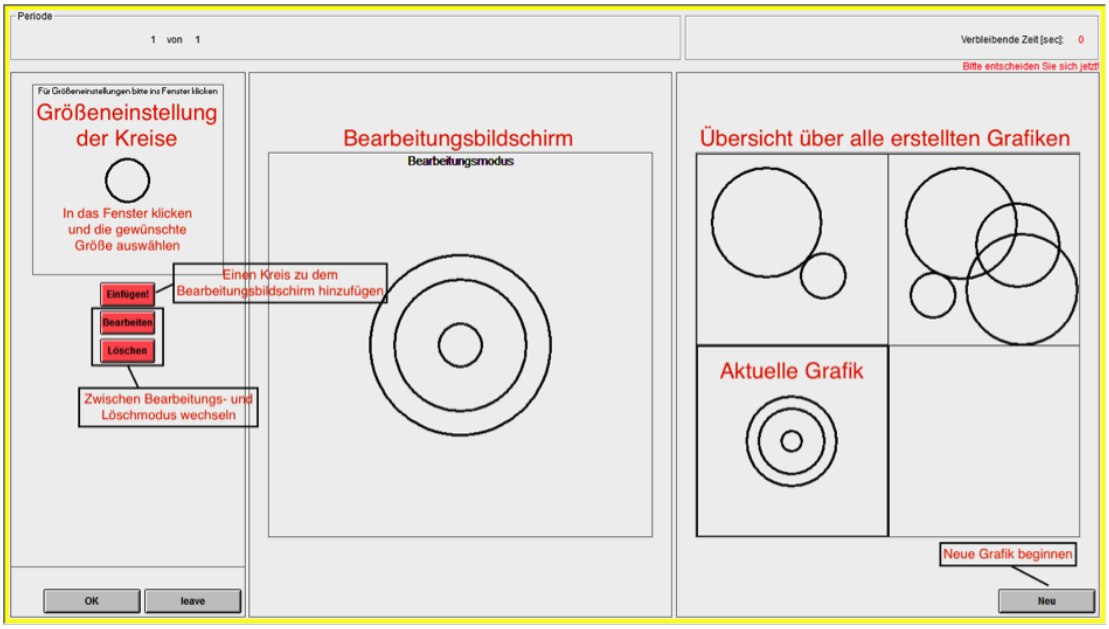
Stage 3 - Graphic Task

This task consists of 1 round. At the beginning of this task you will again be matched with a randomly chosen participant into a group. You have **5 minutes** in order to construct creative graphics using circles. In order to do so you can construct as many graphics as you like.

Example of the task

The size of the circles can be adjusted before inserting them into the edit screen. Afterwards you can move circles (click on the ring of the circle and move the mouse indicator) and delete circles (click on the ring of the circle in the delete mode). You can also edit finished graphics (click on the respective graphic in the overview screen). The framed graphic is always the graphic currently in the edit screen. Should there be no framed graphic you will commence with a new one.

The screenshot below describes and visualises the possibilities with which the graphics can be constructed.



Evaluation of the task

In this task 5 participants act as judges. The judges compare your set of graphics in a pairwise manner with another randomly selected set of graphics. The decision criterion is creativity of the graphics. Judges evaluate creativity as a whole, thus the number of constructed graphics does not have a special role. The judges select the work that is considered as more creative as the winner. For your payoff in this task your relative performance in comparison to your group member is relevant.

The participant who won a larger fraction of pairwise comparisons wins the group and receives 8 points, the loser receives 0 points. In case of a draw both receive 4 points.

Part 2

At the end of the second part **one of the stages** will be randomly selected and determines your earnings in part 2.

Stage 1

Please note that for this task the conversion rate is **50 points = 1 Euro**. This task consists of 6 rounds. At the

beginning of this stage you will be matched with a randomly chosen participant into a group. In each of the 6 rounds you have to make a decision in which you allocate points to yourself and your assigned partner.

Evaluation of the task

In case stage 1 in part 2 is selected as relevant for payment 2 rounds will be randomly selected to determine your earnings. In the first randomly selected round your earnings is a result of the points you allocated to yourself. In the second round the points that your partner allocated to you counts. All points are then added.

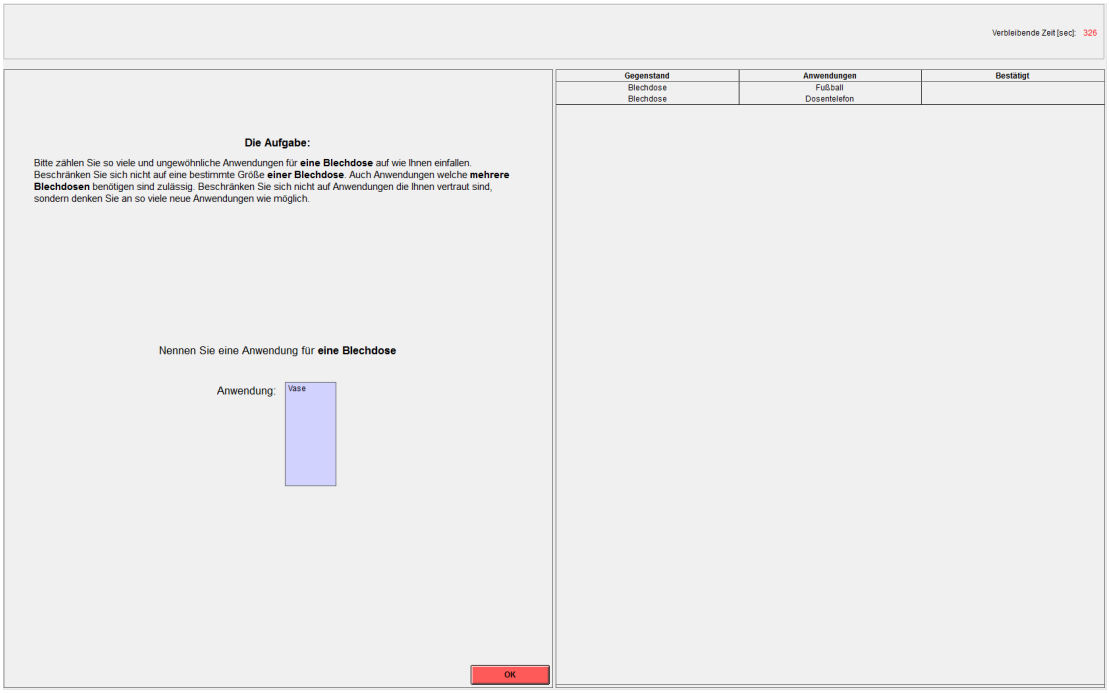
Stage 2

In this task the conversion rate is as at the beginning **1 point = 1 Euro**. The task consists of 1 round. You have to make 23 decisions between 2 different options. Option A is always a lottery in which you receive 5 points with 50% chance and 0 points with 50% chance. Option B always guarantees you a fixed amount. This amount decreases in 0,20 point steps from 5 points to 0,60 points and you have to make a decision in each step between either option A or option B.

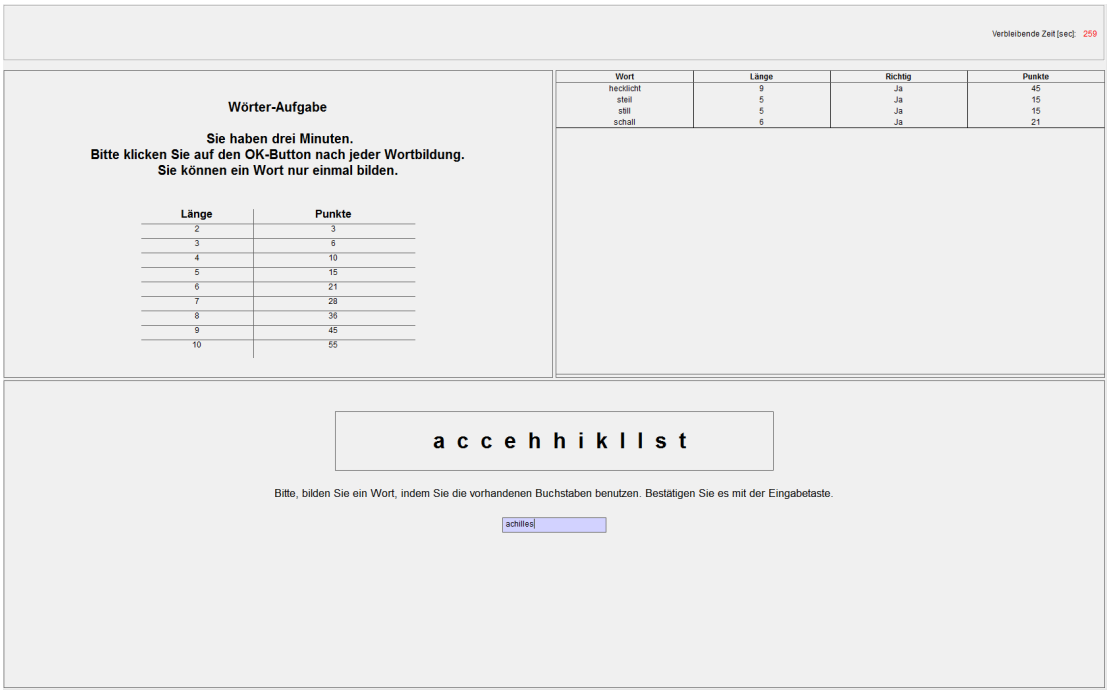
Evaluation of the task

In case stage 2 in part 2 is selected as relevant for payment 1 of the 23 decisions will be randomly selected, the chosen option executed and disbursed.

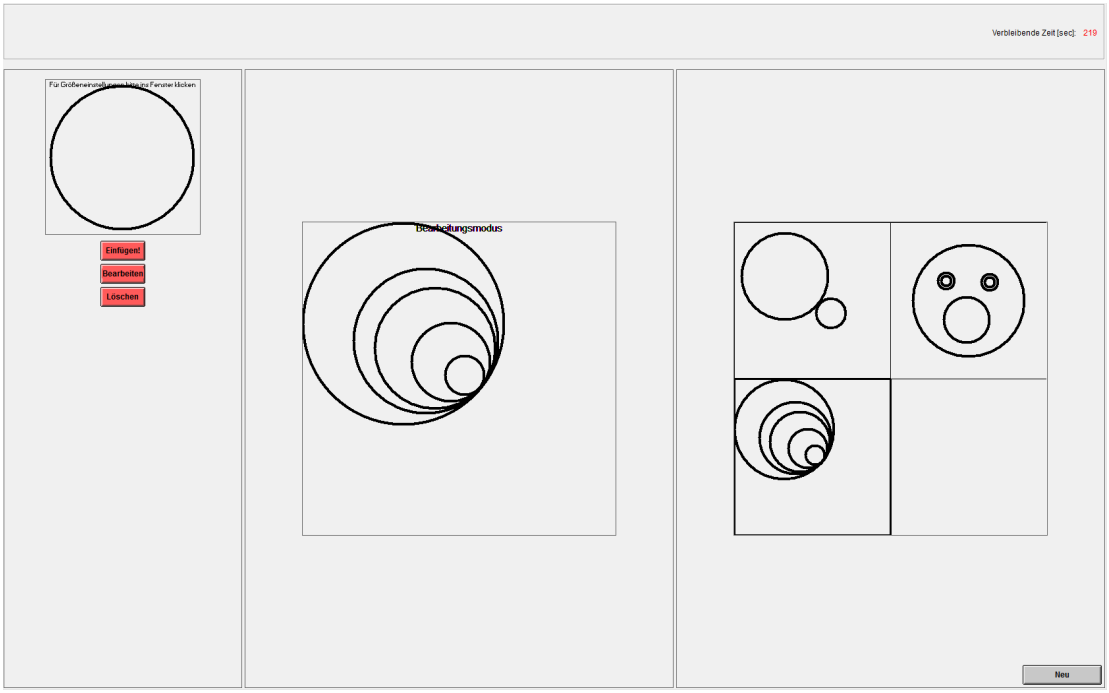
Subject screen: Unusual uses task



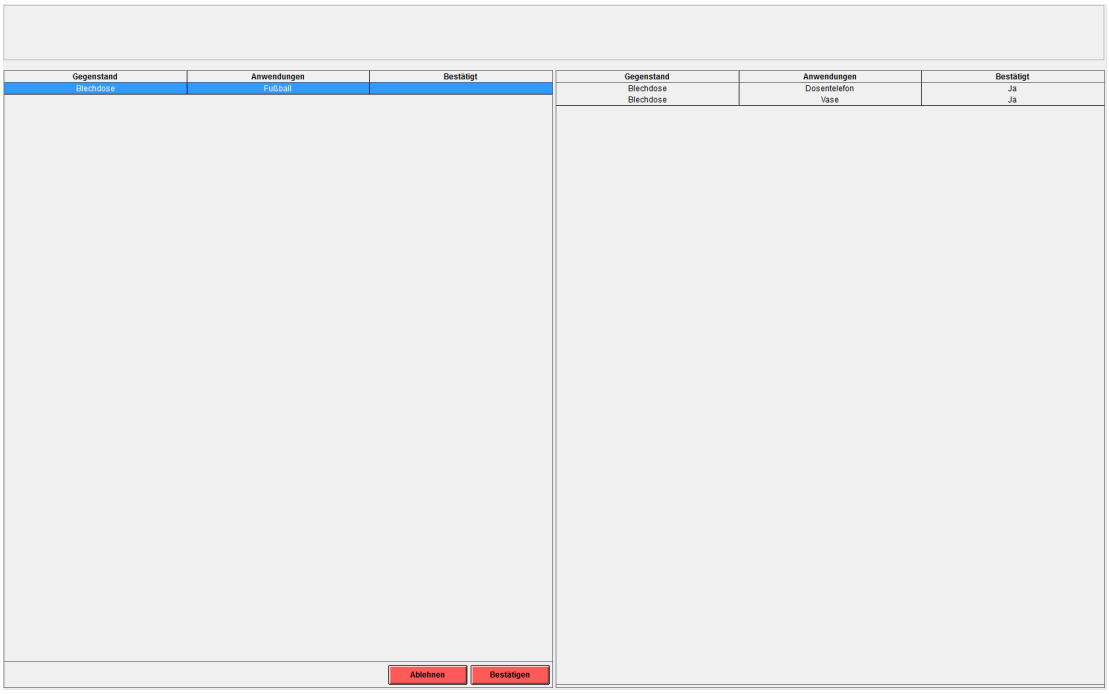
Subject screen: Word task



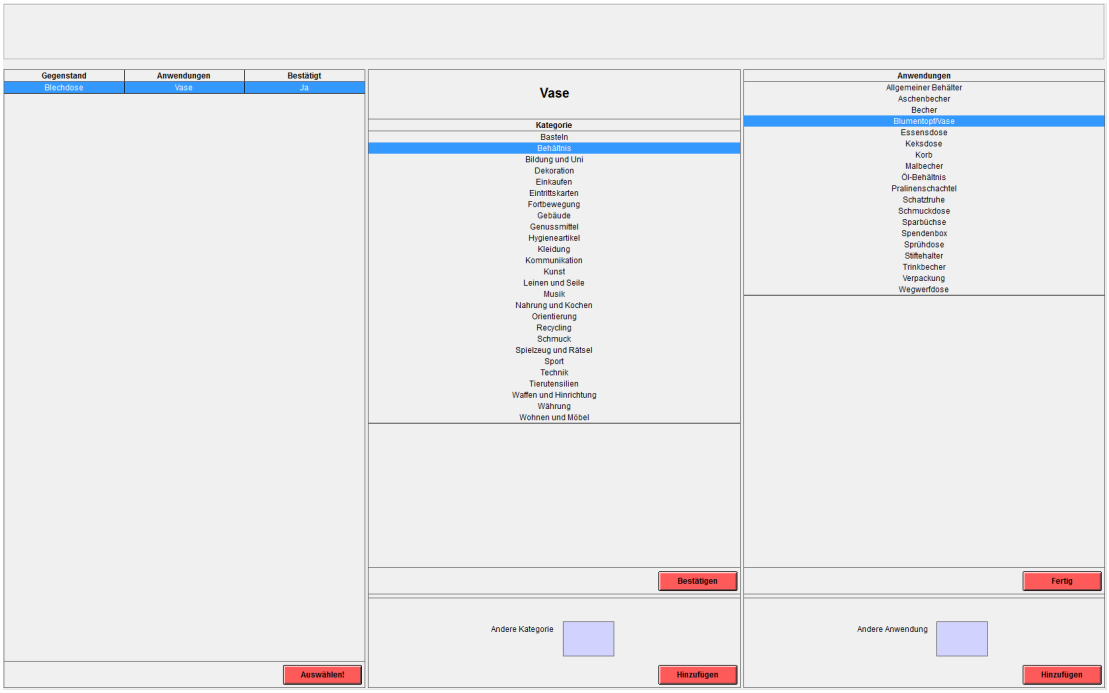
Subject screen: Figural task



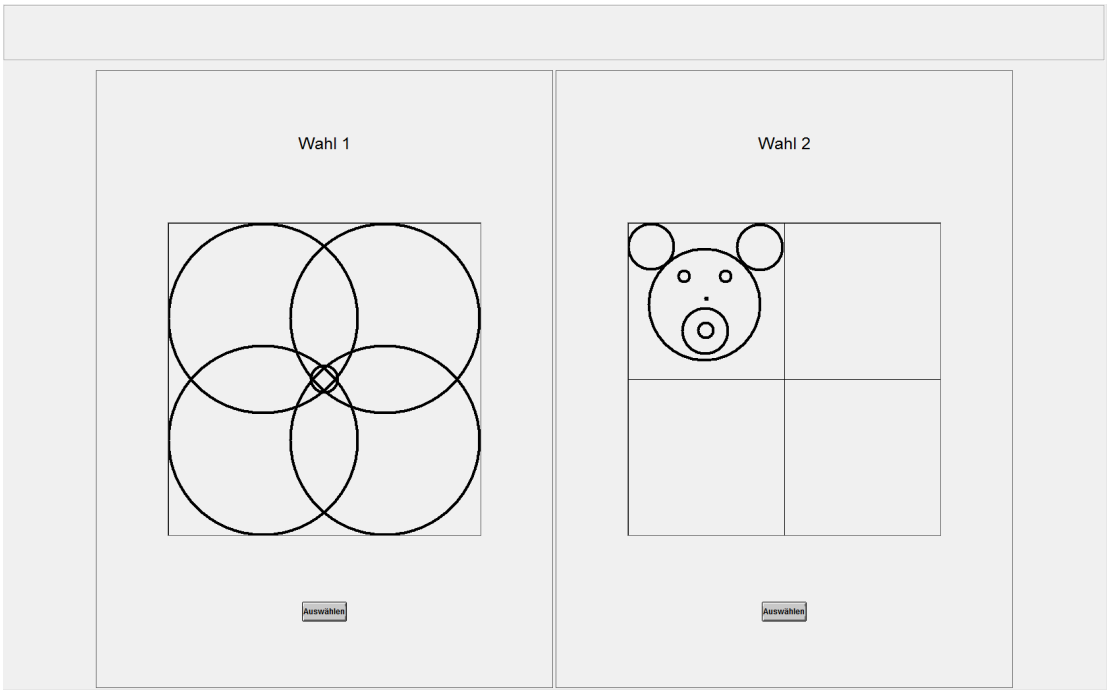
Examiner screen: Unusual uses task



Judge screen: Unusual uses task



Judge screen: Figural task



D.2.2 - Study 2

Welcome to this study

In this study we will ask you to answer several questions. You will be paid a flat fee of £2.50 for completing this study.

Please click continue to proceed.

Instructions - Part 1 [Tower of London]

Please read through the instructions below carefully.

This study consists of 3 parts.

These are the instructions for **part 1**:

You will be asked to play a game for 8 rounds. You have **60 seconds** to complete each round.

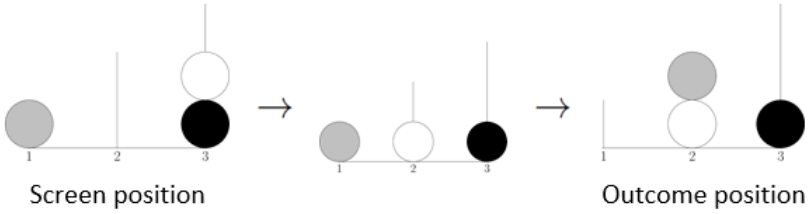
In each round you must adjust the starting position to match the outcome position. On the screen there are 3 balls and 3 rods , which differ in length, 1 to 3, from left to right. The smallest rod can hold 1 ball, the middle can hold 2, and the largest rod can hold 3 balls at one time.

You can move the top ball from each rod to any of the other rods as long as the rod is not full. You can move the ball to any rod, you do not have to move the ball to a neighbouring rod.

The aim of the game is to use the **fewest moves possible**.

Below you find an example of the task screen. You can move balls by clicking on the chosen ball and then the rod you wish to move it to. On the next page you will be able to practice the game.

Once you finished a round or run out of time you will be able to move on.



Comprehension Questions

In order to move on please answer the following 3 questions

1) What is the goal in each round?

- Using the fewest moves
- Using the most moves
- Does not matter

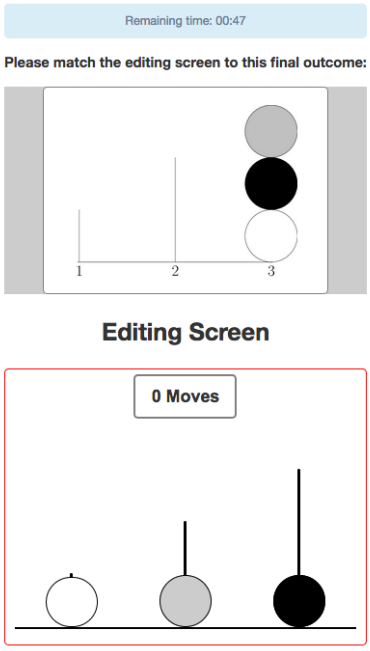
2) How much time do you have for each round?

- 30 seconds
- 120 seconds
- 60 seconds

3) Can you only move a ball to its directly neighbouring rod?

- Yes
- No

[**Subjects complete 1x Practicec and 8x Tower of London in random order**]



Instructions - Part 2 [Hiking task]

Please read through the instructions below carefully.

In this part you will be given a question which you have to answer. You have **5 minutes** to provide as many diverse answers as possible.

The question is about a real-life situation. Try to imagine yourself in that situation and answer the question to the best of your ability.

When thinking of your answers try to come up with as **many diverse** ideas as you can.

After **each idea**, you must click 'Submit'. You cannot submit by clicking 'Enter'. And you cannot change an answer once it is submitted.

Once the time is up you will **automatically** be moved on.

Comprehension Questions

In order to move on please answer the following 3 questions

1) What is the goal of the task?

- To come up with as many ideas as possible
- To come up with as diverse ideas as possible
- To come up with as many diverse ideas as possible

2) How much time do you have to come up with answers?

- 10 minutes
- 5 minutes
- 2 minutes

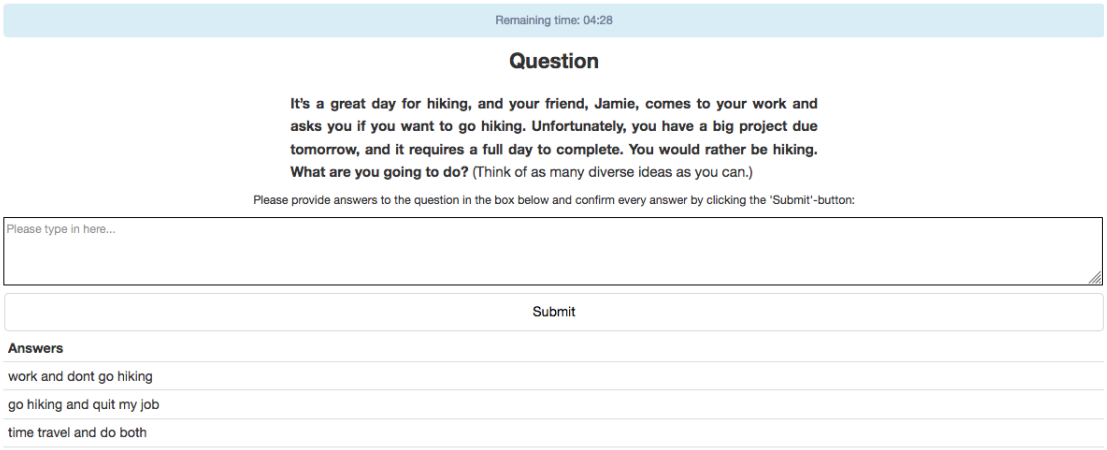
3) When do you have to click 'Submit'?

- After every single idea
- After providing multiple ideas to the question
- Never

.....

.....

[**Subjects complete 1x Hiking task**]



.....

.....

Instructions [Evaluation of Hiking task]

Please read through the instructions below carefully.

In this study you will be asked to **evaluate solutions** to a problem. These solutions have been taken from responses of previous study participants.

The scenario that the previous participants faced was the following:

It's a great day for hiking, and your friend, Jamie, comes to your work and asks you if you want to go hiking. Unfortunately, you have a big project due tomorrow, and it requires a full day to complete. You would rather be hiking. What are you going to do?
(Think of as many diverse ideas as you can.)

You have to evaluate two aspects of the solutions:

1. **Validity**:
The solution addresses the problem even if it is a bit fanciful. You must indicate, Yes or No, as to whether the answer is valid or not.
2. **Originality**:
How original is the solution to the provided problem? Please use your own judgement as to what you consider original. You should score originality on a scale from 1 to 7, where 1 is "not original" and 7 is "very original".

For solutions that you evaluate as invalid, you must indicate the reason for classifying it as invalid. In case you classify something as valid you must give it an originality score.

After each evaluation is complete, please click '**Submit**' to confirm the evaluation and move on to the next answer.

On the following page we present you with a couple of examples to evaluate so that you can familiarise yourself with the task. Moreover, we indicate what score we would have provided so you can compare your evaluation with ours.

Comprehension Questions

In order to move on please answer the following 3 questions

1) Which aspects do you have to evaluate?

- Quality & Length
- Validity & Originality
- Elaboration & Effectiveness

2) What makes a solution valid?

- Existence of an answer
- A perfect solution to the problem
- The solution addresses the problem, even if it is a bit fancyful

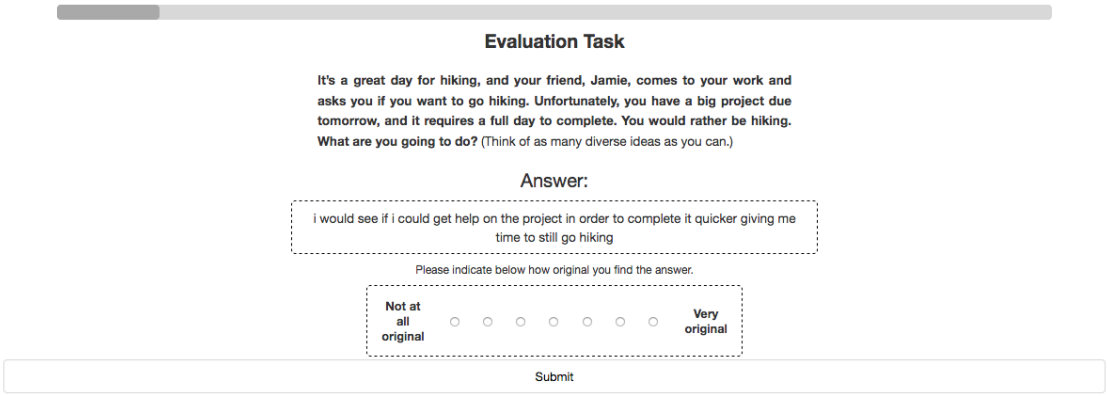
3) Do you have to score originality for invalid answers?

- Yes, I have to score all answers
- No, but I have to give a reason for invalidity

.....

.....

[**Subjects complete 4x practice evaluations and then move to evaluating the a set of answers by other subjects**]



.....

.....

Questionnaire

Please answer for each question below where you sit by indicating "not at all" on the left and "very" on the right. [**For 1) - 4): 7-item Likert scale; 1 - Not at all creative; 7 - Very creative**]

1) In general how creative are you?

2) In general how good are you in problem-solving?

3) Consider a problem that has a perfectly specified goal and there are multiple ways to solve it, such as: “Come up with a way to protect a raw egg, that prevents it from cracking at any height of a fall." How good are you in solving these kind of problems?

4) Consider a problem that has no perfectly specified goal and there are multiple ways to solve it, such as: “Come up with a household item that does not yet exist but everyone needs." How good are you in solving these kind of problems?

.....

.....

Questionnaire

Please indicate for each statement the extent to which you see that the statement is true about you, where "not at all true" is on the left and "very much true" is on the right.[**For 1) - 7): 7-item Likert scale; 1 - Not at all true; 7 - Very much true**]

1) I am good at coming up with new ideas for solving problems.

2) I have a lot of good ideas about how to solve problems.

3) I have a good imagination when it comes to solving problems.

4) I am confident that I can usually produce multiple solutions to a problem.

5) I am confident that I can produce useful solutions to problems.

6) I am not good at coming up with new ideas for solving problems.

7) I am confident that I can solve non-routine problems.

.....

.....

Questionnaire

1) Please tell me, in general, how willing or unwilling you are to take risks? Please use a scale from 0 to 10, where 0 means 'Completely unwilling to take risks' and a 10 means you are 'Very willing to take risks'. Pick a response from 0 to 10 to indicate below where you fall on this scale.[**11-item Likert scale; 1 - Completely unwilling; 7 - Very willing**]

.....

.....

Questionnaire

1) What is your age? (Numeric input)

2) What is your gender? (Text input)

3) What is your ethnicity? (Text input)

4) What is your first language? (Text input)

5) What is the highest degree or level of school you have completed?

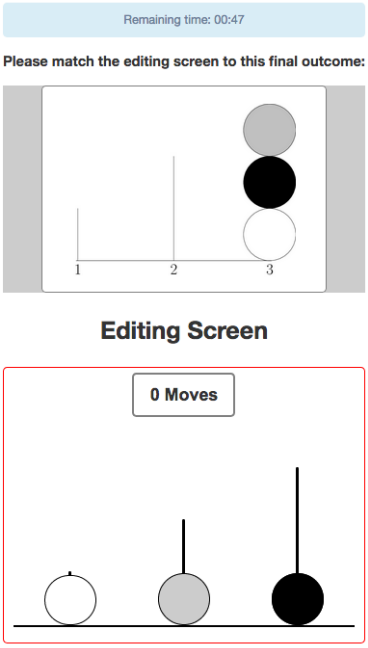
- Undergraduate degree (BA/BSc/other)
- College/A levels
- Secondary school/GCSE
- Graduate degree (MA/MSc/MPhil/other)
- No formal qualifications
- Doctorate degree (PhD/MD/other)
- Prefer not to say

Any feedback regarding this study is highly appreciated. Please leave comments below. (Text input)

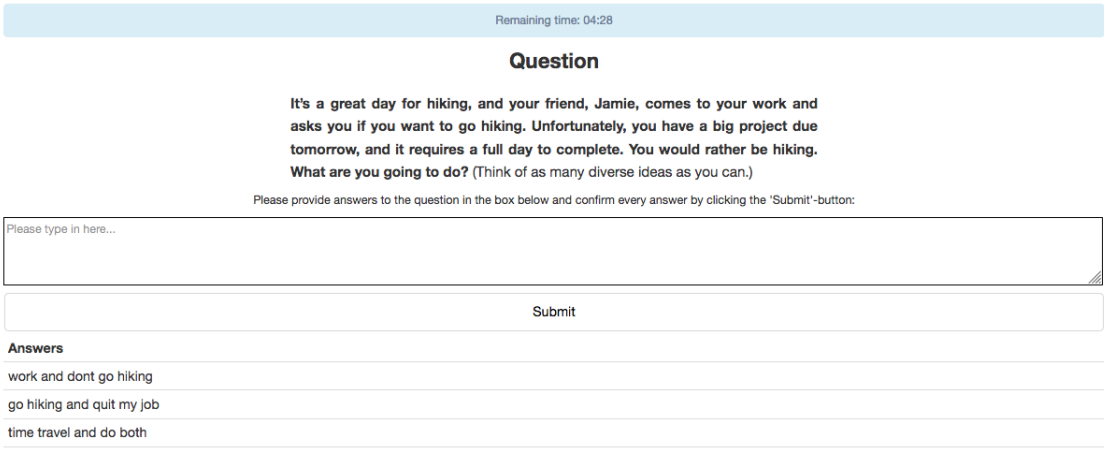
.....

.....

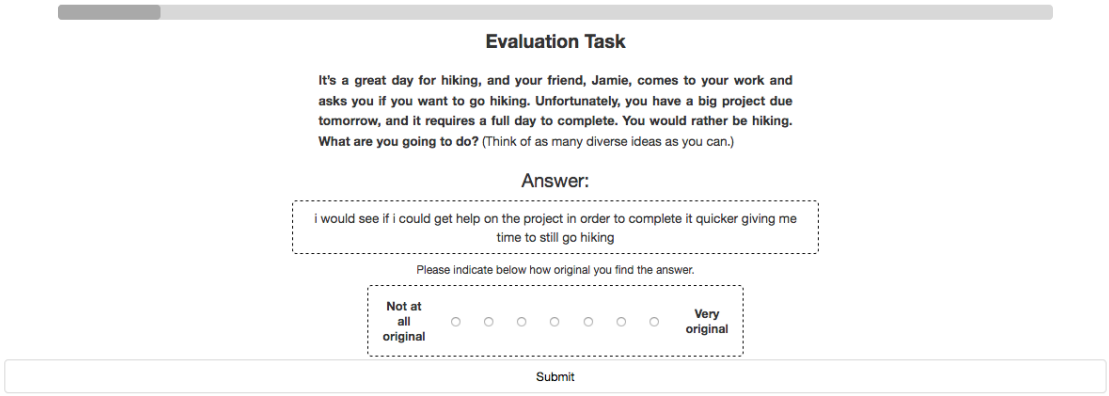
Subject Screen: Tower of London



Subject Screen: Open Problem-Solving



Judge Screen: Evaluation of Open Problem-Solving



D.3 - Creativity tasks

The subsequent sections provide details of the creativity tasks and an overview of output generated. For all detailed data please contact the author.

D.3.1 - Word task

[This is the letter-set we used in the experiment. The method of generating the letter-set and corresponding words can be found in (Eckartz et al., 2012). All words used by subjects within our study are in bold]

letters	points	words	similarity within
accehhikllst	5585	330	0.888436

ach achilles achse achsel acht achte achteck achtecks **achtel** achtes achtle ahle ai **akt** akte aktie akts alice alices **all all alle alles** alls **als alt alte** altes **asche** asket **ast** at **ca** cache caches call calls cellist ch chalet chalets chate chi **chic** chice chices chicste chile cia **echt eh ei eilst eilt eis eiskalt** eklat **elch** elchs eli elias elis **es** esc et etc eth ethik ethisch **hacke hackst hackt** hackte **hai** haie haies hais **hake** hakst hakt hakte **hall halle** halls hallst **hallt** hallte **hals halt halte** hasche hascht haschte **hase** haskell hast **haste hat he hecht** hechts **heck** hecklicht hecklichts hecks heckst heckt **hehl** hehlst hehlt **heil heilst heilt hektisch hell** hellst hellt **hielt hit ich ist** it **kachel kahl** kahle kahles kahlheit **kai** kais kali kalis **kalt** kalte kaltes **kastell keil** keils **keilst** keilt **kelch** kelchs **kiel** kiels **kies kille killst killt** killte **kiste** kit kits **kitsch** klatsch **klatsche kleist** kt **lach lache lachs lachse lachst lacht** lachte **lack lacke** lackes lacks laiche laichst laicht laichte **laie las lasche last laste** latsche least **lech** lechs **leck** lecks **leckst** leckt leica **leicht leihst** leiht leis **lest licht lichte** lichts lieh liehst lieht **lies liest lila** lisa **list liste** lsi lt **sache** sachlich sachliche **sacht sachte sack** sacke sackt sackte **sah** saht saite **schach schacht schachtel schah schal schale** schalheit **schalk schalke** schalkheit **schall** schalle **schallt** schallte schalt schalte **scheck scheich scheit** schellt schi **schicht** schichte **schick** schicke schickt schickte **schielt** schilt schlacht schlachte schlacke schlackt schlackte **schlecht** schleckt schleicht schlich **schlicht** schlichte schlick **seht sei seicht seil** seilt **seit** sek **sekt** set sh shell **sich sichel sicht** sichte **sie** siech siecht **sieh sieht** siel skat sketch **ski** st **stach stachel** stachle **stack** stahl stak **stall** stck steak **steil stich stiche** stichel stichle **sticke stiel stil stile still stille** taille takel takels takle **tal** tales **talk** talks tals **tasche** task **teich** teichs **teil** teils tel **tick** ticke ticks **tisch tische**

D.3.2 - Unusual uses task

[Below you can find all categories of submitted uses by their respective object. A list of all uses can be obtained upon request.]

A piece of paper

'crafting' 'toys & riddles' 'sanitary' 'shopping' 'containers' 'art' 'living & furniture' 'education & uni' 'currency' 'technology' 'clothes' 'decoration' 'food & cooking' 'orientation' 'communication' 'weapons & executions' 'music' 'sports' 'tickets' 'recycling' 'stimulants' 'animals' 'jewellery'

A tin can

'technology' 'music' 'art' 'containers' 'communication' 'living & furniture' 'animals' 'food & cooking' 'technology' 'jewellery' 'decoration' 'clothes' 'buildings' 'weapons & 'executions' 'sanitary' 'movement' 'sports' 'crafting'

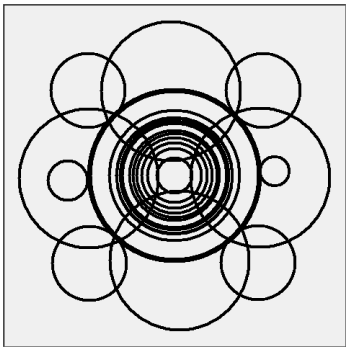
A piece of rope

'toys & riddles' 'leashes & ropes' 'technology' 'sports' 'clothes' 'jewellery' 'weapons & executions' 'containers' 'animals' 'decoration' 'living & furniture' 'stimulants'

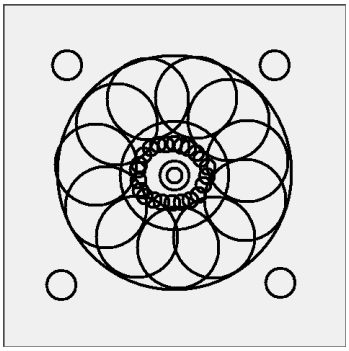
D.3.3 - Figural task

[All figure-sets submitted in the figural task.]

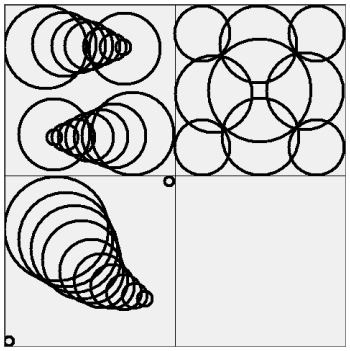
Rank: 1 / Fraction: .955



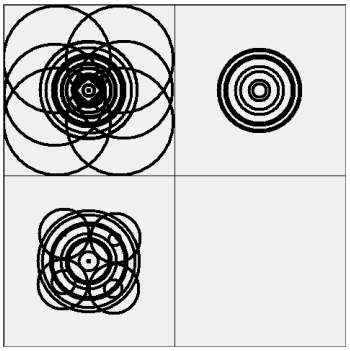
Rank: 2 / Fraction: .944



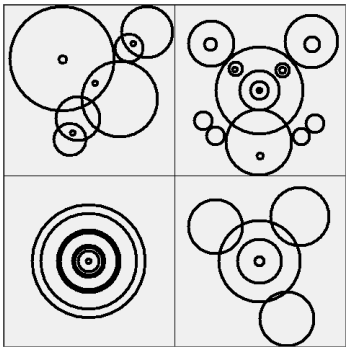
Rank: 3 / Fraction: .875



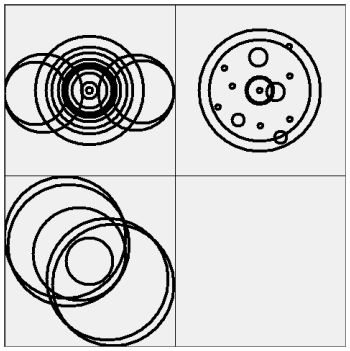
Rank: 4 / Fraction: .862



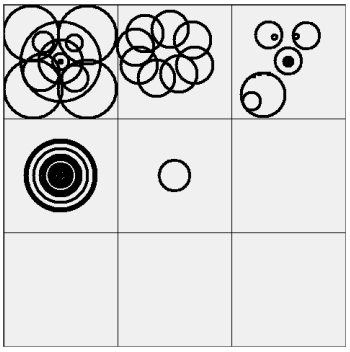
Rank: 5 / Fraction: .833



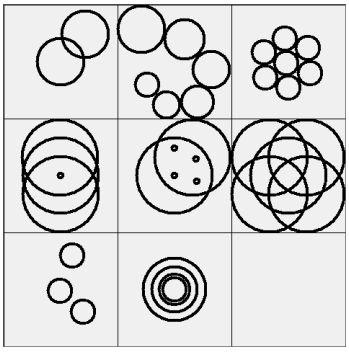
Rank: 6 / Fraction: .824



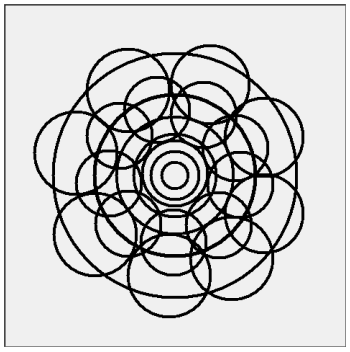
Rank: 7 / Fraction: .810



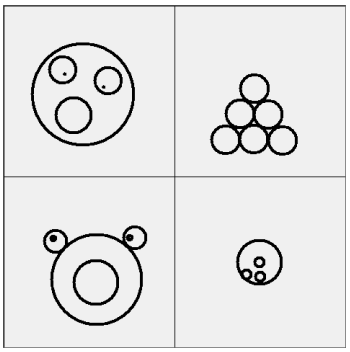
Rank: 8 / Fraction: .808



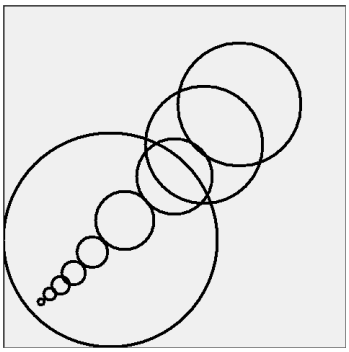
Rank: 9 / Fraction: .765



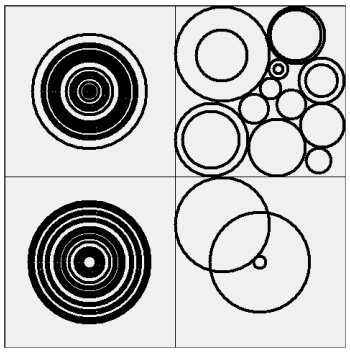
Rank: 10 / Fraction: .750



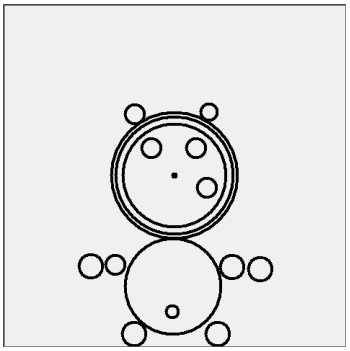
Rank: 11 / Fraction: .741



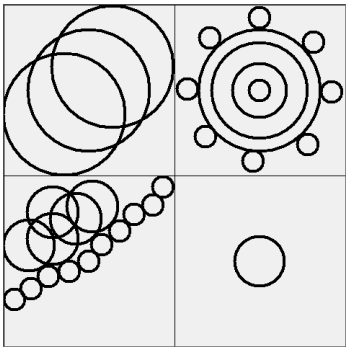
Rank: 12 / Fraction: .724



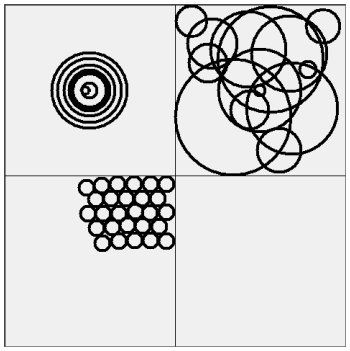
Rank: 13 / Fraction: .710



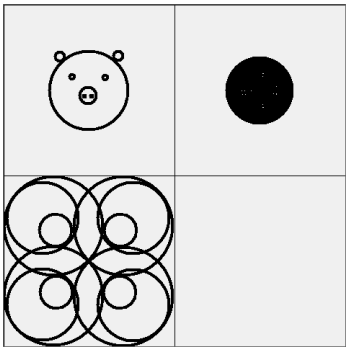
Rank: 14 / Fraction: .629



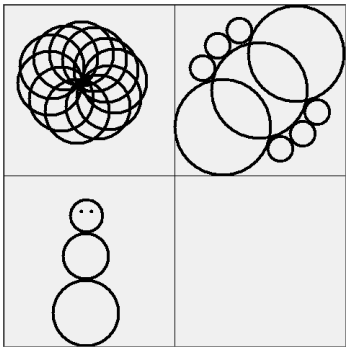
Rank: 15 / Fraction: .619



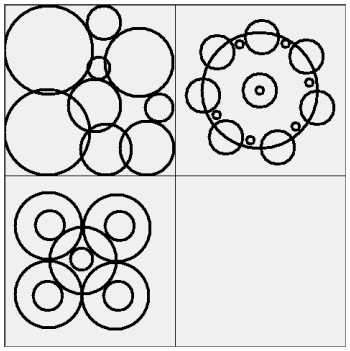
Rank: 16 / Fraction: .611



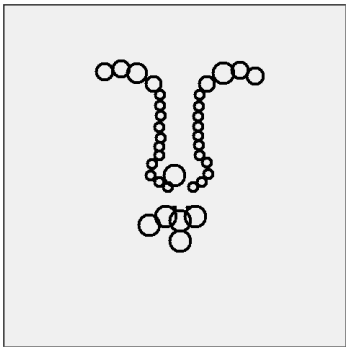
Rank: 17 / Fraction: .609



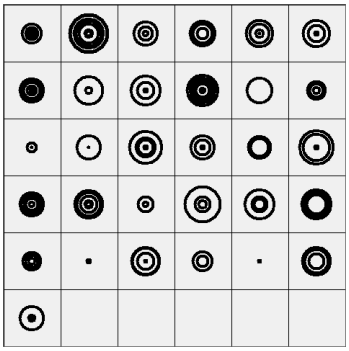
Rank: 18 / Fraction: .593



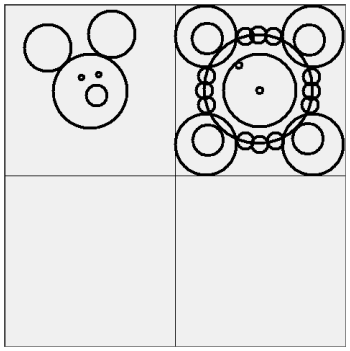
Rank: 19 / Fraction: .579



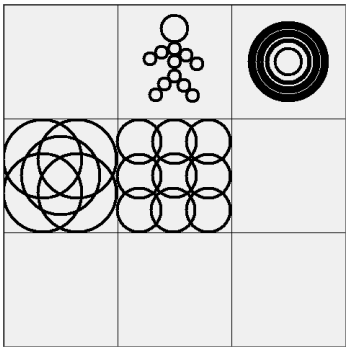
Rank: 19 / Fraction: .579



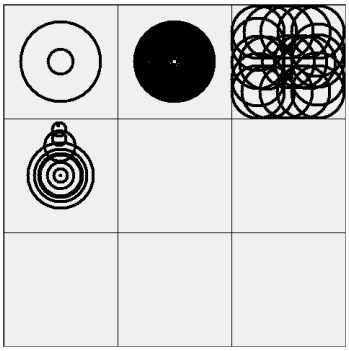
Rank 20: / Fraction: .565



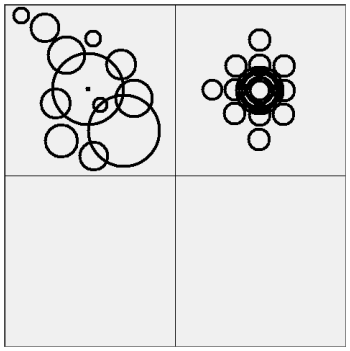
Rank: 21 / Fraction: .556



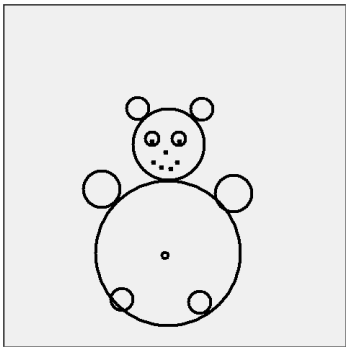
Rank: 22 / Fraction: .552



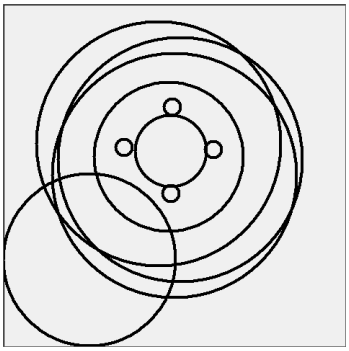
Rank: 23 / Fraction: .548



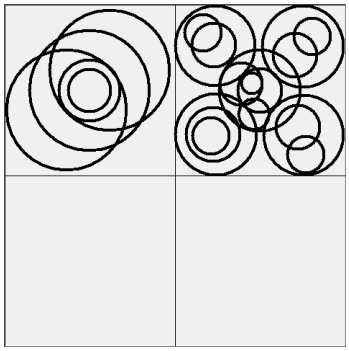
Rank: 24 / Fraction: .524



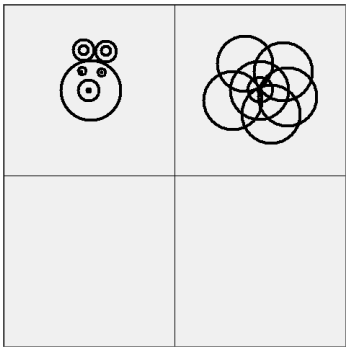
Rank: 25 / Fraction: .522



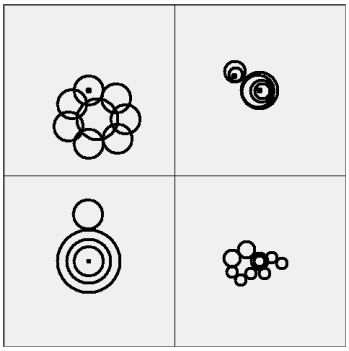
Rank: 26 / Fraction: .500



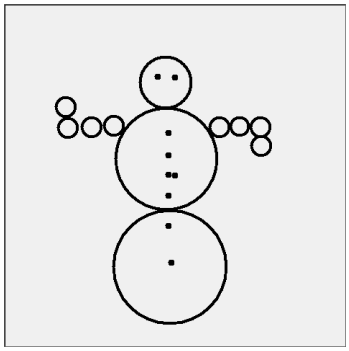
Rank: 26 / Fraction: .500



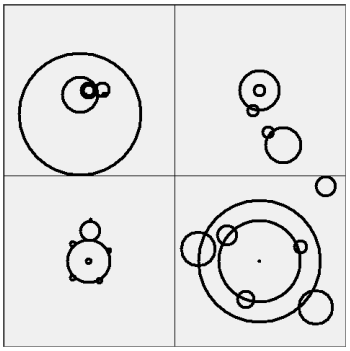
Rank: 26 / Fraction: .500



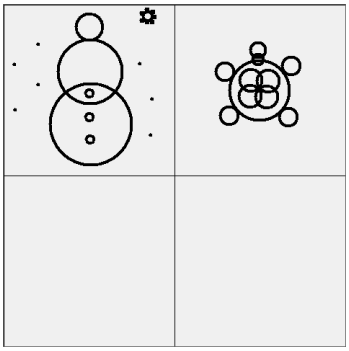
Rank: 27 / Fraction: .481



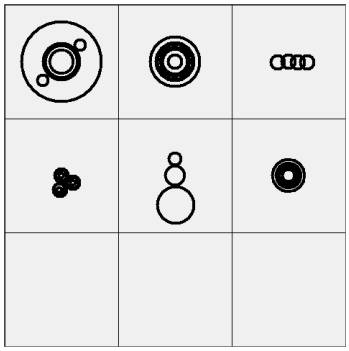
Rank: 27 / Fraction: .481



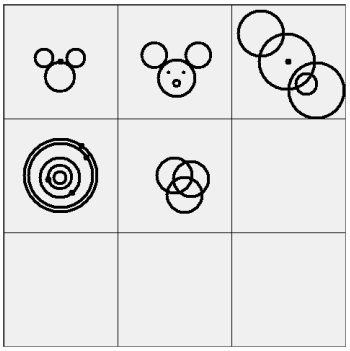
Rank: 28 / Fraction: .471



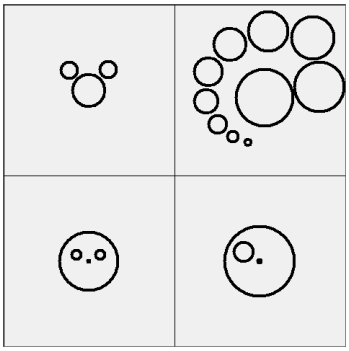
Rank: 29 / Fraction: .462



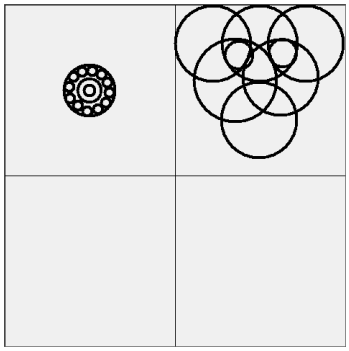
Rank: 30 / Fraction: .429



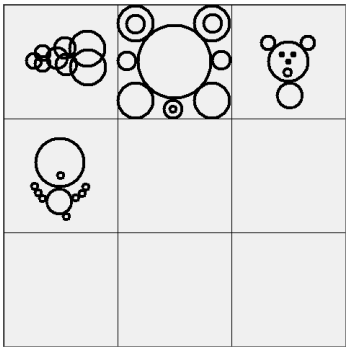
Rank: 31 / Fraction: .414



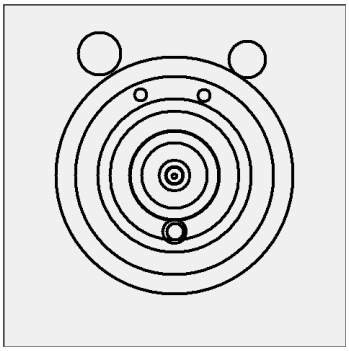
Rank: 32 / Fraction: .407



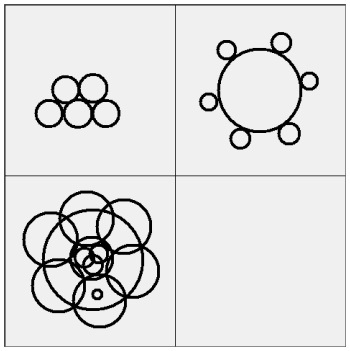
Rank: 33 / Fraction: .400



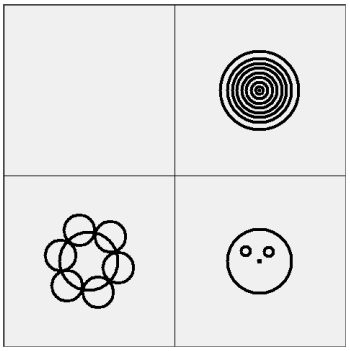
Rank: 33 / Fraction: .400



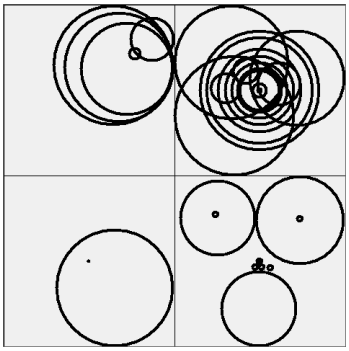
Rank: 34 / Fraction: .393



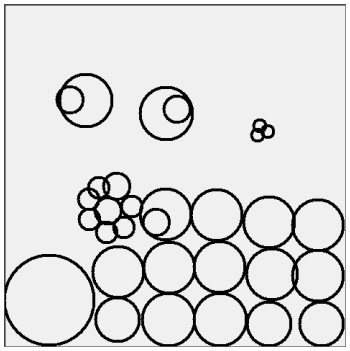
Rank: 35 / Fraction: .387



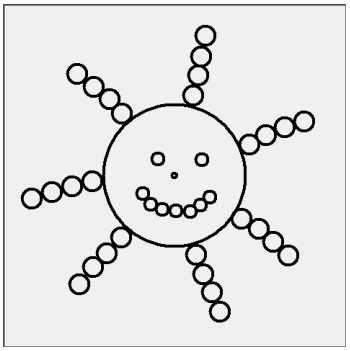
Rank: 36 / Fraction: .375



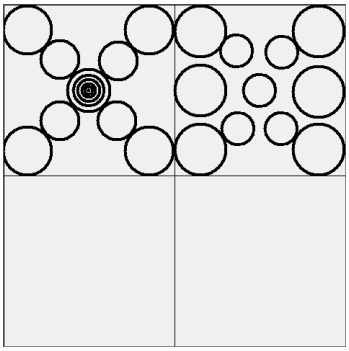
Rank: 36 / Fraction: .375



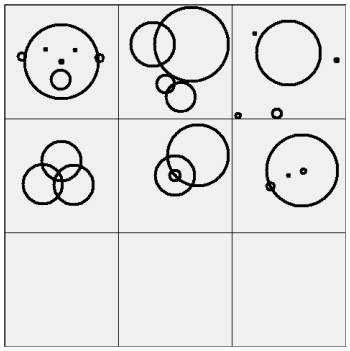
Rank: 37 / Fraction: .345



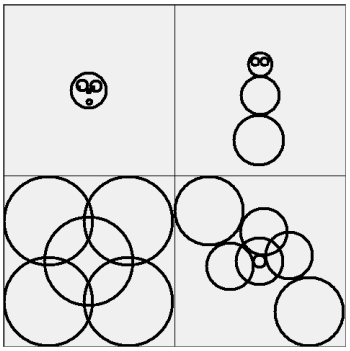
Rank: 38 / Fraction: .333



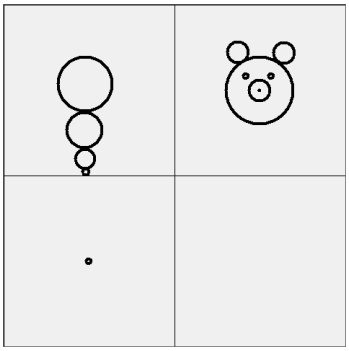
Rank: 38 / Fraction: .333



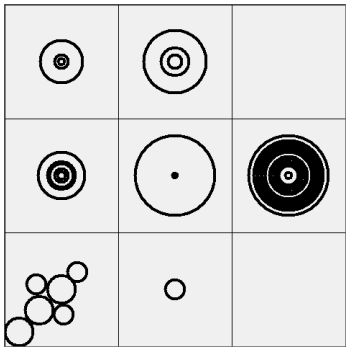
Rank: 39 / Fraction: .321



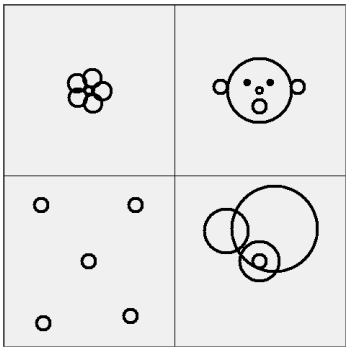
Rank: 40 / Fraction: .318



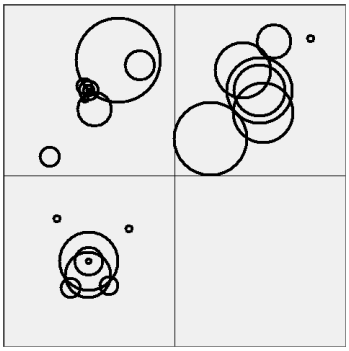
Rank: 41 / Fraction: .300



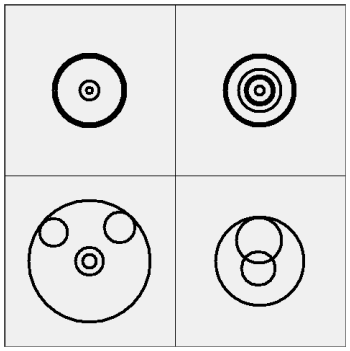
Rank: 42 / Fraction: .258



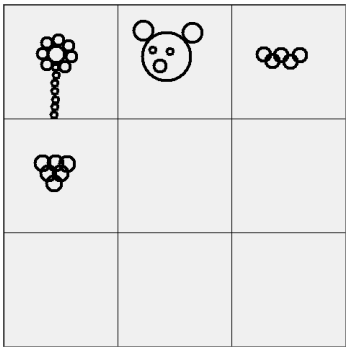
Rank: 43 / Fraction: .250



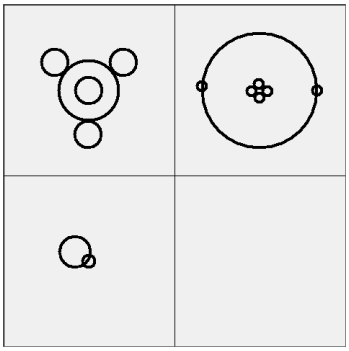
Rank: 44 / Fraction: .241



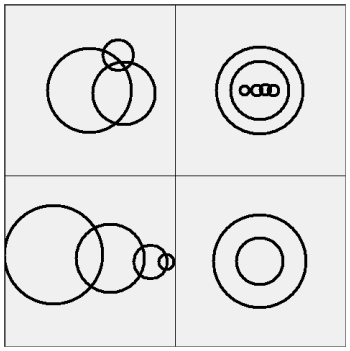
Rank: 45 / Fraction: .231



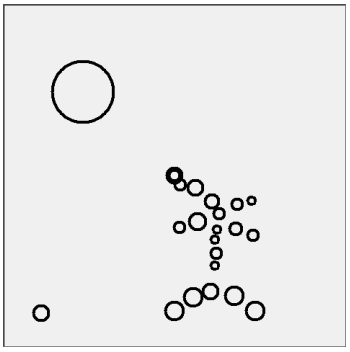
Rank: 46 / Fraction: .206



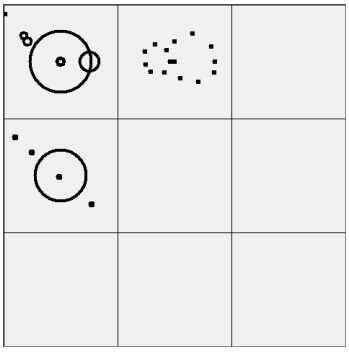
Rank: 47 / Fraction: .167



Rank: 48 / Fraction: .111



Rank: 49 / Fraction: .071



D.3.4 - Tower of London

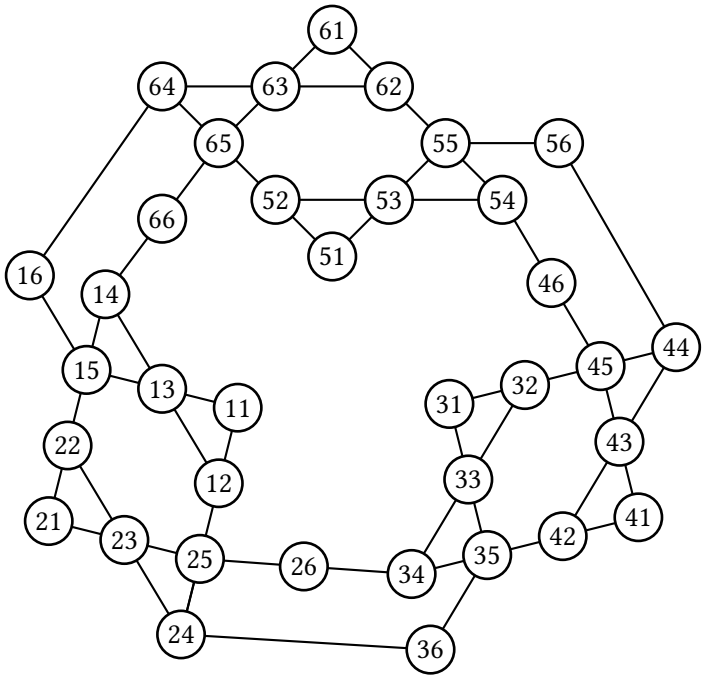
[Below you can find a description of the 8 different problems used in the Tower of London task and their properties. All numbers listed in the table correspond to the network and the problem space depicted in figure D.3.1. The network describes the entire problem space, meaning that it shows all possible moves from each of the 64 possible positions. Please see Fimbel et al. (2009) website for a detailed analysis of the task: http://tolspace.googlepages.com/.]

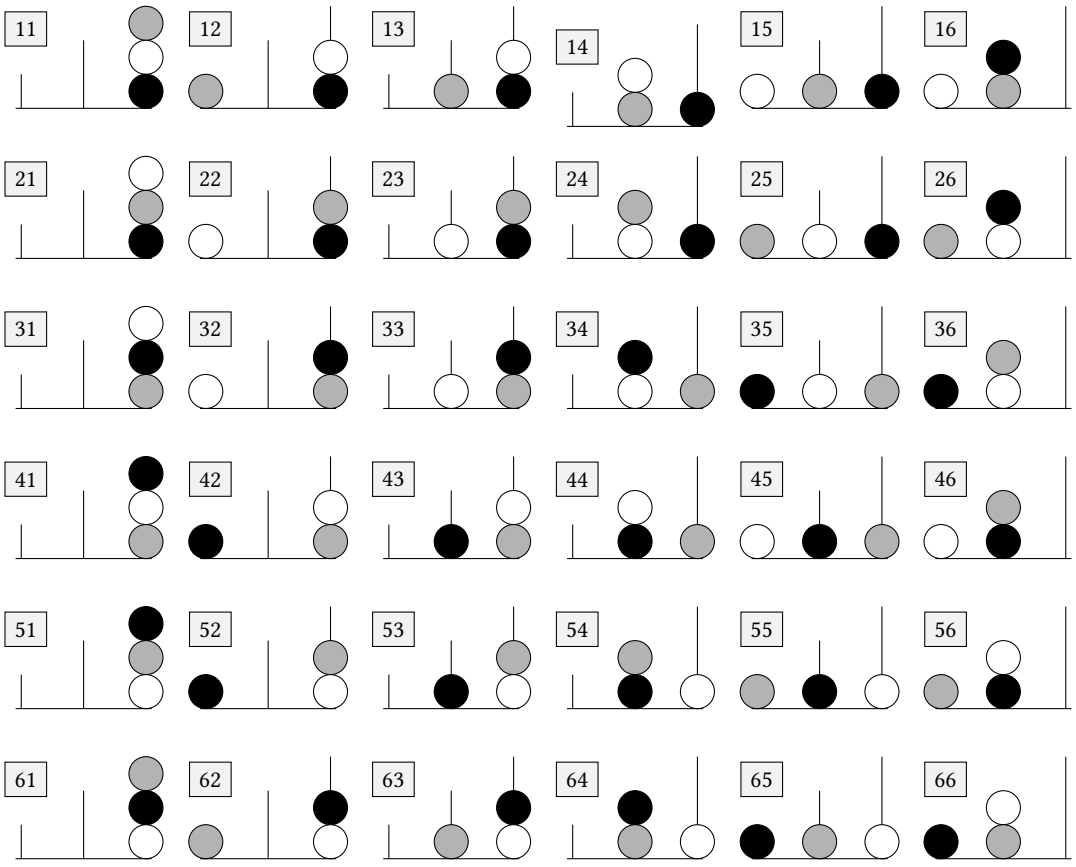
Table D.3.1: Problem Overview for the Tower of London task

Initial	Outcome	min # Moves	Solution Paths
		Standard	
15	26	4	15 - 13 - 12 - 25 - 26 15 - 22 - 23 - 25 - 26
15	33	6	15 - 13 - 12 - 25 - 26 - 34 - 33 15 - 22 - 23 - 25 - 26 - 34 - 33 15 - 22 - 23 - 24 - 36 - 35 - 33
15	46	7	15 - 14 - 66 - 65 - 52 - 53 - 54 - 46 15 - 16 - 64 - 65 - 52 - 53 - 54 - 46 15 - 16 - 64 - 63 - 62 - 55 - 54 - 46
15	45	8	15 - 13 - 12 - 25 - 26 - 34 - 33 - 32 - 45 15 - 22 - 23 - 25 - 26 - 34 - 33 - 32 - 45 15 - 22 - 23 - 24 - 36 - 35 - 33 - 32 - 45 15 - 22 - 23 - 24 - 36 - 35 - 42 - 43 - 45 15 - 14 - 66 - 65 - 52 - 53 - 54 - 46 - 45 15 - 16 - 64 - 65 - 52 - 53 - 54 - 46 - 45 15 - 16 - 64 - 63 - 62 - 55 - 54 - 46 - 45 15 - 16 - 64 - 63 - 62 - 55 - 56 - 44 - 45
		Isomorphic	
55	66	4	55 - 53 - 52 - 65 - 66 55 - 62 - 63 - 65 - 66
55	13	6	55 - 53 - 52 - 65 - 66 - 14 - 13 55 - 62 - 63 - 65 - 66 - 14 - 13 55 - 62 - 63 - 64 - 16 - 15 - 13
		Unique	
15	61	4	15 - 16 - 64 - 63 - 61
15	56	6	15 - 16 - 64 - 63 - 62 - 55 - 56

Note. The *solution paths* above correspond to the respective positions in the search space matrix depicted in figure D.3.1.

Figure D.3.1: Problem Space of the Tower of London Task by Fimbel et al. (2009).





D.3.5 - Hiking task

[**Below you can find all categories of submitted solutions to the 'hiking problem'. A list of all answers can be obtained upon request.**]

'split time between' 'work/postpone hiking' 'work while hiking' 'hike' 'extension' 'arson' 'family emergency' 'quit job' 'get help' 'delegate/ask colleague' 'other' 'do neither' 'fake an excuse' 'work over night/stay late' 'work faster/rush project' 'don't finish project' 'send pictures while working' 'ask boss' ' take a day off' 'extraterrestrial' 'combination of categories' 'damage the computer' 'fire alarm/bomb threat' 'copy project' 'ask twin' 'doppelganger' 'punch/hurt boss' 'hurt/blame Jamie' 'time travel' 'suicide' 'have lunch with Jamie' 'sneak out of work' 're-calculate time/reorganise' 'fake a robbery' 'reschedule in future' 'team hike' 're-discuss with Jamie' 'find an error in project to delay' 'make a project about hiking'

Appendix for Chapter 5

[**All subsequent analyses are referenced in the document. Brief descriptions are provided above every analysis.**]

E.1 - Additional analyses

D.1.1 - Results

Table E.1.1 presents regression results of the relationship between the Local Search task and the UUT (Models 1 and 2) and the Depth Search task and the ToL (Models 3 and 4). Model 1 has all submitted words as dependent, Model 2 only valid words. Model 3 uses Moves for the Depth Search task and ToL and Model 4 completion times.

Table E.1.1: Regression for comparisons within pairs of tasks

	LS & UUT		DS & ToL	
	(1)	(2)	(3)	(4)
# of Uses	0.509***	0.304***		
	(0.0723)	(0.0480)		
Tower of London			-0.0111	0.155*
			(0.0624)	(0.0910)
Incentives	1.163*	0.586	0.564	-1.790
	(0.615)	(0.385)	(0.538)	(3.156)
Additional Measures				
Vocabulary Test	0.100***	0.0952***	-0.0748***	-0.925***
	(0.0300)	(0.0196)	(0.0255)	(0.151)
Self-reported Risk	-0.157	-0.0721	0.159	1.020
	(0.142)	(0.0925)	(0.151)	(0.686)
Enjoy	0.510**	0.229*	-0.330**	-5.240***
	(0.199)	(0.132)	(0.153)	(0.959)
Difficulty	-0.736***	-0.412***	0.418***	5.041***
	(0.197)	(0.128)	(0.152)	(1.092)
Demographics				
Age	-0.127***	-0.0830***	-0.0287	0.384**
	(0.0259)	(0.0181)	(0.0261)	(0.155)
Student	-0.666	0.120	-0.844	-6.093
	(0.827)	(0.504)	(0.574)	(4.270)
Gender (Reference: Female, n = 230)				
Male (n = 166)	0.207	0.104	0.252	0.674
	(0.652)	(0.426)	(0.499)	(3.156)
Non-binary (n = 4)	6.530***	4.083***	-1.528	-7.982
	(2.095)	(1.072)	(2.433)	(20.45)
Education (Reference: College/A levels, n = 110)				
No formal qualifications (n = 8)	0.0135	-1.509	1.718	21.28
	(2.940)	(1.207)	(1.665)	(13.33)
Secondary school/GCSE (n = 48)	-0.409	-0.0251	-0.855	4.992
	(1.028)	(0.618)	(0.978)	(5.638)
Undergraduate degree (n = 158)	0.182	0.201	-1.218	-9.465**
	(0.721)	(0.466)	(0.784)	(3.844)
Graduate degree (n = 64)	0.439	0.137	-0.158	-7.817
	(0.880)	(0.570)	(0.911)	(4.806)
Doctorate degree (n = 8)	1.138	2.490	-1.750*	-13.34
	(2.368)	(1.662)	(0.952)	(9.005)
Order Controls	Yes	Yes	Yes	Yes
Additional Controls	Yes	No	Yes	Yes
Observations	759	759	760	760
R^2	0.427	0.380	0.195	-

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Note. Model 1 has all submitted words as dependent, Model 2 only valid words. Model 3 uses Moves for the Depth Search task and ToL and Model 4 completion times. Additional controls: employment, nationality, problem dummies, order of tasks. and control question fails.

We present correlations for Local Search and UUT across rounds in Table E.1.2. The correlation coefficient presented is the relationship between first and second execution of the task within-subject.

Table E.1.2: Within-subject correlation across rounds for Local Search and UUT

	Round 2	
	Local Search	*Unusual Uses Task*
Round 1	0.529***	0.750***
	(400)	(400)

* p < 0.1, ** p < 0.05, *** p < 0.01.

As above, we present correlations for Depth Search and the ToL across rounds in Table E.1.3. The correlations are somewhat smaller in Depth Search, indicating lower levels of robust behaviour across rounds. Howver, the coefficient is still significant.

Table E.1.3: Within-subject correlation across rounds for Depth Search and ToL

	Round 2	
	Depth Search	*Tower of London*
Round 1	0.142***	0.471***
	(400)	(400)

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table E.1.4 presents regression results for our collected survey measures. Both questionnaires are utilised with two factors as included in the table. We only find a marginal effect of the 'Ideational Behaviour Scale' with the Depth Search task, but no systematic relationship can be observed.

Table E.1.4: Regression survey measures

	Unusual Uses Task	Tower of London	Local Search	Depth Search
	(1)	(2)	(3)	(4)
Ideational Behaviour Scale (Runco et al. 2001)				
1st Factor	0.523	0.812	0.358	-6.784*
	(0.471)	(1.954)	(0.631)	(3.876)
2nd Factor	-0.175	0.297	0.0822	1.815
	(0.399)	(1.499)	(0.434)	(3.315)
Additional Controls	Yes	Yes	Yes	Yes
R^2	0.160	0.150	0.154	0.102
	(5)	(6)	(7)	(8)
Modes of Problem-Solving (Jabri, 1991)				
Associative	-0.0137	0.219	-0.286	2.675
	(0.247)	(1.113)	(0.293)	(2.379)
Bisociative	0.479	-0.313	0.121	-1.688
	(0.325)	(1.301)	(0.352)	(2.419)
Additional Controls	Yes	Yes	Yes	Yes
R^2	0.165	0.149	0.154	0.0981
Observations	390	390	390	390

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

All previous results between tasks hold when only examining the incentivised treatment. Table E.1.5 presents correlations across all tasks only for the incentivised subjects.

Table E.1.5: Within-subject correlation for incentivised treatment

	# of Uses	ToL Time	# of Correct	Depth Time
Creativity Tasks				
# of Uses	1			
ToL Time	-0.237***	1		
Network Tasks				
# of Correct	0.468***	-0.352***	1	
Depth Time	-0.196***	0.183***	-0.247***	1

* p < 0.1, ** p < 0.05, *** p < 0.01.

E.2 - Experimental instructions

[All instructions below were seen by the subjects. Each subject saw two randomly drawn tasks, for brevity we only present one in this section. . Two dotted lines represent a new screen in the programme. Below we only present the incentivised treatment. The instructions can be easily altered to remove the incentives.

The order of the tasks was randomised, we present just one order below. In this example the order is *UUT*, *Local Search*, *ToL*, *Depth Search*.]

Welcome to this study

In this study we will ask you to complete several tasks. You will be paid a fee of **£2.50 plus a bonus payment** depending on your performances for completing this study.

The study consists of **4 parts and a questionnaire**.

At the end of the study **one performance will be randomly chosen** to determine your bonus payment.

Instructions for each task will be provided as you go along. Please read all instructions carefully, answer the comprehension questions and complete the associated task.

Please click continue to proceed.

.....

.....

Instructions - Part 1 [Unusual Uses Task]

Please read through the instructions below carefully.

This part consists of 2 rounds. In each round, you have 2 minutes to come up with as many diverse uses for an object as possible.

For example

Please list as many, different and unusual uses for a rubber tyre as you can think of. Do not restrict yourself to a specific size of a tyre. You can also list uses that require several tyres. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible.

When thinking of your answers try to come up with as many diverse uses as you can. Answers are valid when they are either practicable or when their realisation is at least vaguely conceivable.

In this task for each valid use you submit, you receive a bonus of 50 pence. Thus, the more valid uses you submit, the larger your bonus will be. This means that if you submit 24 valid uses you would earn a bonus of 24 times £0.50 which equals £12.

After each use, you must submit your answer. You can submit by clicking 'Enter' or by using the submit button. You cannot change a use once it is submitted.

Once the time is up you will automatically be moved on.

Comprehension Questions

In order to move on please answer the following questions:

1) What is the main goal of the task?

- To come up with as many uses as possible

- To come up with as different uses as possible
- To come up with as many,different and unusual uses as possible

2) How much time do you have to come up with uses?

- 5 minutes
- 3 minutes
- 2 minutes

[Subjects complete 2x2 minutes of the UUT with different objects]

Instructions - Part 2 [Local Search]

Please read through the instructions below carefully.

This part consists of 2 rounds. In each round you have 2 minutes to come up with as many word associations as possible.

A valid association is a word that is connected to a pre-defined network based on dictionary definitions: Whenever two nouns appear in the same definition they are connected.

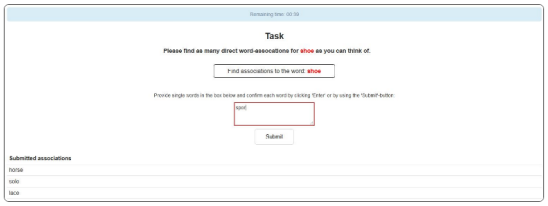
For each round, we will provide you with a specific word that you need to find associations to. Your task is to find as many words as possible that are connected to this specific word.

Please note that you can only enter words that are nouns (*What is a noun?*[**links to a definition of a noun**]) and singular. Examples are: mouse, colour or event. Terms with spaces (i.e. white noise) or nouns in plural are not valid. Please ensure the words are spelt correctly.

In this task for each correct association you submit you receive a bonus of 50 pence. Thus, the more correct associations you submit, the larger your bonus will be. This means that if you submit 24 correct associations you would earn a bonus of 24 times £0.50 which equals £12.

After each association, you must submit your answer. You can submit by pressing 'Enter' or by clicking the 'Submit' button. You cannot change an answer once it is submitted. You can find an example of the task screen below.

Once the time is up you will automatically be moved on.



Comprehension Questions

In order to move on please answer the following questions:

1) How much time do you have?

- 3 minutes
- 2 minutes
- 4 minutes

2) What is the main task?

- Copy words in a given time
- Do a mathematical calculation
- Provide as many associations to a word

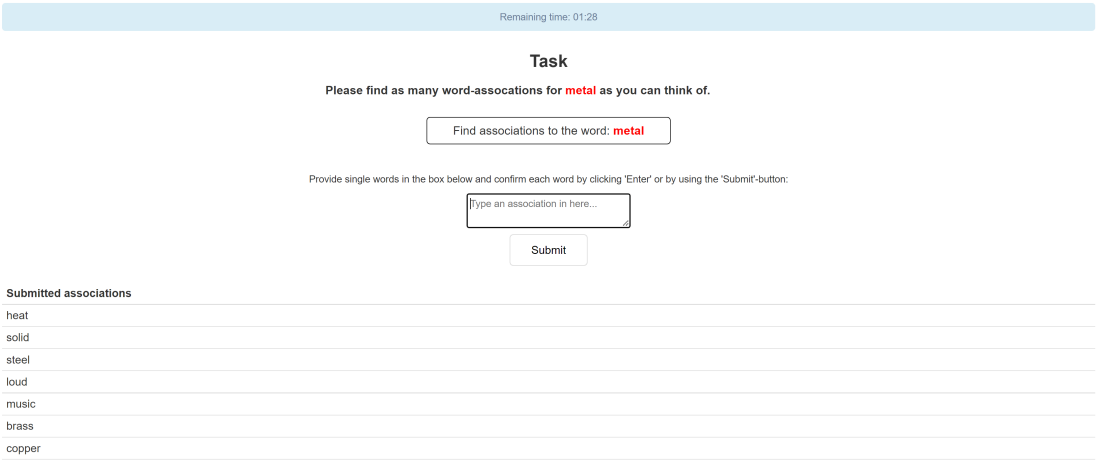
3) What is the role of the network?

- It does not have a special role
- It only provides a general direction of associations
- It defines valid associations

4) Which words can you enter?

- Any word can be entered
- Only nouns that are singular
- Only adjectives count as words

[Subjects complete 2x2 minutes of the Local Search task with different words]



Instructions - Part 3 [Tower of London]

Please read through the instructions below carefully

You are asked to play this part for 2 rounds. You have 60 seconds to complete each round.

On the screen there are 3 balls and 3 rods, which differ in length, 1 to 3, from left to right (see image below). The smallest rod can hold 1 ball, the middle can hold 2, and the largest rod can hold 3 balls at one time. In each round, you must move the 3 balls positioned on different rods to adjust a starting position to match an outcome position.

You can move the top ball from each rod to any of the other rods as long as the rod is not full. You can move the ball to any rod and you do not have to move the ball to a neighbouring rod.

Your task is to reach the outcome position **as fast as possible**.

In this task for each second left after you have finished, you receive a bonus of 20 pence. Thus, the faster you complete the task, the larger your bonus will be. So, imagine that you could solve the task in 0 seconds, then you would earn a bonus of 60 times £0.20 which equals £12.

Below you find an example of the task. You can move balls by clicking on the chosen ball and then the rod you wish to move it to. On the next page you will be able to practice the task.

Once you finish a round or run out of time you will be able to move on.

Comprehension Questions

In order to move on please answer the following questions:

1) What is the goal in each round?

- Reaching the outcome position as fast as possible
- Reaching the outcome position using the most moves
- I just have to reach the outcome position

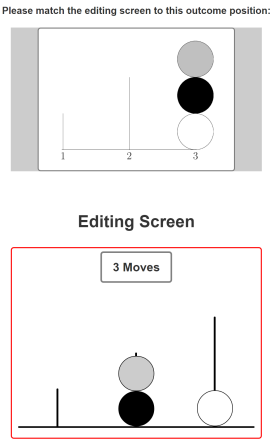
2) How much time do you have for each round?

- 30 seconds
- 120 seconds
- 60 seconds

3) Can you only move a ball to its directly neighbouring rod?

- Yes
- No

[Subjects complete 1x practice and 2x60 seconds of the Tower of London with different paths]



Instructions - Part 4 [Depth Search]

Please read through the instructions below carefully

This part consists of 2 rounds. In each round, you have 120 seconds to use connections to move from a start word to an end word. All connections are drawn from a pre-defined network based on dictionary definitions: Whenever two nouns appear in the same definition they are connected.

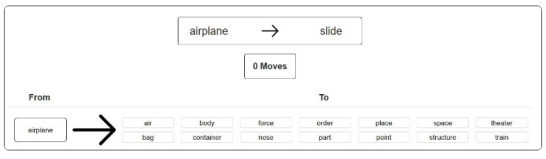
Your task is to reach the end word **as fast as possible.**

In this task for each second left after you have finished, you receive a bonus of 10 pence. Thus, the faster you complete the task, the larger your bonus will be. So, imagine that you could solve the task in 0 seconds, then you would earn a bonus of 120 times £0.10 which equals £12.

You are always provided with all possible next connections so your task is to find the connection that leads you towards the end word.

Below you find an example of the task screen. On the top of the screen you see the start and end word. Lower down the screen, on the left-hand side, you can see your current word (which in the example coincides with the start word) and on the right hand side all connections in alphabetical order. To make a move, just click on the connection you want to move to. On the next page you will be able to practice the task.

Once you finish a round or run out of time you will be able to move on.



Comprehension Questions

In order to move on please answer the following questions:

1) What is the main goal of the task?

- Avoid arriving at presented words
- Reach the end word as fast as possible
- Find the longest path between words

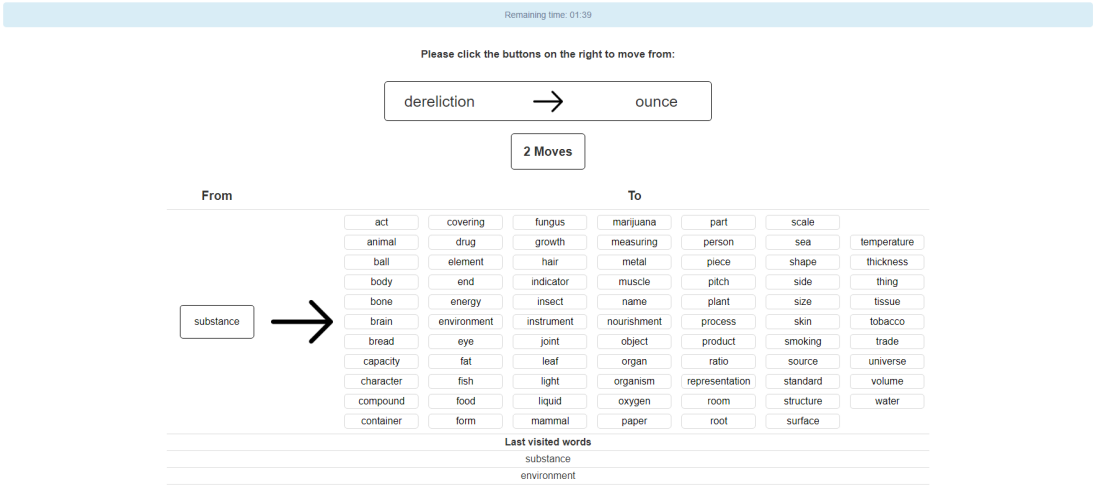
2) How much time do you have per problem?

- 90 seconds
- 120 seconds
- 60 seconds

3) How can you identify the next connections?

- I am provided with all direct connections
- I will have to type it in
- I am provided with a random selection of connections

[**Subjects complete 1x Practice and 2x120 seconds of the Depth Search task with different paths in distinct sub-networks**]



Questionnaire - Part 1/4 Please read through the instructions below carefully.

This questionnaire consists of about 60 trials, in each of which you will see a string of letters. Your task is to decide whether this is an existing English word or not. If you think it is an existing English word, click on “yes”, and if you think it is not an existing English word, click on “no”.

If you think that the word exists, even though you don’t know its exact meaning, you may still respond “yes”. But if you think this is not an existing word, you should respond “no”.

In this questionnaire, we use British English rather than American English spelling. For example: “realise” instead of “realize”; “colour” instead of “color”, and so on.

You can take as much time as you like for each decision but overall this part of the experiment should take about 5 minutes.

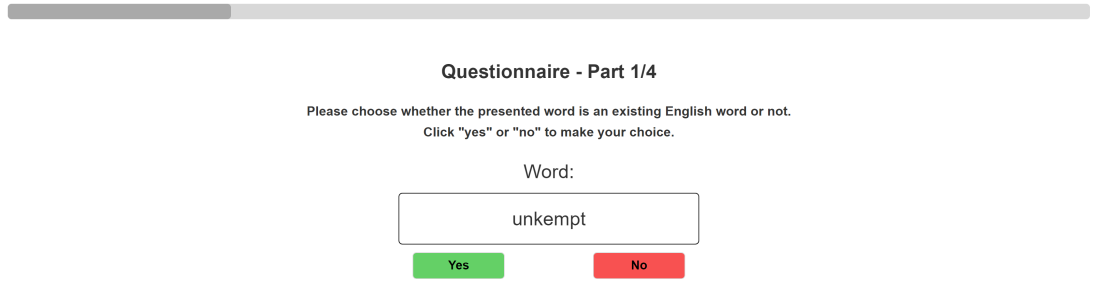
If everything is clear, you can now answer the comprehension question and then start the questionnaire.

Comprehension Questions

In order to move on please answer the following questions:

1) What is the main goal of the task?

- To spell check English words
- To say whether a word exists in English or not
- To create new English words



Questionnaire - Part 2/4 Please answer each statement by indicating "**I am a person who enjoys...**" ranging from (**not at all**) to (**very much**). [**7-item Likert scale, 1 - Not at all; 7 - Very much**]

1) ...adhering to the commonly established rules of my area of work.

2) ...following well-trodden ways and generally accepts methods for solving problems.

3) ...being methodical and consistent in the way I tackle problems.

4) ...paying strict regard to the sequence of steps needed for the completion of a job.

5) ...adhering to the well-known techniques, methods and procedures of my area of work.

6) ...being strict on the production of results, as and when required.

7) ...accepting readily the usual and generally proven methods of solution.

8) ...being precise and exact about production of results and reports.

9) ...adhering carefully to the standards of my area of work.

10) ...being fully aware beforehand of the sequence of steps required in solving problems.

11) ...being confronted with a maze of ideas which may, or may not, lead me somewhere.

12) ...pursuing a problem, particularly if it takes me into areas I don't know much about.

13) ...linking ideas which stem from more than one area of investigation.

14) ...being fully occupied with what appear to be novel methods of solution.

15) ...making unusual connections about ideas even if they are trivial.

16) ...searching for novel approaches not required at the time.

17) ...struggling to make connections between apparently unrelated ideas.

18) ...spending time tracing relationships between disparate areas of work.

19) ...being 'caught up' by more than one concept, method or solution.

Questionnaire - Part 2/4 Please indicate for each statement how often from (never) to (very often) a statement occurs to you. [**7-item Likert scale, 1 - Not at all; 7 - Very much**]

1) I have many wild ideas.

2) I think about ideas more often than most people.

3) I often get excited by my own new ideas.

4) I come up with a lot of ideas or solutions to problems.

5) I come up with an idea or solution other people have never thought of.

6) I like to play around with ideas for the fun of it.

7) It is important to be able to think of bizarre and wild possibilities.

8) I would rate myself highly in being able to come up with ideas.

9) I have always been an active thinker — I have lots of ideas.

10) I enjoy having leeway in the things I do and room to make up my own mind.

11) My ideas are often considered "impractical" or even "wild."

12) I would take a college course which was based on original ideas.

13) I am able to think about things intensely for many hours.

14) Sometimes I get so interested in a new idea that I forget about other things that I should be doing.

15) I often have trouble sleeping at night, because so many ideas keep popping into my head.

16) When writing papers or talking to people, I often have trouble staying with one topic because I think of so many things to write or say.

17) I often find that one of my ideas has led me to other ideas that have led me to other ideas, and I end up with an idea and do not know where it came from.

18) Some people might think me scatterbrained or absentminded because I think about a variety of things at once.

19) I try to exercise my mind by thinking things through.

20) I am able to think up answers to problems that haven't already been figured out.

21) I am good at combining ideas in ways that others have not tried.

22) Friends ask me to help them think of ideas and solutions.

23) I have ideas about new inventions or about how to improve things.

Questionnaire - Part 2/4

1) Please tell me, in general, how willing or unwilling you are to take risks? Please use a scale from 0 to 10, where 0 means 'Completely unwilling to take risks' and a 10 means you are 'Very willing to take risks'. Pick a response from 0 to 10 to indicate below where you fall on this scale.[**11-item Likert scale, 0 - Completely unwilling to take risks; 7 - Very willing to take risks**]

1) What is your age? [**Numeric input**]

2) What is your gender? [**Text input**]

3) What is your ethnicity? [**Text input**]

4) What is your first language? [**Text input**]

5) What is the highest degree or level of school you have completed?

- No formal qualifications
- Secondary school/GCSE
- College/A levels
- Undergraduate degree (BA/BSc/other)
- Graduate degree (MA/MSc/MPhil/other)
- Doctorate degree (PhD/MD/other)
- Prefer not to say

[**For the four questions below. 7-item Likert scale, 0 - Not at all; 7 - Very much**]

How difficult did you find the part where you move from a start word to an end word?

Did you enjoy the part where you move from a start word to an end word?

How difficult did you find the part where you entered associations for a specific word?

Did you enjoy the part where you entered associations for a specific word?

.....

.....

[We then provide feedback about all performances and bonus payments. For brevity we do not present this screen in this appendix]

.....

.....

E.3 - Creativity tasks

E.3.1 - Local Search task

[Below we present all possible word associations in the *Local Search* task. For details on the words named by subjects please contact the authors.]

Word: metal

argent asteroid amount application animal action artifact article appearance attachment adornment air activity arrangement art affiliation armor armour alpenstock aglet aiglet ammunition area anvil argyle argyll apparatus appliance andiron armilla artillery awl armpit angle amalgam alchemist assayer analyst ash alloy asphalt bolt bone bent base bimetal bowl ball bullet belt box bonding bank basket bob blackjack basketball block battery babbitting bar bird bond ballpoint ballpen band beam backsaw bandsaw barrel baton bell bit bitumastic blade bicycle body bonnet bracelet boiler bullion brazier brasier brigandine buffer breech breechblock bushing brand brightness bloodstream badge bailiff bimetallism concrete cloisonne comal charcoal color current cutting center compound camp chelate chemical carbonyl ceramic conversion clip casting crutch cooking circle circuit chase cartwheel coating covering charge coupling construction combat competition cock coat cat carriage cage cosh cupel coop can club candle calk calkin canister cannister carbon cap casing cartridge caster carabiner chain chassis clapper clarinet circlet chimney cleat cleats coil column collet cockpit conductor component computer container corset cord crystal cramp crosshead cowl cringle cuff cutter cup cylinder cloth cowcatcher cowling chains cabasset coign coigne character currency coffee clay cavity cockfight crevice coal constable conductivity coin calcination chelation corrosion centimeter century device dash decoration dancing drink decalcomania design dog drum die disk disc diestock door dixie drawbar duct dulcimer drug drumstick doorknocker distance dimension diameter dancer dentist debaser degrader decalescence deterioration disease dross exchange enclosure extraction examination end echo element eyelet electrode edge enamel engine electroplate envelope emblem equipment earth elasticity electricity elixir engineer fluid form food firing flux filling frame foliation friction fatigue fabrication forging fire field fighting firedog fender fan fabric factory firearm fence ferrule fireplace fillet film fireguard fastener filter finger fingerboard flue fishplate flagon forge flashing foil front foundry framework furnace fret fretwork flour founder fume ferroconcrete game gold group gun golf galvanization galvanisation guard gamecock gas gong grommet grummet glass ground gridiron garment gaff gear gearbox globe glockenspiel gem grinder gurney gusset grate graver grid grill griddle grille grillwork guest gravity glaze gravy gunner gallon germanium garnet gemstone home hook handling holding head hanging hammer hand hammering horn hood horse hoop hearth handcuff hacksaw handlock helmet handsaw handspike heraldry heating heat hardware heel handle hunter honeycomb holder hole hilt hinge horseshoe hydrofoil hardness household hydrogen imitation iron interaction instrument incorporation interference image ironware ingot irons implement ivory ink ion inch impurity insulator joint junction jackstones jack jacket joist jewel jewelry jewellery kind key knucks knuckles karabiner keyring kickstand kern kettle knocker light living leading liquid lame locomotive ligature loop leg line lock lining lithography lead lid lamp lathe lamppost lattice latticework lacquer lacquerware leaf link ligament length log lever lunula lockbox letterpress lump lithium leather matte mass medium material mean metallic mineral mechanism mesh move mount mixture method music metalworking metalwork measuring miller magazine machine mail metalworks manacle match metalware missile mouth morion microwave motorcycle mirror mouthpiece manner mercury member metallurgy magnetohydrodynamics membership majorette melter metalhead metallurgist metalworker money tincture rock sun orbit staff point plastic sheath shoelace ribbon steel silver vessel water teeth test sheet net player weight string piece surface wood pen transfer paper reinforcement saw wheel rod sound rein weapon part wristwatch wrist pot ring track status rank unit standard value ratio strip monoxide page type time spoke vehicle wire tea worker product spring sailing rope striker side projection sole shoe pole thing object transport piston shackle trophy winner soldier wedge printer oxide result temperature process poisoning state oxidation shipping storage thread zither ornamental presence quality person phenomenon slowing rate scum pottery protection substance tooth work production stress screen plate workplace roof strength support preciousness superiority velocity tube use zinc rusting percussion utensil shell train set tool stretcher partition science pattern sphere saddle pommel vent smoke pipe watch shape necklace way one project printing reed saxophone stone post streetlight tumbler ornament strongbox valuable sulfide ore yarn place table technology plasma study wax refining solid monazite nail screw nut nugget phosphorus nitrogen nitride platinum osmium plumb oil piping press pressing perpendicular pick sugar purification timber upright position poker plating plastron pig pricket pin sheathing thickness shield pilot pan penicillamine patina setting pave plectrum piton plectron plexor pleximeter plessimeter pipage planchet plummet porringer round sonar ping plater potassium tin pewter quoin rail rapper radiator pair rings rule ruler refinement rubidium sport shot metal salt sculpture stabile stopping slug rigging spar screening shielding stay salamander scale projectile sap stove stovepipe smithy urn spigot samovar sand sandbox scarf scauper scorper silicon semiconductor scriber scribe shaper shoetree shim shear snips slat spline runner snow ski spike stake speculum rider stirrup straightness straightedge sword spread swage scrapheap sliver splinter workman textile shearer smith sodium strontium solder slag scoria toe tap tinker tongue slot token system tape plural tinsnips tapeline while sewing thimble tinsel thruster thermometer thermocouple voltage thermojunction

trampoline triangle trivet trowel order tension turnbuckle temper toughness tarnish tip sign office tipstaff tapper troy upset xylophone resonator vibrato vibraphone vibraharp vibes pressure welding path waveguide watchband watchstrap wristband screwdriver woodscrew workpiece weld windowpane wall window welder zarf

Word: skin

anterior adult arch attribute animal act air attempt arrangement affusion absence anaspid arthropod argasid armor ant aurochs antelope armadillo antenna abaya accelerator agal actuator arm appendage artery atrium attic ax axe arrow arrowhead aventail aura atmosphere aureole anteriority authority acetabulum angle artichoke army amphisbaena adviser ageratum andryala agueweed achillea acephalia acephalism acephaly asynclitism abscess alopecia back bolt bottom brachycephalic bone bowing base beetle bust boss bovid bull ball beanball beaner bullet basket beheading bob baseball basketball bow block baby bridle bird blackpoll beak baldpate bladdernose bulldog bluebottle bison bullhead bluehead boxfish branch balaclava band beam bark bedstead bearing basinet board boat body braid bollock bullfight blowhole blood brain bullethead bean bonce bemusement breed birth broccoli butter baldhead baldy brainworker butler bud borecole boneset banksia brodiaea beebalm bract balanitis balanoposthitis bighead baldness chicken capitate center composite cephalopod cetacean crustacean cause clout comb contraction covering charge crime class cephalometry ceremony carrot congregation crest cephalaspid crampfish cassowary carrion cockscomb coxcomb cockateel cockatiel clypeus calf cachalot collie corgi cat claw chlamyphore capuchin catfish checkrein club cabinet cam camail cap casque capote chanfron chamfron chest chessman chador chadar chaddar chuddar circlet clothing cloak coil college collider constituent computer container cord crook crown cushion cylinder cloth clubhead chapeau cervix cavity caul coxa chin condition concert cauliflower cabbage cole cos chicory clan cluster community colony country city comet cosmos chancellor capo chieftain chief colewort chrysanthemum cosmea catananche cornflower capeweed capitulum coin century clinocephaly clinocephalism cephalalgia cobalt calfskin dolichocephalic discharge device denial decapitation development decoration dilation demonstration danger dinosaur duck dog drum disk disc doll dome doornail drug drumstick defect distance diameter disbelief department datum divot diplomat dean don dahlia disease dengue evergreen empty execution expression elevation elongation examination end embryo eagle emperor elbow edge eye engine emblem ear equilibrium enarthrosis etiquette endive enemy empire educator edelweiss edema firm fluid fitting female flip feeling form feat formation foot feather flex friction fetometry foetometry fetus fowl foretop fin fish feeler frontstall face fighter fastener feature foliage foremast fusee fuzee front fruit forehead foreskin femur family foam froth father floret flower fleabane feverfew freshwater general group golf government gudgeon griffon gastropod gnu goat grenadier goujon grouper goby gas gurnard ground garment gaff gear governor grass garland glory grain gloriole genu growth gryphon griffin goldilocks gayfeather gosmore goldenrod groundsel goldfields high hip human hundred hook headshot header headshake headshaking headstand headship hat headlock water tail side part mouth shield pair insect robe toe wind hair kaffiyeh term tool handle tip mail hood neck light indication saint quality hipbone joint leaf heart plant serpent mythology sap herb native monster medicine labor presentation loss wool knee sign reverence submission shame hammer sculpture shoulder person victim inclination sac seal jaw mane horn sculpin spine male plate limb topmast opponent profession problem manservant servant household wine table tree shrublet seed inflammation top wrist shape metacarpus mollusk tentacle nail measurement skin organ order ray parrot whale spermaceti oil sperm ruff sheepdog pink monkey monk horse rein position state overcoat woman metal particle wreath victory vertex headdress weather organism rest vertebrate membrane lettuce root head salad university syndicate tribe weed pain muscle reaction turf piece skimming tropic orientation structure system sphere monarchy star platform tapering sucker lion shell ox sea mast spar sail panicle raceme shot hand photograph net soccer machine wagon help wrestling procession stream source viewer mass user individual pressure one school stimulation pus projection size height length liquid ship toilet representation thought pattern pole science passage text line inactivity matter subject movement pointer mark juncture role word percussion instrument kettledrum marimba ungulate protuberance shark hammerhead holocephalan holocephalian prey hooter hawk leather place hydra pike halberd helmet headpiece headrest kerchief headscarf headstall pillar stone statue herm headband panel headboard headgear harness hobnail halo standing surface headful information quantity margin hairline headache loaf sausage meat tongue pig headcheese savage trophy headhunter headman horseweed hortensia helianthus horsemint iron project ixodid point inion iceberg name region purple hue ironweed return vein jugular weight jerk scale jewfish triangle scarf knight kale kail queen king kingdom livelong lizardfish lid laurel loft meter stick match linstock stalk stem leopardbane time latency job minister ruler nation right monarch muzzle mudcat mallet sport masthead middle scalp mohawk thing metonymy organization title matriarchy matriarchate pond mill wheel millpond scorpion manticore mantichora manticora mantiger matriarch materfamilias mayor mayweed matchweed mistflower matchbush muskwood milkweed pendant molle megacephaly macrocephaly megalocephaly microcephalus microcephaly nod numbfish neb nailhead nimbus nucleus noodle noggin mind noddle napa nanocephaly overhead obeisance osteostracan owl ornithomimid ostrich notion reference truth reality sand missile rocket payload heat ogive obliquity skull occiput owlclaws orpin orpine spasm heel opisthotonos premier pick poll portfolio placoderm inch pachycephalosaur pachycephalosaurus reptile plesiosaur plesiosaurus puffbird pochard proturan mean spark plug pichiciago pichiciego pinche pickax pillow pickaxe pin pinhead rubber reflex plexor plessor percussor hole poncho sweater pullover puppet track positioner style ponytail precava pate protocol patriarchy patriarchate pantryman lance horseman picador patriarch paterfamilias officer president prexy prior pope pontiff pyrethrum umbel primrose primula platan phalacrosis queue iridescence paradise riflebird ringtail redhead weevil rostrum rattail object remora linen woolen sling rebozo teeth row rake

rivet variety radicchio romaine republic subshrub rabbitweed smash superior sovereign opening scapular somersault summersault summerset somersaulting sclaff thread shank screw phrase relation subordination shovelhead pectoral skate sauropod sticktight termite soldier snout snakefish suckerfish stargazer scorpaenoid scorpionfish slipover scapulary warmth secretary hub motion shaft setscrew humerus scapula shawl string snare skullcap snowball number sphinx spear spearhead spearpoint stature skeleton sinciput vertebra splenius seek savoy salsify scorzonera meeting summit rock supremacist skinhead stooper sunray sneezewort safflower succory snakeroot snakeweed sunflower sycamore priest tonsure member tack tape mechanism transport torpedo wing titmouse tomtit process wattle pheasant tragopan telsontail trunk toadfish thorax trunkfish testiere turban tabor tabour vessel topgallant tudung round thrift torso thoroughwort tidytips tansy toetoe teasel torticollis univalve undersecretary vulture ventail veil unit van vanguard hemisphere songbird waxwing wryneck wisent wildebeest wig witloof work power waterpower

Word: head

astringent avocado adult amphibian attribute abdominoplasty application animal action act appearance attack accumulation analysis acupuncture abnormality art absence anole arthropod antelope area araroba antiperspirant artery achromia achromasia account ability anus agnail aubergine apple aguacate anjou albino abrasion aspergillosis albinism algidity acne anemia acanthosis antifreeze back blond blonde bone buff blistering base break browning bite burn burning bag breakdown browse branding bird beak buffalo butterfish bonito branch blowfish band bark bearing bit border body brake bulla brand blood blemish birthmark blubber blackhead bile blister bleb birth blackheart blackamoor blastomycosis bedsore bruise bilirubin benzoin benjamin cut cosmetic chicken cherry color colour crack cancroid cell contact contusion cookery care circulation cosmetology cauterization cryosurgery cupping covering collection crest chimaera chameleon chamaeleon cobra coat cuticle calf cattle chamois coronet colugo chrysarobin canoe chest cephalexin cigar cigarette ciprofloxacin clothing coldcream container cream cup cloth cataplasm corn complexion callosity cheek clay comedo corium cartilage cortex canthus cranium clitoris cyst creepiness condition chafing crookneck cocozelle cherimoya cherimolla cataract child craftsman cumquat callus chromoblastomycosis coccidioidomycosis coccidiomycosis cancer carcinoma costiasis cyanosis cowpox cheilitis cellulitis cicatrix cicatrice clavus ceratin calfskin chammy concealer darkening demulcent distress development depilation dash deposit dilation damage dermabrasion diving doctor display dressing design dinosaur deodorant deodourant dermatome dapsone developer drug drip delicacy dryness depigmentation dewlap dermis derma dermatology degeneration dander dandruff dermatologist desquamation deterioration diaphoresis disorder dermatoglyphic disease dermatomycosis dermatophytosis dermatitis dermatosis dermatomyositis dracunculiasis dye exterior emollient escape exposure ending epilation enlargement elevation example erection ectoblast ectoderm exoderm edmontosaurus elephant eel eye embrocation effect elasticity exuviae epidermis eschar ear epicanthus eyelid epicranium extremity excoriation eruption eggplant exfoliation efflorescence epicarp exocarp ecchymosis ecdysis elastosis erythroderma erythema eczema erysipelas exanthem exanthema edema firm fluid fat fatty fell feeling form food failure force flare formation foot feather fold frame friction function fetus fleece filaria flea fin fish filefish face finger fruit focus fingerprint fissure flesh freckle follicle foreskin fingernail fingertip faculty fever fugu family flake fungus fermentation furuncle furunculosis formication freshwater gooseflesh goosebump germ goat garment gelatin grape gall gum gland genu growth grapefruit guava genip ginep goatskin gelatine glycerin glycerogelatin glycerogel hyaline human hybrid head horripilation hanging hand horn herpes hood hair hide hookworm hoof heart heat hangnail healing hidrosis hypothyroidism hypoadrenalism hypoadrenocorticism hickey hemorrhage hyperpigmentation hypopigmentation hives hyalin humectant hepatoflavin item interior impression identification instrument injection inflammation irritation injury inhalation insect implement infection integument intestine itch intertrigo infestation ichthyosis impetigo icterus injectant joint jacket jaundice juice kind kayak kumquat keratoderma keratoacanthoma keratodermia keratosis kraurosis keratin light liquid life location loss leap leg line liposuction layer lizard lip louse lappet lamb lemur limb lid liniment lotion lukewarmness luridness lividity lividness lentigo lung lymph lump loather lichen leprosy lesion lupus livedo leukoderma lack lambskin leather lactoflavin marine minute mass material mammalian mark medicine massage medication molter moulter mite mammal mole meloid mosquito makeup membrane mouth margin melanoderma muscle mummy macula macule milium melanocyte marrow melatonin mutation meal morello medlar mandarin mango mamoncillo molt molting moult moulting myxedema myxoedema molluscum melanosis melanism melanoma mange melanin name neck nail needle node nose nevus nerve nodule nectarine nettle neurofibromatosis neurofibroma neoplasm nitrogen orange opening organism obstruction order oxen ointment organ oil odor overgrowth overactivity onchocerciasis ovoflavin plum part patient protozoan person plant process pressure preparation prostration practice pattern petrissage puncture presence plethodont plug pachyderm pastern plectognath patch pit plate point product pigment pack peeler print plaster powder poultice purpose peach paleness pigmentation pallidness pallor posture physiology pore pulp palpebra perspiration pus plica prepuce palm problem pain pathology potato pumpkin pear peeling pericarp pox pityriasis pemphigus psoriasis prurigo papule pustule pimple pock petechia paresthesia paraesthesia pockmark pachyderma progeria pyrogallol protein parchment pigskin phenol propanediol quick round rash rose recognition reaction removal rock roughness response ray ranid reptile roundworm rhino rhinoceros race ridge rubefacient redness receptor result rind ringworm rhagades rosacea roseola rhinophyma resin riboflavin southeast sore stone surface substance spot solution salve snorkeling squash strip slice sweat surgery skincare stylostixis sting suction stimulation stinger shark salamander snake sarcoptid scale shell spine seed wrinkle stomach tip treatment vegetable spore worker thickening vesicle thrust wound victim torture type whale ulcer wart tissue use tail tongue toenail vertebrate suede

sheepskin side trade throat tract woman top toe shoe texture sole sensation soreness shrublet tree sputum udder scleroprotein tattoo scar skin toiletry scalp weakness worm water stiffness scab shedding vessel sunburn term symptom sheep stress weal vitamin weight thigh white tumor turkey temperature warmness wool secretion taste swelling variety teeth tone sickness vine smallpox swine wine table ungulate snout viscera vascularity snorkel touch sense site silurid soapfish sunscreen sunblock sign stigma scratch scrape sallowness stuffer sapodilla sapota satsuma slough scurf skinner sweating sudation scabies sporotrichosis steatocystoma seborrhea scleredema sun suntan shammy tan tanning toadfish triggerfish workplace tannery toner tepidness tepidity taxidermy tactility stimulus topognosis topognosia tangerine tangelo tanner taxidermist vapor transpiration typhus tinea undergarment underwear underclothes underclothing urtication ugli ulceration urticaria vein venipuncture verruca visibility vesiculation vesication vitiligo vellum vernix webfoot waterskin whip wale wattle wineskin welt wanness whitehead wen wind windburn wheal xanthosis xanthoma xeroderma xerodermia zit

Word: instrument

ablative anterior alto amount action act attachment assignment agency adjustment air acceleration activity analysis accompaniment arrangement accelerator accelerometer accordion altazimuth airplane altimeter actinometer aircraft alidade adapter alidad area aperture analyzer analyser arm autofocus atmometer artillery audiometer automobile auriscope auroscope asdic accord atmosphere accuracy altitude angle agent arranger arc azimuth bottom breathing bone bronze bent bowing base baritone bass boot bank burning bow breath browse building bar banjo bagpipe ballistocardiograph band balalaika beam barometer bassoon bend bathymeter bathometer bell board blade binoculars bore body bones bolometer bronchoscope brass burette buret bugle bulb blood brain bean bellow bamboo cut color content current cross criminal clarion chemical cause comparison change completion circle circuit contraction combination contract chair cautery cryocautery charge capacity control class calculation composition contrivance calibration claim case copper chamber cardiograph castanets clappers caliper calliper cangue calliope cabinet console calorimeter camera cauterant ceiling celesta cello chronoscope clinometer cittern cithern cither citole clarinet choir clavier clavichord chime chordophone colorimeter coil column compass component cornet computer contrabassoon contrafagotto concertina cryoscope craniometer curette curet cymbal cloth cymograph curvature cavity cornea corpuscle chord cloud consort craftsman cocuswood cocoswood centrifugation centimeter century derivative drumhead differential drawing device deflection deflexion diffusion division dance dip display direction drum dose diamond declinometer dermatome dashboard densimeter densitometer dilator dilater disk dial dialyzer divider document dosemeter dosimeter dynamometer dulcimer drumstick density distance duration depth datum declination debt dancer entity event execution exposure electrocution elevation examination end echo evaporometer electrocardiograph extractor ergometer electroencephalograph edge electrodynamometer electromyograph esophagoscope endoscope eye engine eyepiece ensemble electroscope euphonium eudiometer earth embouchure energy eyeglasses effect esophagus etude earthquake evaporation fluid force fingering falsification foot flow frame figure function forgery field float fascia firearm fiber film fiberscope finger fingerboard flugelhorn fluegelhorn flute flagpole fret fiddle fluctuation floater flourish fanfare folk family freezing frequency group gash government gas gittern gong glass ground garrote garotte gallows gauge gage garrotte gourd graduate glockenspiel gravimeter grid guillotine guitar goad glasses graticule gravity graph granadilla gamelan glaucoma head hearing hammer hand horn hair hoop housing heliometer hematocrit haematocrit hemostat haemostat heart heat harp harmonica harmonium handle hodoscope hole horsehair hydrometer hygrometer hautboy hautbois horologe hardness height horizon humidity inside intent interior irregular iron interaction implementation instrumentation sentence manner place verb right interest property person voice set performance music piece particle rocket intensity radiation movable part theodolite telescope plane table line sight object musician style measure rate water sensitivity light measuring pulse term time range member torture leg strand wood skimming neck wind period recoil ventricle string pressure percussion use violin pair thumb rhythm tube orchestra section length mean mouthpiece tap tone tissue substance reason removal wart standard point plural punishment steam whistle keyboard radio television quantity input operator system steel plate measurement interval surveyor order inclination soundbox wire key piano sound lute reagent reed melting skull size scoop paper stylus record tree value security pen pointer movement position slice skin panel opening organ indicator solution membrane segment radioactivity power law voltage wave muscle surgery machine warfare ram lense tenor tuba volume reaction player vision solo note sequence placement optic strip woodwind rod red metal tune strangulation thickness rain variation pole state motion network xylophone kettledrum marimba spiral plasma separation star view vessel row physic trace path score skill relationship process role production intonation inclinometer integrator interference pattern interferometer meter user indication idea philosophy success instrumentalism profession instrumentalist kicksorter tortoiseshell kazoo irregularity surface keratoscope tension kettle zither koto tool weapon knife kymograph keyboardist shoulder strap instance vibrato pitch quality keytar limb lagerphone larynx laryngoscope ship speed log luthier mute mark singer reproduction pomp microscope procedure structure microsurgery membranophone magnetograph image magnifier pebble maraca scale resonator plectrum mandolin mallet microtome reading needle velocity nephoscope ocular incision operation orchestration oesophagoscope pipe oboe octant ocarina otoscope retina ophthalmoscope oxygen oximeter melody playing will requirement probate ivory plastic pick investigation injury probe wound planimeter source photometer paddle panpipe pianoforte number step pedometer regulator peg laser photocoagulation photocoagulator pelvimeter periscope post wrist pillory plectron piezometer plethysmograph picture plotter planet planetarium liquid pipet pipette respiration perspiration polygraph potentiometer prod screen projector protractor lyre psaltery precipitation pluviometer polyphony percussionist quadrant victim rack stop timbre register recorder

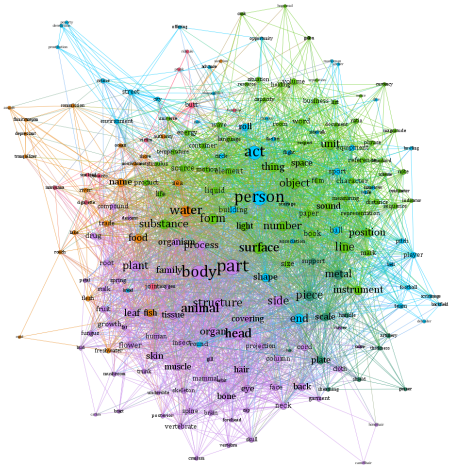
index refractometer rangefinder microwave radar radiolocation reticle reticule retractor rheometer mirror rhinoscope reset kind stopping implement switch slash stab standardization standardisation syrinx sonometer sousaphone saxhorn trombone sackbut samisen shamisen material sclerometer seismograph sextant sensitometer sector tip silverpoint sitar sights slipstick polyp tumor snare spherometer stake recording rapidity spirograph lung spirometer transmission reflection sonograph sonar spectacles specs vibration soundboard spectroscope speculum passage timber offender stocks stroboscope strobe stethoscope pilot propeller synchroscope synchronoscope synchronizer synchroniser variety synthesizer syringe synthesiser semitone strings thrust tripod transit tuning temperament trumpet trump tintometer tympan tympanum timpani tympani inch tape psychologist tachistoscope rotation tachometer tach tapeline information observer telemeter timepiece timekeeper theremin thumbscrew tensiometer moisture soil temperature thermograph thermometer trepan trephine triangle shape tonometer tracing tracer slide tucket run toccata tootle transcriber udometer service voluntary voicing valve violoncello variometer viscosity viscometer viscosimeter vibraphone vibraharp vibes viol viola lash whip wheel watt wattmeter zithern

E.3.2 - Depth Search task

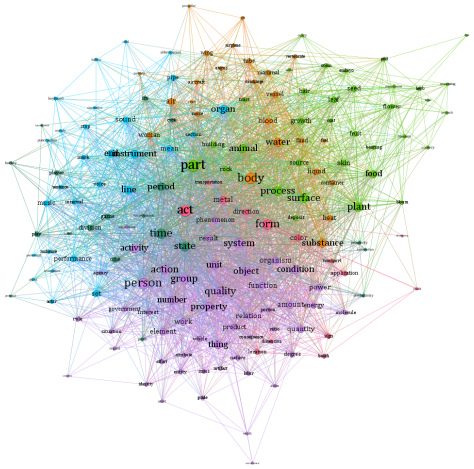
[**Below we present the four sub-networks utilised in *Depth Search*. For details on the implementations and subject behaviour please contact the authors.**]

Figure E.3.1: Overview of all sub-networks

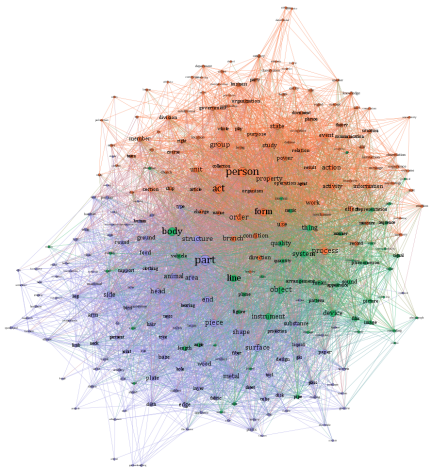
seed: rudd



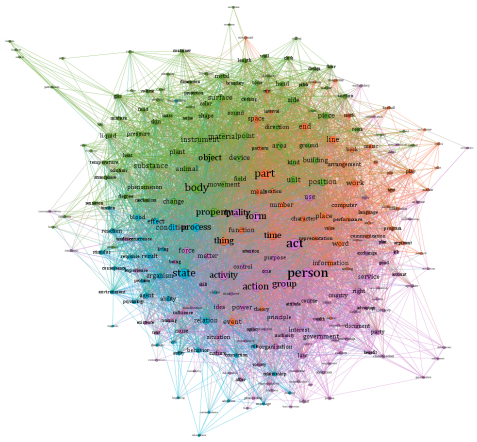
seed: dhal



seed: crinoid



seed: syringe



E.3.3 - Unusual uses task

[**Due to a lack of space, we cannot list all submitted Uses in this section. Please contact the author for the data on uses submitted.**]

E.3.4 - Tower of London

[**For brevity, please see Appendix D.3.4 for an illustration of the ToL network. Problems utilised correspond to these network positions (32 - 61), (25 - 54), (34 - 55), (22 - 31)**]