Shaping the Creative Performance of Teams: Experimental Evidence on the Effects of the Workplace, Incentives, and Gender Composition

Von der Fakultät für Wirtschaftswissenschaften der
Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades einer Doktorin der
Wirtschafts- und Sozialwissenschaften genehmigte Dissertation

vorgelegt von

Lisa Eva Klinkenberg

Berichter: Univ.-Prof. Dr. rer. pol. Christian Grund
Berichterin: Univ.-Prof. Dr. rer. pol. Christine Harbring

Tag der mündlichen Prüfung: 02.09.2026

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Parts of this dissertation have also been published as:

Grund, C., Harbring, C., & Klinkenberg, L. (2025). An experiment on creativity in virtual teams. *Journal of Economic Behavior & Organization*, *231*, 106926.

Grund, C., Harbring, C., & Klinkenberg, L. (2026). Creative Performance: A comparison of individuals and virtual teams – an Experimental Study. *Journal of Business Economics*.

Acknowledgments

Even though I knew early on that I wanted to pursue a doctorate, the journey toward completing this dissertation was not without challenges. Along the way, I was fortunate to be surrounded by people whose support, encouragement, and guidance made this work possible, and I would like to express my sincere gratitude to them.

First and foremost, I would like to thank my supervisors, Christian Grund and Christine Harbring. I deeply value the trusting and close working relationship we shared throughout the entire doctoral process. Thank you for your outstanding support in every phase of this dissertation, for your academic expertise, and for the countless insightful discussions that shaped this work. Beyond your professional guidance, I especially appreciate your ability to encourage me to slow down at the right moments, reconsider perspectives, and approach challenges with clarity and confidence. I have learned a great deal from both of you, not only academically but also personally.

I would also like to express my heartfelt thanks to our secretary, Anja Uttich, who is an indispensable part of our chairs. Her door was always open, she was willing to help whenever needed, and her kindness and positivity brightened even the more stressful days. Her support and warmth contributed greatly to the pleasant and supportive working environment I experienced throughout these years.

Furthermore, I would like to thank my colleagues at the Chair of HRM and Personnel Economics and the Chair of Organization — Alexandra Soboll, Anaïs Breulet, Anna Nießen, Can Ilgün, Lisa Pütz, Marcel Koch, Michael Hirsch, Philipp Monschau, and Stella Simons. I am grateful for the many ways in which you supported me: through discussions of experimental designs, arguments, and data analyses; through proofreading manuscripts; and through your help in conducting laboratory experiments. Equally important were the many enjoyable conversations, conference trips, shared experiences, and the personal support during the inevitable ups and downs that accompany a dissertation journey. Working alongside you made these years both intellectually enriching and personally memorable.

I would also like to thank all student assistants who supported my research over the past years. Your contributions through literature reviews, assistance in conducting laboratory experiments, evaluating creative responses, and formatting this dissertation significantly facilitated my work and are greatly appreciated. Finally, my deepest gratitude goes to my family and my partner. I would like to thank my parents, Monika and Ulrich, my twin sister Sarah, and my partner Malte

for their unconditional support throughout every stage of my life. You have been my emotional foundation during this journey. Thank you for listening whenever I needed to organize my thoughts aloud, for giving me confidence during difficult phases, and for celebrating every small and large success with me. Your love, patience, and unwavering belief in me carried me through this process more than you may realize.

To all of you, I extend my deepest thanks. This journey would not have been possible in the same way without each and every one of you.

TABLE OF CONTENTS

Table of Contents i
Index of Tables v
Index of Figures vii
Index of Appendices ix
Index of Abbreviations xiii
Introduction 1
Part I: Conceptual and Empirical Foundations 15
1 Creativity 15
1.1 Definitions and current research topics 15
1.2 Unusual Uses Task 18
1.3 Droodle-Task 21
1.4 Measuring creativity 22
2 Creativity and the workplace 24
2.1 Working from home 24
2.2 Creativity and virtual work 26
3 Pay-for-Creativity 27
3.1 Pay-for-performance incentives 28
3.2 Creativity and pay-for-performance incentives 30
Part II: Creative Performance: A Comparison of Individuals and Teams – an Experimental Study 37
1 Introduction 37
2 Definitions, current research and theoretical background 39
2.1 Work from home and virtual teams 39
2.2 Creative teams and the workplace 40
2.3 Hypotheses 43
3 Experimental design 44
4 Results 49
4.1 Descriptive results 49
4.2 Hypothesis testing 51
4.3 Potential mechanisms exploring the difference between individuals and teams 54
5 Discussion and conclusion 62
5.1 Limitations 63
5.2 Implications for further research 65
5.3 Practical implications 66
5.4 Contribution 67
Appendix I 69

Part III: An Experiment on Creativity in Virtual Teams 89
1 Introduction 89
2 Definitions and previous literature 91
2.1 Creativity in teams 91
2.2 Workplace setting and creative teams 93
2.3 Self-selecting into a workplace setting 94
3 Experimental design, treatments and measures 94
4 Hypotheses 98
5 Data collection and sample 100
6 Results 101
6.1 Hypotheses testing 101
6.2 Further results 108
7 Discussion and conclusion 112
7.1 Limitations 114
7.2 Implications for further research 115
7.3 Practical implications 117
7.4 Contribution 117
Appendix II 119
Part IV: Pay-for-Creativity: Fostering Creativity in Teams Through Monetary Incentives 139
1 Introduction 140
2 Theoretical background and literature review 141
3 Hypotheses 144
4 Experimental design 146
5 Data collection, sample and descriptive statistics 149
6 Results 151
7 Robustness checks 157
8 Discussion and conclusion 163
8.1 Limitations 166
8.2 Implications for further research 167
8.3 Practical implications 169
8.4 Contribution 169
Appendix III 172
Part V: Creativity in Teams – The Role of Gender and Incentives 203
1 Introduction 204
2 Theoretical background and literature review 205
3 Experimental design 209
4 Data collection, sample and descriptive statistics 210
5 Results 212

6 Robustness checks 218
7 Discussion and conclusion 226
7.1 Limitations 228
7.2 Implications for further research 229
7.3 Practical implications 231
7.4 Contribution 232
Appendix IV 234
Part VI: Concluding Remarks 245
1 Summary of findings and contribution 245
2 Future research 247
References 251

INDEX OF TABLES

Table 1: Overview of Essays 3
Table 2: Objective measures of creative performance 23
Table 3: Mean performance and SD for different measures of creativity of individuals and teams over both phases 50
Table 4: OLS regression on treatment effects for different performance measures 53
Table 5: Average levels of sympathy and trust in dyadic teams over the two phases with SD in brackets 55
Table 6: OLS regression on the effect of Sympathy and Trust on different performance measures 56
Table 7: Average performance of Teams and Simulated Teams with SD in brackets 59
Table 8: Characteristics of Teams and Simulated Teams 61
Table 9: Treatments 97
Table 10: Hypotheses 99
Table 11: Average quantitative and qualitative performance of the exogenous treatments in both phases with standard deviations in parentheses below 101
Table 12: Results of pairwise comparisons of the qualitative performance using Mann-Whitney U-Test to test for the first hypotheses 103
Table 13: Results of pairwise comparisons of the qualitative performance using Mann-Whitney U-Test to test for the second hypotheses 104
Table 14: Multivariate OLS regression analysis on qualitative creative performance of the second phase of the exogeneous treatments. 105
Table 15: Average quantitative and qualitative performance of the endogenous treatments in both phases with standard deviations in parentheses below 107
Table 16: OLS regression on qualitative creative performance and interaction effects between measures of ability and workplace 109
Table 17: OLS Regression analysis on qualitative creative performance of phases 1 and 2 including different measures of heterogeneity 111
Table 18: Structure and Treatments 147
Table 19: Average individual performance with standard deviation in parentheses, minimum, and maximum 151
Table 20: Average performance measures in the incentive phase with standard deviations in parentheses 152
Table 21: Results of pairwise comparisons of the different performance measures using Mann-Whitney U-Tests to test for the first hypothesis 153
Table 22: OLS regression on qualitative performance during the incentive phase 155
Table 23: Results of pairwise comparisons of the different performance measures using Mann-Whitney U-Test to test for the second hypothesis 156
Table 24: Results of pairwise comparisons of the different performance measures using Mann-Whitney U-Test to test for the third hypothesis 157

Table 25: OLS regression differentiating tournament and incentive effects on qualitative performance during the incentive phase 158

Table 26: Absolute performance differences between the first team task and the incentive phase averaged over tasks 160

Table 27: Descriptive statistics of teams with top-ranked ideas 162

Table 28: Experimental procedure 210

Table 29: Sample by gender composition and treatment 211

Table 30: Average creative performance of teams during the incentive phase by treatment and gender composition 212

Table 31: OLS regression on quantitative and qualitative performance in the team-check task and incentive phase 213

Table 32: OLS regression on quantitative and qualitative performance with interaction effects between incentive and gender composition differentiated between incentive type and intensity 215

Table 33: OLS regression on average differences of quantitative and qualitative performance moving from the team-check to the incentive phase with interaction effects between incentive type and gender composition 218

Table 34: Average number of top-evaluated ideas per treatment and gender composition 219

Table 35: OLS regression on top-evaluated ideas during the team-check task and incentive phase with interaction effects between incentive type and gender composition 220

Table 36: Average creative performance ratios per gender composition and treatment 222

Table 37: OLS regression on relative difference in qualitative performance moving from the team-check task to the incentive phase with a self-perceived gender measure 223

Table 38: Distribution of teams over gender composition, STEM involvement, and incentive type 224

Table 39: OLS regression on qualitative performance for incentive type and intensity split by STEM occupation of team members 225

INDEX OF FIGURES

Figure 1: Example Droodle 22
Figure 2: Experimental procedure 47
Figure 3: Average performance results of the two treatments with test results of Mann-Whitney U-Tests 52
Figure 4: Average errors made of Teams and Simulated Teams with test results of Mann-Whitney U-Tests 60
Figure 5: Experimental procedure 95
Figure 6: Average qualitative performance and SD in parentheses of the exogenous treatments in phase 1 and 2 and the absolute difference of performance between those two phases. 102
Figure 7: Average qualitative performance of different ability-paired teams over all tasks .. 110
Figure 8: Average qualitative performance per treatment over the course of the experiment. 153
Figure 9: Average qualitative performance of the team-task and the incentive phase averaged over tasks 160
Figure 10: Comparison of the quality per quantity ratio between the team-task and the incentive phase 161
Figure 11: Average qualitative performance differences between the team-check and the incentive phase per Droodle 217

INDEX OF APPENDICES

Appendix I

AI. 1: Experimental screen of the ability check task 69
AI. 2: Difference between quantitative and qualitative performance between treatments 69
AI. 3: Average qualitative performance of treatments per UUT with average ability over both phases 70
AI. 4: Average quantitative performance of individuals, low- and high-performing teams over both phases 70
AI. 5: Average qualitative performance of individuals, low- and high-performing teams over both phases 71
AI. 6: Comparison of gender composition of teams and simulated teams 71
AI. 7: Comparison of frequency of levels of age difference in years between teams and simulated teams 72
AI. 8: Spearman's rank correlation 73
AI. 9: Pearson correlations of different performance measures, sympathy, and trust 74
AI. 10: OLS regression on interaction effect of extraversion on performance ratio of teams and simulated teams 75
AI. 11: OLS regression on interaction effect of openness on performance ratio of teams and simulated teams 75
AI. 12: Instructions 76
AI. 13: Questionnaires 84
AI. 14: Categories 87

Appendix II

AII. 1: Experimental setting in persence 119
AII. 2: Experimental screen of a task 119
AII. 3: Gender composition of dyadic teams for all treatments 120
AII. 4: Age of the participants divided for all treatments 120
AII. 5: Educational level of the participants divided for all treatments 120
AII. 6: Profession of the participants divided for all treatments 121
AII. 7: Origin of the participants divided for all treatments 121
AII. 8: Average creative ability and standard deviation in parentheses measured on the individual level 121
AII. 9: Multivariate OLS regression analysis on qualitative creative performance of the first phase of the exogeneous treatments 122
AII. 10: Average Sympathy and Trust in both phases for all treatments 122
AII. 11: The potential role of selection effects due to Dropouts 123
AII. 12: Heckman 2SLS selection model 124

AII. 13: Multivariate OLS regression analysis of several variables and interaction effects on qualitative creative performance of the second phase of the exogenous treatments 125
AII. 14: Different measures of creativity and treatments effects of the exogeneous treatments 126
AII. 15: Instructions 127
AII. 16: Questionnaires 136

Appendix III

AIII. 1: Droodle used as ability-check 172
AIII. 2: Droodle used in the team-check 172
AIII. 3: First Droodle of the incentive phase 173
AIII. 4: Second Droodle of the incentive phase 173
AIII. 5: Third Droodle of the incentive phase 173
AIII. 6: Example screen 174
AIII. 7: Average quantitative performance per treatment over the course of the experiment 174
AIII. 8: Average originality of answers per treatment over the course of the experiment 175
AIII. 9: Average quantitative performance in the team-check task and the incentive phase averaged over tasks 175
AIII. 10: Sample characteristics 176
AIII. 11: Means of different experiment-related controls measured on the individual level . 177
AIII. 12: Pearson correlations of individual characteristics, perceptions, and total performance measures 178
AIII. 13: OLS regression on quantitative performance during the incentive phase 179
AIII. 14: OLS regression on qualitative performance during the incentive phase with stepwise inclusion of the control variables 180
AIII. 15: OLS regression on qualitative performance including interaction effects between several control variables and treatment effects 181
AIII. 16: Indirect effects of potential mediators of PfP systems on qualitative creative performance with *Fix* as baseline 183
AIII. 17: OLS regression on qualitative performance including interaction effects between several control variables and tournament effects 184
AIII. 18: OLS regression on quantitative and qualitative performance including interaction effects between tournament incentives and tournament related controls 185
AIII. 19: Mann-Whitney U-Tests comparing different performance ratios between treatments 186
AIII. 20: OLS regression on average top-ranked ideas per Droodle during the incentive phase 186
AIII. 21: OLS regression on quantitative and qualitative performance including interaction effects between ability of teams and treatment effects 187
AIII. 22: OLS regression on qualitative performance during the incentive phase split by previous ability of teams 188

AIII. 23: OLS regression on quantitative and qualitative performance controlling for heterogeneity in teams 189

AIII. 24: OLS regression on qualitative performance controlling for socio-demographic differences between team members 190

AIII. 25: Instructions 191

AIII. 26: Questionnaires 197

Appendix IV

AIV. 1: Socio-demographic statistics by gender 234

AIV. 2: Means of different experiment-related controls by gender measured on the individual level 235

AIV. 3: OLS regression on creative performance during the team-check task and incentive phase of the *Fix* treatment 236

AIV. 4: OLS regression on quantitative and qualitative creative performance with interaction effects between treatments and gender composition 237

AIV. 5: OLS regression on qualitative creative performance with interaction effects between incentive type and several control variables 238

AIV. 6: OLS regression on qualitative creative performance with interaction effects between incentive intensity and several control variables 239

AIV. 7: OLS regression on average differences of quantitative and qualitative performance moving from the team-check to the incentive phase with interaction effects between incentive intensity and gender composition 240

AIV. 8: OLS regression on average differences of quantitative and qualitative performance moving from the team-check to the incentive phase with interaction effects between treatments and gender composition 241

AIV. 9: OLS regression on top-evaluated ideas during the team-check task and incentive phase with interaction effects between incentive intensity and gender composition 242

AIV. 10: OLS regression on qualitative performance during the incentive phase with a self-perceived gender measure 243

AIV. 11: OLS regression on qualitative performance during the incentive phase split by previous ability of team members with interaction effects between gender composition and incentive schemes 244

INDEX OF ABBREVIATIONS

AI	Artificial intelligence
CMC	computer-mediated communication
e.g.	exempli gratia, for example
Endo	endogenous treatment
et al.	et alii (and others)
Exo	exogenous treatment
FtF	Face-to-Face
HO	Home Office
i.e.	id est (that is)
IM	intrinsic motivation
LLM	large language models
LSA	latent semantic analysis
N	number of observations
PfP	pay-for-performance
Plow	low- intensity piece-rate
Pmax	high-intensity piece-rate
PR	presence
RCT	Role Congruency Theory
R&D	research and development
SD	standard deviation
SDT	Self-determination Theory
SRT	Social Role Theory
Tlow	low-intensity tournament
Tmax	high-intensity tournament
TSHC	team-specific human capital
TTCT	Torrance Test of Creative Thinking
UUT	Unusual Uses Task
WFA	work from anywhere
WFH	work from home

INTRODUCTION

> *'[The] process of Creative Destruction is the essential fact about capitalism.'*
>
> - Joseph A. Schumpeter, *Capitalism, Socialism and Democracy* (1942)

Schumpeter stated that creative destruction is the essence of capitalism through which new innovations can lead to radical change, thereby destroying or at least rethinking existing systems and patterns, and with that drive forward industries and organizations (Schumpeter, 2008). The relevance of innovation through creating something new is further underlined by the Nobel Memorial Prizes in economics 2025 of Joel Mokyr, Philippe Aghion, and Peter Howitt for the theory of sustained growth through creative destruction. Such innovations leading to change and growth are induced by creative ideas and actions, also referred to as non-routine work (e.g., West & Sacramento, 2023). An increased demand for such non-routine work in the organizational world as a starting point for innovation, not always being radically destructive, underlines the relevance of non-routine and often creative work tasks (Autor et al., 2003; Englmaier et al., 2024b).

According to the World Economic Forum's Future of Jobs Report 2025, creativity is the fourth most important skill employers consider as a core competency for employees and is expected to increase in relevance for the working life until 2030 (World Economic Forum, 2025). As creativity is on the rise, the relevance of understanding creative processes from an economic and scientific perspective is too. A prominent research gap concerns participation in non-routine work, and which work-related factors can influence creative performance. Previous studies show that non-routine work and creative tasks are increasingly being handled by teams (Kozlowski & Bell, 2013) but existing research lacks an in-depth understanding of such teams, being characterized by higher needs for coordination, cooperation, and risks of conflict than individual work. This is done with special focus on the influence of organizational work characteristics, i.e., the workplace setting, incentive structures, and team composition, on achieved level of creative performance by teams (e.g., Englmaier et al., 2024a). Therefore, this dissertation examines influential factors of the workplace (physically and organizationally) on creative performance of teams using laboratory experiments. To be more concise, in the following essays, creativity is defined as the production of novel and useful ideas (Amabile, 1988), and the focus is on the aspect of divergent thinking, i.e., finding as many creative solutions as possible, reflecting the early stage of innovative processes (Reiter-Palmon et al., 2019). Divergent thinking sets the starting point for innovation, which is the result of harnessing creativity, coupled

with the successful implementation of creative ideas (Ford, 1996). By focusing on such an early stage, this dissertation is of informational value for a wide range of creative issues in most parts of work and industries and thus, for most employers and employees in contact with non-routine work – as opposed to routine, process-driven tasks such as assembly line or conveyer belt operations - whether that non-routine work is radical or incremental in nature. This includes, for example, consultants navigating complex client problems and searching for different strategic options, teams in research and development (R&D) responding to shifting market trends and emerging consumer demands, or knowledge workers across sectors who must regularly adapt, innovate, and generate several novel solutions rather than follow a fixed procedure.

In the following, an overview of the four essays of this dissertation is presented, followed by a summary of each essay separately. Thereafter, an overview of the method used in this dissertation is given, namely experimental economics.

Overview of essays

This dissertation contains four essays evolving around creative teams and determinants of their creative performance with regard to the organization of work. All essays use the methods of experimental economics, in the laboratory as well as online. Table 1 gives an overview of the four essays, containing research questions, design, and creative tasks used, as well as the key findings. Next up, a short overview of each essay and their interrelationships is presented.

A parallel development to the increasing demand for creative work is the rapid increase in technologies enabling virtual communication and other forms of virtual work (Brenan, 2020). This, coupled with the demand for a global workforce (Blay et al., 2026; Kharroubi, 2021) and accelerated by COVID-19 pandemic restrictions, results in an increasing demand for virtual work, especially in teams (Barrero et al., 2021). Therefore, the first two essays of the dissertation focus on the effects of the (physical) workplace on the creative performance of teams. First, the question is addressed whether teams are at all superior to individuals working on a creative task, considering a virtual working environment. Virtual exchange during teamwork might influence team performance differently than working in presence (e.g., Girotra et al., 2010; Janis, 1972; Mullen et al., 1991; Osborn, 1953), thereby, setting the need for this specific research question, enlarging the general debate whether teams or individuals are to be preferred for a creative task. Results show a higher creative performance of teams, especially for qualitative measures.

Building on the positive results in favor of teams, the second essay digs deeper into the role of the place to work at for teams, considering virtual, presence, and hybrid working forms and the

effect of being able to self-select into a workplace setting. Thereby, this essay contributes to closing a gap in previous existing literature (e.g., Dutcher & Rodet, 2022; Englmaier et al., 2024a; Grözinger et al., 2020). The study results are in favor of teamwork being conducted, at least partially, in presence and indicate that not the consideration of performance but rather individual preferences are relevant when self-selecting into a workplace setting. The first two essays both use the Unusual Uses Task (UUT) to assess divergent thinking, thereby contributing to the general understanding of this task and its implementation in the context of direct teamwork in the laboratory, as it was previously mainly used as an individual task.

Table 1: Overview of Essays

Essay	Research Question	Experimental Design	Task	Main Results
1	Do teams show higher creative performance than individuals when working from home?	2-phase online experiment one week apart in which several UUTs are either solved individually or in dyadic teams.	UUT	Teams exhibit significantly higher creative performance than individuals across the two virtual phases.
2	Are teams more creative in presence than they are in an online setting? What is the impact of the initial workplace setting for subsequent creative performance of teams? Does the ability to self-select into a workplace setting affect the creative performance of a team?	2-phase experiment one week apart with dyadic teams in the lab and online, varying between fully online, fully presence, and hybrid settings. Additionally, 2 self-selection treatments for solely working in presence or online are implemented.	UUT	Teams working at least one phase in presence show higher creative performance than those working entirely online. No significant effects of self-selection on creative performance are found.
3	Is the level of creative performance of teams higher under pay-for-performance (PfP) than pure fixed pay? Does a pure PfP system lead to higher levels of creative performance than a system including also a fixed pay component?	1-phase experiment in the lab, with five treatments differing in the implemented incentive schemes, piece-rate or tournament, by altering the share of payment which is determined by performance.	Droodle	Teams under PfP systems outperform teams under fixed payment, whereas creative performance increases with incentive intensity. Tournament systems lead to higher creative performance than comparable piece-rate systems.
4	Do PfP incentives affect the creative performance of teams differently than fixed pay, depending on the gender composition of the team? Does this effect depend on the type and intensity of the incentive?	Same design as in Essay 3.	Droodle	No evidence for an interaction effect between gender composition and PfP on creative performance is found. Female teams show higher performance in an unincentivized task, whereas teams with at least one male catch-up when incentives are introduced.

Following from the general increased need for understanding creative teamwork and the influence of workplace determinants, the additional question arises how creative performance can be enhanced and further incentivized. This shifts attention to the organizational design of one particularly important work characteristic: payment. Existing research results are mixed as to whether there is any effect of monetary incentives on creativity and if so in what direction (e.g., Charness & Grieco, 2019; Erat & Gneezy, 2016; Gross, 2020; Kachelmeier et al., 2024). Essay 3 covers this issue by investigating the role of piece-rate and tournament incentive schemes and their intensity, i.e., the specific wage fraction which is performance dependent. Results show a positive effect of performance-dependent incentives in comparison to fixed payment on creative team performance and the dominance of inter-team tournaments in comparison to just increasing incentive intensity.

Remarkably, women are underrepresented in the field of non-routine work (Iaria et al., 2022), even though no significant difference between the creative potential of males and females exists (Baer & Kaufmann, 2008; Runco et al., 2010). The competitive environment of creative work might explain this difference. A broad literature hints towards differing effects of incentives and especially tournaments on performance of males and females (e.g., Datta Gupta et al., 2013; Niederle & Vesterlund, 2011). Nonetheless, the interaction of gender composition of teams and implemented incentive structures is not clear (e.g., Apesteguia et al., 2012; Hoogendoorn et al., 2013; Ivanova-Stenzel & Kübler, 2011) and those interactions in the creative field have not yet been investigated in a clean laboratory setting focusing on several incentive types and intensities. Therefore, Essay 4 extends the research of the third essay towards the potential moderation of gender composition of a team on the effects of incentive schemes on creative performance. Results show no evidence for moderation by gender composition. Pure female teams show significantly higher levels of creative performance in an unincentivized task, whereas teams with at least one male tend to catch-up as soon as incentives are in place, independent of incentive scheme, which is in line with earlier research findings on contingent conscientiousness (Segal, 2012).

In the following, the four essays are summarized in more detail, giving their motivation, a short conclusion, outlook on future research, and their practical relevance. Thereafter, the method of experimental economics is introduced. The first part of this dissertation introduces the theoretical and empirical background relevant for this thesis, which is then followed by the four essays. Part VI presents a summary of the key findings as well as avenues for future research, i.e., convergent thinking, artificial intelligence (AI) in context of creative thinking, and a potential setting for a field experiment, are discussed.

Essay 1: Creative Performance: A comparison of individuals and teams – an Experimental Study

The first essay addresses the issue of whether individuals or teams should perform a divergent thinking task in a virtual workplace context. This specific research question is motivated by the pressing relevance of virtual work and non-routine tasks in organizations. Existing research in the field of creative teams focuses on the effect of diversity of team members on the creative output (e.g., Dutcher & Rodet, 2022; Hoever et al., 2012; Kurtzberg, 2005), or the optimal communication form and workplace setting for such teams (e.g., Brucks & Levav, 2022; Grözinger et al., 2020). However, there is little research addressing the underlying, more fundamental question of whether it is at all beneficial to assign creative tasks to a team rather than one or more individuals in a virtual work setting over an extended period of work. Building on media-richness theory (Daft & Lengel, 1986), computer-mediated communication (CMC) as part of virtual teamwork can lead to different team interaction and outcomes compared to communication in presence, due to for example reduced communication quality and coordination difficulties (e.g., Girotra et al., 2010; Janis, 1972; Mullen et al., 1991; Osborn, 1953). Additionally, the research moves beyond the one-shot observation of teams to account for the team formation phase, which can be especially affected by CMC (Reiter-Palmon et al., 2021).

The first essay does not only contribute to the previously named research gap. In addition, the well-known UUT, which is typically used to assess individual creativity (e.g., Guilford, 1967; Plucker et al., 2011; Torrance, 1966, 1974), is transferred to a virtual team-setting. Therefore, the first essay contributes to the application of this task in an experimental economic context focusing on team performance.

To answer the underlying research question, an online experiment has been conducted over two work phases, one week apart, using the video communication tool *Zoom*. One important characteristic of the design is that participants have to confirm beforehand to only participate from their homes, thus, reflecting for a real work-from home (WFH) situation, unlike other studies (e.g., Brucks & Levav, 2022; Dutcher & Rodet, 2021; Grözinger et al., 2020). Participants are either randomly sorted into a dyadic team, i.e., a team of two, or worked individually on four UUTs in each working period. In each UUT, participants have to generate as many alternative uses as possible for a given object, e.g., a tin can. The experiment contains two phases, as the intention was to account for the team formation process, which is determined by sympathy and trust, whereas especially the latter develops over time (Meyerson et al., 1996). Next to the main

dependent variable, i.e., quantitative and qualitative creative performance, several control variables concerning individual characteristics as well as team-specific items, like sympathy or trust, are assessed using several questionnaires after each working phase.

The results show significantly higher creative performance of teams in comparison to individuals across the two phases. Furthermore, this effect also holds when the performance of dyadic teams is compared to the aggregated performance of two individuals, whereas this effect is driven by the generated quality of ideas in relation to the sheer quantity of creative ideas. Therefore, when considering teams or individuals for a creative task in a virtual context, the specific creativity goal, i.e., quantity or quality, needs to be considered. Additionally, trust is found to be a significant positive factor on creative performance of teams, whereas sympathy showed a significant negative influence, which can be attributed to a loss of task-focus when team members enact rather unsupervised in a virtual environment (Methot et al., 2016; Pillemer & Rothbard, 2018).

The results of this study underscore the relevance of teamwork for a non-routine task, even in a virtual environment. Furthermore, the positive finding for creative teams is a strong foundation for further research on (virtual) workplace settings and their influence on the creative performance of teams. From a practical perspective, the study hints towards the implementation of virtual dyadic teams working on divergent thinking tasks, if quality of ideas is the performance dimension of interest.

Essay 2: An Experiment on Creativity in Virtual Teams

Enabled by increased digitalization, motivated by a shortage of skilled workers, and enforced by the COVID-19 pandemic, the demand for virtual work increased over recent years (Barrero, 2021; Brenan, 2020; Kharroubi, 2021). Thereby not only completely virtual or presence work-models evolved, but also hybrid models allowing to switch between working at the own home and the facilities of the organization (Iqbal et al., 2021; Naqshbandi et al., 2023). Hybrid models are currently under-researched, due to their relative newness, as well as the need to investigate several work phases during which the change of workplace occurred. Previous studies on teams in the virtual context focus mostly on existing teams, but hint at problems related to trust and coordination during the formation process (Abi Saad & Agogué, 2023; De Guinea et al., 2012; Vuchkovski et al., 2023) which is underlined by social presence theory (Short et al., 1976) and media-richness theory (Daft & Lengel, 1986). Noteworthy, those existing studies have been mainly conducted before the pandemic as well as the fast increase in digital work, both leading to higher levels of familiarization with, for example, CMC or work coordination in a virtual

context. Due to the recent changes in the physical workplace and thereby communication, which is essential for team-specific human capital (TSHC) (Blair, 1999) and direct teamwork in the creative context, the second essay investigates the influence of purely online, presence, and hybrid workplace settings on the creative performance of teams. Additionally, self-selection into a workplace setting is investigated as individuals differ in their preferences for WFH, which can influence performance, perceived autonomy, and the decision to work at an organization (Deci & Ryan, 1987; Kaduk et al., 2019).

To investigate the effect of different physical workplaces on the creative performance of teams, a laboratory experiment with two phases, one week apart, has been designed. Per phase, participants have to work in the same dyadic teams on four different UUTs. Dependent on the treatment, both phases are either conducted fully online via *Zoom* (control), fully in presence, or the workplace is switched between the two phases, reflecting hybrid work. Additionally, two self-selection treatments are implemented, allowing participants to either decide on participating only in presence or online. Both phases are followed by additional questionnaires, covering individual- and team-related controls like personality, sympathy, and trust. For the control treatment, the data of dyadic teams of Essay 1 is used.

Results of the experimental study clearly point towards beneficial effects of working at least one phase in presence during the formation process of a creative team, as creative performance is significantly higher in the fully presence or hybrid sessions than when fully working online. Thus, beneficial effects of working the first phase in presence spillover to the following phase, independent of workplace. Those findings might be explained by the knowledge of participants that they will work together in presence at least once, giving room for stronger reciprocal behavior, due to FtF interaction. Moreover, no significant effect of self-selection on creative performance is found, indicating that rather individual preferences that are not performance related are relevant for workplace decisions.

Building on those findings, future research could investigate the underlying mechanisms, for example, by collecting verbal and non-verbal communication data during the experiment. Additionally, a time horizon beyond the early stages of team formation could shed light on potential tradeoffs of switching workplaces or fully working online on a daily basis. Overall, Essay 2 provides highly relevant information for practitioners, suggesting that newly formed teams working on divergent thinking tasks should meet up at least once during their interaction to develop their creative potential.

Essay 3: Pay-for-Creativity: Fostering Creativity in Teams through Monetary Incentives
Creative professions and teams often have to prove their creativity in competitions around the most creative ideas (Morgan & Wang, 2010; Terwiesch & Ulrich, 2009). Examples for such competitions can be found at online platforms for logo designs or public architecture tenders. Yet despite the practical relevance of such competitions, research on incentivizing creative teams, using absolute incentives like piece-rates or relative incentives like between-team tournaments, remains scarce and yielded mixed findings. Essay 3 of this dissertation addresses this gap by examining how performance-dependent incentives, and in particular their intensity, i.e., the wage fraction depending on performance, affect the creative performance of teams.

While PfP is a well-established mechanism in routine tasks and generally shows positive effects (Bonner & Sprinkle, 2002; Lazear, 2000; Pendergast, 1999), its transferability to creative contexts is uncertain, following Self-determination Theory (SDT) (Deci & Ryan, 1985; Ryan & Deci, 2000). Creative output is associated with high uncertainty in both production and evaluation, which challenges findings from routine work (Bradler et al., 2019). Moreover, prior research has largely focused on individuals (e.g., Malik et al., 2015), even though teams often outperform individuals in creative tasks, as shown in both theory and empirical evidence (Feinstein, 2011; Kozlowski & Bell, 2013), including Essay 1 of this dissertation. As creative teams might react differently towards PfP than individuals, due to channels of communication, cooperation, and inter-team considerations (Baer et al., 2014; Gächter et al., 2025), it is essential to understand incentive effects on creative team performance.

Finally, this essay does not only consider an arbitrarily chosen piece-rate or tournament but systematically varies the intensity of incentives. Following Principal-Agent Theory, the balance between motivating creativity through sufficiently strong incentives and the risk aversion linked to variable pay renders the net effect of PfP on team creativity theoretically ambiguous (Gerhart & Fang, 2015; Jensen & Meckling, 1976; Kim, 2025).

Existing research focuses merely on individuals performing creative tasks, showing different results. Incentive systems can have a positive (Charness & Grieco, 2019; Kachelmeier et al., 2024), a negative (Attanasi et al., 2026; Huo, 2020), or no effect at all (Eckartz et al., 2013; Malik et al., 2015) on creative performance. Thereby, the existing results are not that easy to compare and aggregate towards a final statement on the effectiveness of incentives, due to variance in design, incentive scheme, and especially incentive intensity.

To address this gap in current research, I have conducted a one-phase laboratory experiment. During the experiment, participants have to work in dyadic teams on several versions of the so-

called Droodle-Task, which measures again divergent thinking as output of direct teamwork. In total, five treatments are implemented, covering fix, piece-rate, and tournament payment with varying incentive intensities for the latter two. The working phase is followed by some questionnaires, covering individual and team-related controls as well as self-assessment of performance.

Results of this study reveal that creative performance under PfP systems, except for a low-intensity piece-rate system, is significantly higher than under a fixed payment scheme. Additionally, performance increases with incentive intensity, whereas a relative inter-team incentives outperform absolute schemes. The incentive intensity is less decisive for creative performance than the presence of a tournament. Winner-takes-all tournament scenarios lead to highest creative performance of teams, however, with the highest variability in results. Therefore, Essay 3 provides valuable information for practitioners, aiming to increase the creative performance of their teams, potentially increasing performance of an organization and contributing to employees' work life. More concise, if quality but also stability of performance is of interest, the implementation of between-team competitions with a loser prize is suggested.

Future research could dig deeper into team characteristics potentially moderating those effects, as for example done in Essay 4 considering the gender composition of teams. Moreover, not only are external monetary incentives a promising tool to enhance performance, but also non-monetary incentives and feedback should be considered in future research.

Essay 4: Creativity in Teams – The Role of Gender and Incentives

Although research does not hint at gender differences in the ability to perform creative tasks (e.g., Baer & Kaufmann, 2008), women are underrepresented in corresponding fields, e.g., of all listed inventors worldwide, only 13 percent are women (WIPO, 2023). Therefore, the question arises if work-related environmental factors, like incentive structures, might influence the creative performance of women. This is again relevant for creative teams, as often teams work together on a specific R&D project or in the academic field. The fourth essay of this dissertation aims to shed light on the potential interactions between PfP incentive schemes and gender composition of teams on their creative performance and is an extension of Essay 3.

Research on gender differences in the context of PfP incentives often focuses on the preferences to work under a certain incentive scheme (e.g., Datta Gupta et al., 2013; Dohmen & Falk, 2011; Gneezy et al., 2003; Niederle & Vesterlund, 2007). The focus of this essay is on performance under and not preference for specific incentive schemes. Thereby investigating incentives for existing employees which are confronted with a specific work task, i.e., divergent thinking as

a team. Such tasks are often a part of working life without being the only or main task to perform, thus, without active selection in only this field of work under a specific incentive.

Only a few studies exist that investigate the role of gender composition of teams in a creative context (e.g., Dutcher & Rodet, 2022; Rodet, 2025; Verdes & Vásárhelyi, 2023), and to the best of my knowledge, no study focuses on the role of incentive schemes. Results on the role of gender composition on team performance are ambiguous (e.g., Apesteguia et al., 2012; Hoogendoorn et al., 2013; Ivanova-Stenzel & Kübler, 2011; Städter et al., 2022), leaving it open whether and how single-gender or gender-mixed teams outperform the others. In contrast, relative clear results exists that men tend to react more positively towards tournament incentives than women (e.g., Corti et al., 2001; Niederle & Vesterlund, 2011), especially in male-stereotyped tasks (Niederle & Vesterlund, 2011; Shurchkov, 2011), which creativity is, as it associated with traits of independence, self-direction, adventurous and risk taking (Proudfoot et al., 2015). Social Role Theory (SRT) and Role Congruity Theory (RCT) together suggest that because creativity is coded as agentic, and thus male-typed, women face systematic disadvantages in how their creative output is recognized and rewarded, which can be further enhanced by competitive environments (Eagly, 1997; Eagly & Karau, 2002; Hora et al., 2021).

Essay 4 uses the identical design and sample as Essay 3, focusing on gender composition. Results do not show harsh evidence for an interaction between PfP incentive schemes (absolute and relative) nor PfP intensity and gender composition of dyadic teams. Results show that pure female teams show significantly higher performance than teams with at least one male in an unincentivized task, which later on diminishes when incentivizes are into place. This observation is not due to a specific introduced incentive, nor a drop in performance of pure female teams, but a significant increase in qualitative performance of teams with at least one male independent of incentive scheme. These findings are in line with other research, indicating higher performance of females than males working on unincentivized tasks (Lewis et al., 2023; Segal, 2012).

Considering communication channels in such a team and team size are great starting points for future research. From a practitioners' perspective, these findings indicate that a competitive environment when engaging in creative tasks does not seem to be harmful for pure-female or mixed teams while increasing average creative performance. Thus, relative performance incentives can be implemented to increase creative performance of teams working on divergent thinking tasks, without risking negative effects for females being involved.

The following chapter includes an overview of the method of experimental economics, which is used for all four essays of this dissertation. Thereby, the relevance of experimental economics, classifications, characteristics, procedures, as well as limitations are presented.

Method: Experimental Economics

In the four essays of this dissertation economic experiments are used to gather data and answer the research questions. Therefore, the following section introduces different types of experiments and discusses the necessity of conducting a scientific experiment, with a particular focus on real-effort laboratory experiments.

Economic experiments gained in popularity in the last century. Reinhard Selten introduced and established experimental economics in Germany in the 1980s. He used this method of research to develop theories that could better describe and predict the limited rational behavior of subjects than traditional theory could (e.g., Selten, 1989; Selten & Stoecker, 1986). In recent years, two Nobel Memorial Prizes in economics have been assigned to economists for their experimental research, i.e., Daniel Kahnemann and Vernon Smith in 2002, and Abhijit Banerjee, Esther Duflo, and Michael Kremer in 2019.

Economic experiments can broadly be distinguished into laboratory and field experiments. Whereas laboratory experiments are conducted in an artificial and fully controlled environment with actively recruited participants, field experiments are conducted in naturally occurring environments, allowing researchers to study behavior in real-world settings with higher external validity and greater generalizability of findings to broader populations (Harrison & List, 2004). However, this comes at the cost of control: In natural field experiments, it is rarely possible to fully isolate the effect of a single variable, as participants are exposed to a range of confounding influences such as managerial pressure, peer effects, or organizational culture that cannot be held constant across conditions (Harrison & List, 2004).

With respect to conventional laboratory experiments in contrast to field experiments, Falk and Heckman (2009) identify several advantages and disadvantages. The authors emphasize that laboratory experiments provide the highest level of control over external factors and variance, thereby enabling identification of causal relationships, keeping all conditions constant except the independent variable. Therefore, laboratory experiments are suited for explorative research. They allow, for example, a clean measurement of performance-influencing factors in isolation due to random distribution of participants over treatments, instead of assessing performance while not fully controlling the work environment and therefore communication or external in-

fluences like leadership styles or pressure. Furthermore, laboratory experiments are cost-efficient, highly replicable, and incentivized participation ensures that subjects engage with sufficient seriousness (Friedmann & Cassar, 2004).

To conduct a laboratory experiment, the experimenter must first develop a comprehensive experimental design that enables a systematic investigation of the underlying research questions. The design specifies all relevant procedural elements, including experimental tasks and their sequence, structure and content of questionnaires, the physical or virtual setting, and the expected duration of each session. Typically, economic experiments rely on established tasks or games, and several experimental treatments and a control group are prepared that differ solely in the independent variable of interest, thereby ensuring internal validity and comparability across conditions (Guala, 2005).

Participants are typically recruited from an established subject pool, which commonly consists of university students (Harrison & List, 2004). They are randomly allocated to treatments in order to avoid self-selection effects (Guala, 2005). In conventional economic laboratory settings, participants make decisions individually in separated cubicles equipped with computers, thereby minimizing communication and ensuring anonymity of decision-making. Notably, the experimental studies conducted for this dissertation deviate from this traditional setup. As interpersonal communication constitutes a crucial element of the following experimental designs, participants are assigned to small teams in which they are required to interact and jointly make decisions, thus, eliminating the need for individual separated workstations within a team. Depending on the specific experimental condition, sessions are conducted either in the physical laboratory environment or from home via an online video communication platform, facilitating structured group interaction while maintaining experimental control.

At the beginning of each session, participants receive detailed written and verbal instructions outlining the conditions under which decisions will be made. These instructions explain the general structure of the experiment, the rules governing the tasks, the flow of the session, and the determination of payoffs. Participants are informed whether they will make decisions individually or collectively within a group, which information they will have access to, and how their own and others' decisions affect their outcomes. Importantly, participants are not lied to in economic experiments.

Consistent with standard practices in experimental economics, monetary incentives are employed to ensure that decisions have real economic consequences (Croson, 2005). However, while participants in traditional laboratory experiments typically receive their payments in cash

immediately following the session, in the present essays, compensation is transferred via bank payment after the completion of the experiment. This procedural modification reflects two design requirements specific to the present essays: the assessment of creative performance, which necessitates a delayed payment calculation, and the partial use of online experimental settings, which precludes direct cash payment.

Within the domain of laboratory experiments, a distinction is drawn between chosen-effort and real-effort designs. In chosen-effort experiments, participants select an abstract effort level without actually performing the corresponding task. They are often provided with functions or tables that specify the costs associated with different effort levels. By contrast, real-effort experiments, as used in this dissertation, require participants to exert actual mental or physical effort during the tasks. Various types of real-effort tasks have been developed. Carpenter and Huet-Vaughn (2019) distinguish eight categories, including computer-based tasks, counting activities, puzzle solving, office tasks, and typing exercises. They advocate for office-related tasks, as these are familiar to student participants and transferable to actual work settings. Charness et al. (2018) provide an extended classification of eleven categories, ranging from data entry and IQ tests to physical activities, text decoding, and mail sorting. Tasks like mail sorting would not reflect creative performance. Thus, in the essays of this thesis, two different tasks are used to assess divergent thinking as part of creativity via idea generation.[1]

The advantage of chosen-effort experiments lies in their ability to precisely control the cost function, while real-effort designs cannot fully capture individual effort costs. Moreover, perceived costs in chosen-effort experiments are exclusively monetary. Real-effort experiments are considered more realistic and thus potentially yield higher external validity, as participants engage in tasks that resemble real-world activities. Nevertheless, such experiments also entail reduced control due to individual heterogeneity in effort costs. Dutcher et al. (2024) emphasize the importance of the cost function, arguing that when the function is complex or difficult to model accurately, for example, when teams and cost of communication or coordination are to be considered, real-effort tasks are preferable. Brüggen and Strobel (2007) likewise found no significant differences in effort levels between the two designs but noted greater variance in outcomes under real-effort conditions due to lower experimental control. Comparative studies by Charness et al. (2018) suggest that differences between chosen- and real-effort experiments are often negligible. They recommend real-effort designs, if task performance is not strongly dependent on individual abilities and behavior itself is the central variable of interest. Given

[1] The tasks used are discussed in more detail in Part I, section 1.2 and 1.3.

that the present thesis focuses on behavioral aspects such as the interaction between team members, contexts in which cost functions go beyond purely monetary components, a real-effort laboratory experiment is deemed most suitable. This approach allows for a partially realistic representation of workplace behavior while maintaining a high degree of control over specific influencing factors, allowing for investigating causal relationships.

A frequently mentioned limitation is the external validity of experiments (Guala, 2005). Within experiments, it is necessary to simplify real-world conditions to measure and vary the dependent variables. For example, in the experimental study of Weber and Camerer (2003), the authors simplified organizational culture towards the language established and used in teams. The concern of external validity is further enhanced by the reliance on student-dominated samples, as students are assumed to lack relevant professional experience and may therefore behave differently from individuals in the labor market (Villeval, 2016). Nonetheless, research indicates that data generated by students is rather reliable when compared to professionals (e.g., Exadaktylos et al., 2013; Falk & Heckman, 2009). Additional concerns include limited monetary stakes. Therefore, results of the following experimental studies should be transferred to the real world with caution to their generalizability.

PART I: CONCEPTUAL AND EMPIRICAL FOUNDATIONS

The first part of this dissertation provides an in-depth discussion of the basic underlying concepts and fundamental literature relevant to the essays, exceeding the level of detail feasible in the subsequent essays of this thesis.[2] As the essays all focus on the aspect of creativity, this field of research is presented focusing on defining creativity, current literature, and different methods of measuring creative performance. Thereafter, a general overview of research on workplace factors in focus of the dissertation, i.e., workplace and incentive systems, is given and the connection to creative performance is made. Essay-specific literature and theories are presented at suitable positions in the other parts.

1 Creativity

The following sections introduce the concept of creativity by presenting several definitions and related literature. Furthermore, different ways of assessing creativity, i.e., the UUT and the Droodle-Task, in an experimental setting are introduced as well as the general debate on how to measure creativity, i.e., objectively or subjectively. This section sets the required basic understanding of the world of non-routine work.

1.1 Definitions and current research topics

The concept of creative work is the main focus of this dissertation. Creativity as a concept is often intertwined with intelligence, as it was thought to exist a direct correlation between those two, leading to a focus of psychologists on intelligence rather than on creativity. Intelligence was considered a predictor of real-world creativity, as highlighted by various studies (e.g., Harris et al., 2019; Sawyer, 2011; Vincent et al., 2002). This notion and focus on intelligence in research began to shift in the middle of the 20th century, notably initiated by Guilford (1950) who tended to treat creativity and the construct of problem-solving as the same topic (Guilford, 1967). Thus, creativity is understood as a separate concept, whereas the focus of this thesis is not on creativity as a personality trait but as performance outcome. The emerging field of creativity research is remarkably divers, ranging from cultivating and identifying creativity in early education as a trait to the facilitation of creativity-enhancing leadership behaviors and workplaces focused on creative performance.

[2] As the individual essays are self-contained contributions that address different facets of the same overarching topic, i.e., creative performance of teams, some overlap in definitions and theoretical constructs across this dissertation is inevitable.

Distinguishing between creativity and innovation can be somewhat challenging, as these concepts are frequently conflated. One effective approach to differentiate these two constructs involves defining innovation as the result of harnessing creativity, coupled with the successful implementation of those ideas (Ford, 1996). Creativity, as a driver of innovation, plays a pivotal role in gaining competitive advantage in various domains of the business world (e.g., Ko & Butler, 2007; West & Sacramento, 2023).

Creativity itself is often understood as a process that combines elements that had not been combined before (Feinstein, 2011). Mumford and Gustafson define creativity as a complex construct influenced by multiple factors, ranging from the generation of an idea to the delivery of a completely new and innovative product (Mumford, 2003; Mumford & Gustafson, 1988). Weisberg indicates that creativity can solely be defined as something new to the inventor, without the need for the new to be valuable (Weisberg, 2006). Other definitions, more narrowed down to the workplace, define creativity as the generation of novel and useful ideas by employees concerning procedures and processes used at work (Zhou, 2003). Alternatively, as the ability to produce work that is both new and valuable, with creativity being a thought process itself (Martens, 2011). In this work, creativity is defined in line with the widespread definition of Amabile (1988), as the production of novel and useful ideas by an individual or a small group of individuals working together. This definition has prevailed among a variety of others (e.g., Zhou, 2003) and also includes creative teams. All those definitions include the requirement for something to be new, indicating that unconventional thinking, out of the box of the existing norm, is needed (Weisberg, 2006).

One important differentiation regarding creativity is the distinction between convergent and divergent thinking. Convergent thinking describes the process of finding the single best solution to a problem and elaborating on this solution as effectively as possible.[3] Divergent thinking, in contrast, focuses on generating as many new ideas as possible and is viewed as a suitable indicator of the creative potential of individuals (Guilford, 1950). It is perceived as a type of thinking that takes place before convergent thinking comes into play, as divergent thinking produces a wide variety of ideas and convergent thinking can narrow those ideas down to one solution or product (Guilford, 1950; Weisberg, 2006). The notion of divergent thinking as an indicator of creativity and the accordingly developed tests have prevailed in (experimental) research (Plucker et al., 2011; Runco & Acar, 2012).

[3] See Harvey & Berry (2023) for a meta-theory about usefulness and elaboration in the creative process.

The concept of divergent thinking covers four categories: fluency, flexibility, originality and elaboration (Guilford, 1959, 1960). Fluency describes the ability to generate a large number of new ideas, whereas flexibility is about covering different categories in which the generated ideas fit. The component of originality is also known as the uniqueness of the generated ideas, referring to the rarity of the given ideas. The last category, elaboration, describes the ability to transform abstract ideas into solutions that are implementable in the real world. Some critical notions on the construct of divergent thinking are presented in the section covering methods to assess creativity.

Other models and concepts view creativity, or the development of a creative product, as a result of different components coming together. Amabile's componential model suggests that an individual needs three types of skills to be creative: domain-relevant skills; creativity-relevant skills and general task motivation (Amabile, 1983). Lubart and Sternberg (1995) differentiate this model in their investment theory, indicating that a creative person is able to buy ideas at a low cost and sell them high. They highlight different abilities a creative individual needs, such as knowledge, intellectual abilities, a global style of thinking, a risk-taking personality and intrinsic motivation, as well as circumstances, as the authors also implement the need for an environment which supports new ideas (Lubart & Sternberg, 1995).

Some research emphasizes the need to view creativity within a specific context, considering not only the direct context but also culture and geographical differences when researching the field of creativity (e.g., Cabra & Guerrero, 2022; de Vries & Lubart, 2019). Meanwhile, Sawyer states that creativity does not only vary between cultures but also over historical time periods (Sawyer, 2011).

Existing research focuses, among other things, on the workplace and the environment in general potentially affecting creative performance (Guilford, 1967). Amabile considered three components of the work environment that influence creativity: organizational motivation, resources, and management practices (Amabile, 1988). Moreover, from a more physical perspective, it is found that possibilities of distraction as well as the social climate at work are important for creativity (Stokols et al., 2002) and that, in general, overly warm, loud and small workplaces can inhibit the creativity of employees (Martens, 2011). More recently, the focus shifted towards virtual workplace characteristics affecting creative teams, like communication channels (Brucks & Levav, 2022; Grözinger et al., 2020).

Another factor of interest in the world of economic research is the potential effect of incentives, tournaments and rewards on creative performance. Bradler, Neckermann and Warnke (2019)

investigated the effect of rewards on individual performance in simple and creative tasks and found a beneficial effect of a bonus payment for both tasks, whereas a gift was not effective in the creative task. Charness and Grieco (2019) studied the effect of different incentive schemes on the performance in closed and open creative tasks, finding that no payment scheme had a significant negative effect on individual creative performance, while piece-rate payment worked best in an open creativity task and overall creativity was higher in incentivized closed tasks. Contrary, results exist that incentives in the form of piece-rates and competition increase individual effort but not the level of creativity, whereas competitive incentives work counter-productive for women (Erat & Gneezy, 2016). In contrast, a field study find beneficial influences of rewards on the quality of innovative and creative ideas of teams, with a decrease in quantity, indicating a possible trade-off between quantity and quality when trying to incentivize ideas (Gibbs et al., 2017). Tournaments are shown to increase creativity in individuals to a certain extent but can lead to a crowding-out effect of intrinsic motivation to invest creative effort when the competition gets too intense (Gross, 2020).

1.2 Unusual Uses Task

After defining what is understood as creative, it is necessary to assess how it can be measured in a controlled setting. In psychology and economic research, several creative tasks have been developed over the last decades. One important differentiation when it comes to assessing creativity is to decide whether one wants to approach creative performance with the help of convergent or divergent thinking as indicator for creative potential and capacity. In general, three approaches to measure creativity exist: measuring with the help of achievement inventories and creative product evaluation, personality tests, and cognitive tests, whereas a broad range of different measures positively correlate with each other (Carson et al., 2005). Some approaches opt for measuring creative potential via questionnaires covering the personalities of participants, like the CPS (Gough, 1979), CAQ (Carson et al., 2005) or the TCPQ-12 (Penagos-Corzo & Saucedo, 2021). All those questionnaires have in common that they rely on the self-evaluation of participants and refer to needed personality traits or the ability to be creative and not actual creative behavior of individuals. Especially the personality trait of being open for new experiences is often connected with creativity (Carson et al., 2005; Feist, 2019; Goldberg, 1992; King et al., 1996; McCrae, 1987).

When researching actual creative behavior, measures in the form of real effort tasks exist for research in the lab or field.[4] A task used to measure creativity via convergent thinking is for

[4] See Charness et al. (2018) for a list of some tasks.

example the so-called *Remote Association Test* where participants get three words and have to find the overarching category in which all three words belong (Mednick, 1962). Another possibility to assess creativity is to let participants write a creative story (Charness & Grieco, 2019) or subject individuals to an escape room scenario where the way out can only be found with creative ideas and actions (Englmaier et al., 2024a; 2024b). All those covered approaches to measure creative ability and behavior are almost exclusively used at an individual level. A common approach to measure creativity at team level in the field is to use student courses at university where students have to work on group projects over the time of the course and their final solutions are evaluated regarding uniqueness and creativity, (Huo, 2020). No such commonly used task for creative teams in a more controlled environment exists.

One of the most famous tests to assess divergent thinking and with that testing for general creative thinking skills, is the Torrance Test of creative thinking (TTCT). The test was developed for children and early recognition of creativity and is a further development of Guilfords (1967) creativity tests. The test consists of a verbal and figural part (Torrance, 1966). The figural part of the TTCT includes three activities in which subjects have to construct a picture, complete an existing picture and use repeated figures of lines and circles to create a picture. The TTCT-Verbal in contrast consists of five tasks in which individuals get a picture and have to guess causes, guess consequences, state product improvements, give unusual uses for everyday objects, i.e., the UUT, and answer just-suppose situations. This part of the TTCT assesses the three dimensions of divergent thinking: fluency, flexibility and originality (Torrance, 1966, 1974).

In the first two of the four essays of the dissertation, the UUT as measure for divergent thinking is implemented. In the UUT individuals have to generate as many creative and useful ideas as possible for everyday objects, like a tin can, whereas a creative use is a use for which the object was not designed for. This task is often used to assess the quantity and quality of creative ideas and functions as an indicator for general creative performance (Runco & Acar, 2012). It closely mimics work which does not fall into the categories of mundane or dull tasks (Dutcher, 2012) and resembles work which is expected in creative and innovation oriented careers (Woodman et al., 1993). This task is often implemented in experimental laboratory research, where individuals have to solve one or more UUTs in a given time frame. More specifically, the UUT is argued to be a task which enables researchers to measure creativity in a domain-free context and with that it is a suitable instrument for research in a laboratory (Dutcher & Rodet, 2023).

The implementation of the UUT is diverse as no clear guide, which researchers had agreed on, exists. The execution of an UUT can for example differ in the time participants have to work

on one object. The range commonly lays between 2 up to 10 minutes per item, with a bandwidth between 5 and 10 minutes used most commonly. The items used in the tasks are diverse too, with the tin can as mostly stated object, at least as example task. Other frequently used objects are a brick, a piece of paper, an umbrella, or a cardboard box (e.g., Brucks & Levav, 2022; Dutcher & Rodet, 2022; 2023; Kurtzberg, 1998). Some instructions only stress that as many uses as possible need to be found (e.g., Baird et al., 2012; Manske & Davis, 1968), whereas other instruction focus on the component of being creative or unusual (e.g., Bradler et al., 2019; Rodet, 2021). As the UUT was developed as part of the TTCT which was designed to assess the creative potential of individuals, it is mainly used in an individual context. Sometimes, the task is also used in a team environment but often in a context where team members still work separately on the tasks (e.g., Grözinger et al., 2020).

Divergent thinking tests in general and the UUT in particular underly some critical notions, as they only measure one component of creativity and miss the component of idea implementation, following the definition of Guilford (1950). Another argument in this debate is that there exists some evidence that the creative thinking process does not include generating a large number of ideas at the beginning of a creative process and with that the UUT cannot be an indicator for creative potential (e.g., Weisberg, 2006). Weisberg suggests that a correlation between divergent thinking and creative performance only exists, because both components are influenced by the same factor, namely a need to be different and no real underlying connection between divergent thinking and creativity exists. Moreover, no official standards and ways to measure the UUT were agreed on upon researchers, leading to the possibility of different results, even if using the same task (Plucker, 2023).

Nonetheless, several researchers see no reason for such harsh criticism on divergent thinking tasks like the UUT, as divergent thinking was able to be shown as a measure with high predictive validity concerning creative achievements in the future (Plucker, 2023; Runco et al., 2010). Other researchers see the point that divergent thinking is not the same as creative thinking, but it is a measurable indicator to assess creativity and with that its usage is validated (Runco & Acar, 2012). Mumford (2003) mentions the importance of considering the research goal when choosing a creativity measure, as research with the aim of analyzing creative potential in children for their future educational path and research investigating creative performance of teams in the workplace should consider taking different measures. Another favorable aspect when debating about the UUT is that it is a rather general task which does not involve special skills and is thus detached from the notion that creativity is only domain specific (Plucker, 2023; Sawyer, 2011), making the UUT implementable in a mixed sample, like a pool of students

(Dutcher & Rodet, 2022). Moreover, the UUT generates a high variance of output, making it easy to detect significant treatment effects (Dutcher & Rodet, 2021). Furthermore, it is implementable on a team level, as the generated uses can be discussed, communicated and written down in a team during the task. Additionally, in a comparison of three divergent thinking tasks, the UUT shows the highest reliability of ratings during the subjective evaluation process, highlighting common understanding of individuals involved in this task (Silvia et al., 2008). Additionally, it scores high on measures of validity and reliability (Carson et al., 2005; Silvia, 2011). Moreover, Plucker (2023) criticizes the high rate of new instruments of the same quality to measure creativity being developed, even if highly functional and repeatedly used models and tests, like the TTCT, exists, leading to less consensus in research and low comparability of results.

1.3 Droodle-Task

An alternative task to assess divergent thinking with is the Droodle-Task, which is rather unknown in experimental economic literature. The origin of Droodles can be traced back to the American humorist and author Roger Price, who published his first collection in 1953. In it, Price describes doodles as drawings that make no sense until the viewer knows the title (Price, 1953; Price & Lovka, 2000). A Droodle is an abstract picture without color and no clear objects shown, allowing for several interpretations and perspectives (Figure 1). In research, Droodles have been used primarily in experiments with children. Brand (1978) and Doherty and Wimmer (2005) used Droodles to investigate the ability to empathize with others in children. Kahn et al. (2008) asked subjects to find the most appropriate and imaginative title for various Droodles within a given time and to create Droodles themselves under time pressure. Ali et al. (2019) asked children to name as many unique and creative ideas for a Droodle as they could think of in 30 seconds. None of these experiments considered the effects of teamwork or incentive systems on divergent thinking. The design of the experiment developed in this dissertation therefore differs from existing literature and contributes to the usage of Droodles to capture divergent thinking.

The implemented Droodle-Task in the third and fourth essay is to find as many creative titles describing a given Droodle as possible and type these into a text box (Ali et al., 2019; Kahn et al., 2008; Kaivola-Bregenhøj, 2001). Thereby, participants are not restricted in the perspective they take as they are able to rotate the picture on their screen. For example, creative solutions for the Droodle in Figure 1 could be “Blender rotors” or “Four elephants sniffing at an orange”.

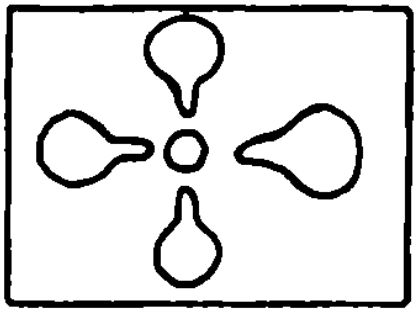


Figure 1: Example Droodle

The task is similar to the UUT, with respect to the process of idea generation. In both tasks participants receive a guideline, either an object or a picture, on what the divergent thinking process should focus on. Additionally, both tasks allow for several highly creative solutions which can either be given with a single word or a more detailed description, making the direct verbal exchange of ideas between team members easy to handle. Thus, both tasks enable direct verbal teamwork. Moreover, answers given for a Droodle also cover the categories of fluency, flexibility and originality which are crucial for defining divergent thinking (Torrance, 1966, 1974).

1.4 Measuring creativity

Next to specifying the task used to capture creative performance, it needs to be set how one can measure creative performance. A debate about the correct or at least most favorable way to assess the results of divergent thinking tasks exists, with the main concern if the use of subjective raters, or a more objective and quantified approach is most suitable and if so, how a quantified measure can look like (Reiter-Palmon et al., 2019; Silvia, 2011; Weisberg, 2006). Nonetheless, subjective evaluations are the commonly used method to evaluate creative performance (Silvia, 2011; Silvia et al., 2008). Reiter-Plamon et al. (2019) suggest that subjective ratings can work, if the raters are well trained and experienced.

When opting for the objective approach, the three categories of creativity can be quantified, as frequency equals to the quantity of valid answers, flexibility is measured as different categories which are covered by the given answers and originality can be measured with the rarity with which answers are given by the whole sample of participants. However, the decision process made beforehand, what is a valid answer and in which category does it fall into, is subjective and can be biased.

Table 2: Objective measures of creative performance

Measure	Weight	Description
Quantitative Performance		Sum of points generated for valid answers
	1 point	given for answers net of answers that are counted as invalid, because they were given twice, are not understandable or include the intended and common usage of an object
Qualitative Performance		Sum of the generated points by covered categories, rare and very rare answers
Categories	1 point	given for each category covered with the answers for one UUT. In total, 29 categories were determined beforehand during a pre-test. Categories are for example: Sports, animal, cooking, construction, fabric or crafting.
Rare answers	2 points	given for rare answers, i.e., answers that were given only three times for one specific UUT.
Very rare answers	4 points	given for very rare answers, i.e., answers that were given maximum twice for one specific UUT.
Errors	0 points	given for answers which include the intended usage of an object, were given twice or were not comprehendible.

In the first two essays, the performance when working on a UUT is assessed using the objective evaluation method of Bradler et al. (2019). The objective measurement method allows for forming two measures of creative performance of teams. Table 2 gives an overview of the measures of creative performance and the corresponding performance points assigned to them for the UUTs. Firstly, a quantitative measure consisting of the number of valid answers given during one UUT is computed, assigning one performance point to each valid answer, which is in line with Bradler et al. (2019). Secondly, a qualitative performance measure consisting of the number of different categories covered and frequency of answers is computed. The subjects receive one point for each category covered (like sport, cooking, animals, etc.), two points for rare answers, which are answers that are given only three times, and four points for very rare answers, which are answers that are only given once or twice when solving the specific UUT.[5] Those evaluations are conducted manually by student assistants. Responses that are far away from implementation and do not have any reasonable use, as well as nonsense, and redundant answers, are excluded, following the recommendation from Reiter-Plamon et al. (2019). As objective measures sometimes risk underlying subjective decisions when defining the rating tables and counting words in the UUTs (Reiter-Plamon et al., 2019), all decisions are documented and double-checked and a guideline for deciding for similarity or defining the category

[5] Those weightings of rare and very rare answers are oriented to the weighting of Bradler et al. (2019) who doubled the weighting of the very rare answers in contrast to rare ones, as they are the main indicators for creativity.

has been developed upfront. A total of 29 categories has been predefined beforehand (see AI. 14 in Appendix I).

For the Droodle-Task implemented in the third and fourth essays, the evaluation method shifts to subjective ratings using the consensual assessment technique (Amabile, 1982), conducted by a jury of four independent assistants. Jurors first determine individually the validity of each answer, i.e., whether it is comprehensible, before rating its creativity on a scale from 0 to 3 points. Overall creative performance per answer is then calculated as the sum of points awarded by each juror, plus additional points for validity. To ensure consistency, jurors are trained on pre-test data prior to rating and assess answers independently of one another, following the recommendations of Reiter-Palmon et al. (2019). Together, the evaluation approaches employed across the four essays cover both objective and subjective assessment methods, offering a comprehensive methodological overview of how creative team performance can be measured through different divergent thinking tasks.

2 Creativity and the workplace

Having established how creativity is conceptualized and measured, this dissertation now turns to the contextual factors shaping creative performance in teams. A first and increasingly relevant factor is the physical location of work. The following sections introduce the concept of WFH, outline its key characteristics and empirical effects, and examine what existing research reveals about its relationship with creative performance

2.1 Working from home

Remote work, also referred to as telework or telecommuting, has attracted substantial scholarly attention over the past decades, driven by technological advances enabling digital work and, more recently, accelerated by the COVID-19 pandemic. Within this concept, a key distinction exists between work from anywhere (WFA) and WFH (Kurland & Bailey, 1999). WFA describes arrangements in which employees are entirely unconstraint in their choice of location, offering complete flexibility across region or countries, time independence and increased autonomy (Charalampous et al., 2019; Choudhury et al., 2021). The dissertation here presented focuses specifically on WFH, where employees perform their work task at home rather than at a company site (Kurland & Bailey, 1999; Lim & Teo, 2000). WFH is closely associated with the use of information and communication technologies, including portable devices, internet connectivity, and video conferencing tools, which make this form of work practically feasible (Messenger & Gschwind, 2016; Stanworth, 1998).

In practice, hybrid forms of WFH, i.e., employees split their working days between home and the office, are a widespread specification of WFH (Gajendran & Harrison, 2007). A growing body of evidence suggests that such hybrid models can yield meaningful benefits for both employees and organizations. In a field study conducted in China, Bloom et al. (2015) find that hybrid WFH leads to a 13% performance increase among call center workers, a 9% increase in daily working hours, and a significant reduction in absenteeism, provided that adequate working conditions at home are available. Similar preferences for hybrid arrangements have been documented in other national contexts: a study of 10,000 Polish workers finds that employees are willing to forgo, on average, 5% of their earnings in exchange for two to three WFH days per week (Lewandowski et al., 2022), while comparable results emerge from a large US sample of 30,000 employees willing to sacrifice 7% of gross salary for a hybrid model, with associated self-reported productivity gains of 4.6% (Barrero et al., 2021). Together, these findings suggest that WFH is not merely a worker preference but can translate into tangible organizational benefits under the right conditions.

The effects of WFH are, however, not uniform across employees, as individual and demographic characteristics can play a moderating role. Gender differences in WFH preferences have produced mixed results across contexts: while Mas and Pallais (2017) find that women with children place higher value on WFH, a German study reports that employees with children actually engage in remote work less frequently (S. Zhang et al., 2020), and a Singaporean study finds no significant gender-based difference in WFH preferences at all (Lim & Teo, 2000). Beyond gender, household size, lower educational attainment, and lower household income are negatively associated with the uptake of mobile work, while employees aged 35 and older engage in remote work more frequently. Those patterns have already been evident in US data from 2002 (Walls et al., 2007). At the level of personality, conscientiousness is consistently linked to both stronger preferences for hybrid WFH (Clark et al., 2012) and higher performance in WFH settings (Gavoille & Hazans, 2022), while extraversion shows the opposite pattern (Gavoille & Hazans, 2022). These findings underscore the importance of accounting for individual heterogeneity when assessing the implications of WFH policies.

When research focuses on the effect of a virtual workplace, virtual teams are a natural unit of analysis, as teamwork inherently requires communication, which is the key distinction between work in presence and WFH. Such teams are defined as an assembly of individuals, often dispersed across various geographical locations and potentially operating in distinct time zones, who unite to collaborate on a shared project. They heavily rely on information and communication technologies, CMC in general, to jointly create and work together (Garro-Abarca et al.,

2021; Horwitz et al., 2006; Peñarroja et al., 2013). Studies on remote teams underscore the importance of initial or at least frequent FtF interaction (Coenen & Kok, 2014; Horwitz et al., 2006; Kennedy et al., 2010; Paulus et al., 2012), with FtF contact proving particularly valuable in competitive rather than cooperative team environments (Hill, et al., 2009).

A critical determinant of virtual team success is the level of trust between team members, as trust enhances the quality of communication and interaction, particularly in virtual settings (Chang et al., 2011; De Guinea et al., 2012). Notably, however, Peñarroja et al. (2013) find no significant difference in trust generated through FtF versus video conference interaction. Building such trust can nonetheless prove challenging, as virtual communication tools may feel less personal than in-person interaction (Kennedy et al., 2010), a notion supported by media richness theory (Daft & Lengel, 1986). These challenges are especially pronounced during the initial formation phase, where virtual teams frequently struggle with communication and coordination before such difficulties gradually diminish over time (De Guinea et al., 2012; Lurey & Raisinghani, 2001). Despite its relevance, the initial formation phase of virtual teams remains understudied, as the existing literature predominantly focuses on teams that are already established.

2.2 Creativity and virtual work

Beyond the general effects of WFH discussed above, the work environment has been identified as a key determinant of individual creative performance. For example, workplace conditions such as reduced monitoring and the presence of creative coworkers are shown to enhance individual creativity (Zhou, 2003), while leadership style is similarly identified as an influential determinant of creative behavior among employees (e.g., Hughes et al., 2018; Leroy et al., 2022). Work design, particularly the role of goals and time constraints, represents another key factor shaping individual creative output (Rodet, 2021). Extending this line of inquiry, a growing body of research has begun to examine whether the physical location of work exerts a systematic influence on creative performance.

Initial evidence on work location and creativity comes from Dutcher (2012), who compares individual performance on dull and creative tasks in two settings: working in a lab versus working remotely. Results show a significant increase in creative performance of 11–20% outside the lab, while performance on the dull task decreases significantly in the same setting, with both effects being predominantly driven by male participants. These findings suggest that the optimal work location is task-dependent rather than universally fixed. Complementing this, Hunter (2019) argues that remote work fosters creativity in scientific research by improving work-life

balance, given that greater personal well-being is associated with higher creative output. Taken together, these studies provide early evidence that WFH may be particularly conducive to creative work at least for individuals; a premise that the present dissertation examines further in a team context.

As creative performance of teams is characterized by active communication and exchange of ideas (Feinstein, 2011), virtual work might affect such teams' output. One of the first studies on creativity of virtual teams examines the effect of CMC via chat versus FtF interaction on the creative output of four-person teams (Kerr & Murthy, 2004). Results show that CMC outperforms FtF communication in divergent thinking tasks, while the reverse holds for convergent thinking tasks; nonetheless, qualitative creative performance is highest during FtF communication. More recent studies largely reinforce the advantage of FtF interaction for creative team tasks. Dutcher and Rodet (2023) find that a FtF team outperforms both a chat-based team and individuals, though no direct comparison between individuals and teams is drawn and communication was limited to chat than videoconferencing tools. Brucks and Levav (2022) demonstrate that video conferencing inhibits team creativity relative to FtF interaction, as the medium leads participants to focus on the screen, resulting in fewer distal associations. In contrast, Grözinger et al. (2020) find no significant difference in creative performance between virtual and FtF teamwork.

It should be noted, however, that much of this research predates the widespread adoption of chat tools and video conferencing in both professional and private life, which may limit the generalizability of these findings to contemporary work settings. Moreover, the existing literature predominantly examines one-shot teams or teams that are already formed and functioning, leaving the initial virtual formation phase largely unexplored. These gaps motivate the first two essays of this dissertation, which specifically examine the creative performance of newly formed virtual teams.

3 Pay-for-Creativity

Next to the physical workplace, this dissertation also covers another essential component of work conditions: payment. By this additional, and highly relevant component of working life, this dissertation allows a broad understanding of working conditions for teams in the field of non-routine tasks, potentially fostering innovation.

Incentives can broadly be categorized as monetary or non-monetary, with the latter further divided into tangible and intangible forms. A further distinction concerns what is being incentivized, i.e., effort provided or output generated, and whether the incentive structure is discrete, continuous, or relative (Lazear, 2018). The focus of this thesis lies on monetary PfP incentives. The following sections introduce the general concept of PfP, provide an overview of the broader research landscape, and discuss its specific implications for creative work.

3.1 Pay-for-performance incentives

The concept of PfP incentives is grounded in the theoretical framework of Principal–Agent Theory (Holmström, 1979; Prendergast, 1999). In the classical setup, a principal delegates a task to an agent whose actions are only imperfectly observable. This informational asymmetry can create a moral hazard problem, as the agent might shirk or deviate from the principal's interests, which is to produce as much output as possible. In a team setting this can lead to free-riding behavior. To mitigate this, principals design variable payment schemes that tie compensation to measurable performance outcomes. The optimal contract model distinguishes between first-best and second-best solutions, depending on whether effort is observable or not. In the second-best case, variable pay becomes a mechanism to align incentives and increase effort, albeit at the cost of exposing the agent to risk.

A well-studied form of PfP is the piece-rate scheme, where workers are paid directly for each unit of output produced. Lazear (2000, 2018) demonstrated that piece rates can generate both incentive effects (increasing individual effort) and selection effects (attracting more productive workers and dropping less productive ones). However, piece-rate systems also expose workers to exogenous risks, such as fluctuations in demand or production conditions, that affect output independently of effort. As a result, the agent's risk aversion becomes a crucial consideration for designing optimal incentive schemes, balancing motivational benefits against the disutility of income variability.

Another particular form of PfP incentives are rank-order tournaments (Bull et al., 1987; Lazear & Rosen, 1981; Nalbantian & Schotter, 1997), prominently used for promotions. In such settings, individuals or teams compete against each other, whereas performance is assessed relative to other competitors. Therefore, rewards are allocated based on ordinal ranking rather than absolute output. The basic mechanism is analogous to promotion tournaments in firms, where winning a contest grant not only a monetary prize but also opportunities for higher follow-up tournaments. The prize spread, i.e., the difference between winner and loser rewards, deter-

mines the strength of the incentive effect (Harbring & Irlenbusch, 2008). A larger spread typically induces greater effort but can also increase unproductive behaviors such as sabotage or collusion between competitors.

An extreme version of tournament incentives is the winner-takes-all structure. In those tournaments, no loser prize exists, such that the potential risk when participating is higher. Krishna and Morgan (1998) analytically show that such contests are optimal when competition involves only a few participants (typically two or three), as the marginal gains from additional competitors diminish rapidly. Such tournaments are often applied in non-routine or innovation-driven contexts, for example at logo-design platforms or public architecture tenders (Terwiesch & Ulrich, 2009).

A substantial body of research has further examined gender differences in both preference for and performance under tournaments. The seminal study by Niederle and Vesterlund (2007) finds that men tend to enter tournaments more often than women, largely due to higher levels of overconfidence. Subsequent research by Geraldes (2020) reveals that when women compete only against other women, or can choose the gender of their opponents, the gender gap in competitiveness disappears. This suggests that women's lower entry rates in mixed-gender tournaments partly reflect belief distortions about others' preferences, creating a self-fulfilling prophecy. Buser et al. (2023) provide further evidence that gender differences in tournament participation stem primarily from women's reluctance to compete against men whenever alternative options exist.

When tournaments are organized among teams, additional behavioral dynamics in teams emerge. Majerczyk et al. (2019) show that intergroup competition can simultaneously foster individual effort and intra-team cooperation, though at the potential cost of increased free-riding over time. Sutter and Strassmaier (2009) find that communication within teams enhances cooperative behavior and improves overall team output under tournament conditions. Similarly, Ai et al. (2023) report that workers in the gig economy exert higher effort and work longer hours when embedded in teams competing against other teams. Eckel and Grossman (2005) demonstrate that team contests are among the most effective mechanisms to strengthen team identity in laboratory settings. In a meta-analysis, Condly et al. (2003) find that team-directed incentives yield performance increases of up to 48%, compared to 29% under individual PfP, an effect attributed to psychological processes that enhance collective motivation.

The gender composition of teams also affects behavior and performance in tournaments. Hardt et al. (2025) show that gender diversity within teams influences information exchange and coordination quality. Experimental findings by Ivanova-Stenzel and Kübler (2011) indicate that performance differences between men and women are most pronounced in single-gender competitions and diminish as team diversity increases.

Over the past decades, several meta-analyses and literature reviews have synthesized the extensive research on monetary incentives (e.g., Bonner & Sprinkle, 2002; Cerasoli et al., 2014; Holmström, 2017; Prendergast, 1999). The consensus emerging from this literature is that PfP schemes generally enhance effort and performance in routine tasks, where outputs are easily measurable. However, exceptions exist. Empirical evidence shows that stronger incentives do not always translate into higher performance. For example, Ariely et al. (2009) documented cases of choking under pressure, where excessively high stakes can impair performance due to inefficient increases of effort. Gneezy and Rustichini (2000) found a negative effect of introducing (low) piece-rate incentives on the effort provided for tasks, which are rather intrinsically motivated. For a more detailed overview of incentives in such a context, which creativity also belongs to, see the chapter below.

Finally, recent developments in machine learning have opened new avenues for designing incentive systems. Opitz et al. (2024) demonstrate that algorithmically assigned, personalized incentives can significantly outperform uniform pay schemes, particularly for individuals with clearly defined personality traits. Such adaptive incentive systems represent a promising frontier for aligning pay structures with individual motivational profiles, potentially enhancing performance in both routine and non-routine tasks.

3.2 Creativity and pay-for-performance incentives

Moste studies and research results mentioned in the section above base their findings and conclusions on the results of PfP in the context of routine tasks, where effort and output are measurable and performance criteria unambiguous (e.g., Lazear, 2000). Nonetheless, considering results and theories for non-routine tasks, like divergent thinking, is important as the complex motivation, i.e., intrinsic vs. extrinsic motivation, for and the specific characteristic of non-routine tasks, i.e., risk, may change the effectiveness of PfP systems (Gerhart & Fang, 2015; Gibbons, 1998; Jensen & Meckling, 1976; Kim, 2025). Consequently, the mechanisms through which PfP incentives operate, and their eventual effectiveness may substantially differ in such contexts. In the following section, theoretical considerations and recent research findings in this field are presented.

Engaging in creative tasks entails a higher degree of risk and uncertainty than routine tasks. This is for several reasons: First, the outcome of a creative task (partly) depends on external evaluation, whereas creativity is perceived subjectively different (Dewett, 2006). Second, individuals may be unsure about their own ability to solve creative tasks, as such tasks are non-routine and less frequently encountered in daily work settings (Kim, 2025). Third, and more importantly, in a competitive environment, individuals face additional uncertainty regarding their relative performance, since they cannot easily predict the creativity of their competitors. As discussed earlier, variable pay exposes agents to income risk, which interacts with individual risk aversion (Jensen & Meckling, 1976; Prendergast, 1999). In creative contexts, the combined effect of task-related uncertainty and pay variability may attenuate the positive incentive effects of PfP schemes. Additionally, the intensity of such PfP incentives needs to be considered due to this described interplay between incentive and risk effects.

Another essential theoretical perspective when examining PfP in creative contexts is motivation theory, particularly SDT and its subcomponent, Cognitive Evaluation Theory (CET) (Deci et al., 1999; Deci & Ryan, 1985; Ryan & Deci, 2000). These frameworks differentiate between intrinsic motivation, i.e., engaging in a task for its inherent satisfaction, and extrinsic motivation, i.e., engaging in a task for external rewards or recognition. CET posits that external rewards can, under certain conditions, undermine intrinsic motivation, a process referred to as "crowding-out". Since creativity is typically driven by intrinsic motivation (Amabile, 1988), PfP incentives may risk reducing creative performance by shifting the locus of motivation from internal to external. Notably, CET was primarily developed to explain motivational dynamics rather than performance per se (Gerhart & Fang, 2015).

Empirical findings over the past two decades have nuanced this rather pessimistic view. Eisenberger and Shanock (2003) show that expected monetary rewards can enhance creativity, if participants understand that creativity is required and valued. Similarly, Byron and Khazanchi (2012) find in a meta-analysis that creativity-contingent rewards, which explicitly emphasize the value of originality, tend to increase creative performance. In another study, the perceived relevance and meaningfulness of a reward is found to moderate the effect of extrinsic incentives on creativity (Fischer et al., 2019). Cerasoli et al. (2014) further demonstrate that the crowding-out of intrinsic motivation is less pronounced when rewards are directly tied to performance outcomes. Their meta-analysis suggests that intrinsic motivation predicts the quality of creative output, while monetary incentives predominantly affect quantity, yet intrinsic motivation remains a robust predictor of overall performance regardless of the presence of rewards.

Other studies support a synergistic view: extrinsic motivators may complement intrinsic motivation under certain conditions. Hennessey and Amabile (2010) and Malik et al. (2015) argue that extrinsic incentives can enhance creativity when intrinsic motivation is already high, suggesting an interaction effect rather than a simple substitution. Reflecting this evidence, Amabile and Pratt (2016) revised the original Componential Model of Creativity, introducing the concept of synergistic extrinsic motivation. This updated framework distinguishes between informational and controlling extrinsic motivators. Informational motivators, such as recognition, feedback, or rewards that signal competence, can enhance intrinsic motivation, whereas controlling motivators, such as rigid performance monitoring or coercive incentives, can diminish it. Consequently, certain forms of PfP, especially those framed as acknowledgement of creative value rather than external control, may foster rather than hinder creativity.

Beyond motivational considerations, theoretical work in organizational economics highlights an additional challenge for incentivizing creativity: the multitasking problem (Bonner & Sprinkle, 2002; Holmström, 2017; Holmström & Milgrom, 1991; Prendergast, 1999). Divergent thinking tasks often require a trade-off between quantity and quality of creative answers. When performance is assessed on a limited set of observable metrics (e.g., number of ideas generated), PfP systems risk incentivizing easily measurable aspects of creativity at the expense of less tangible but (at least) equally important dimensions (e.g., originality). This trade-off makes optimal incentive design in creative settings inherently complex, even before psychological factors such as intrinsic motivation or risk aversion are considered.

Recent empirical research examining the effects of PfP incentives on individual creativity and divergent thinking has produced highly heterogeneous findings. While numerous studies demonstrate positive effects of extrinsic rewards under certain conditions (e.g., Bradler et al., 2019; Charness & Grieco, 2019; Englmaier et al., 2024a; 2024b), others find null (e.g., Eckartz et al., 2013; Malik et al., 2015) or negative results (e.g., Attanasi et al., 2026; Gross, 2020). To explain the diversity of results in this research area, scholars have identified several task- and design-related factors that moderate the relationship between PfP and creativity.

One major distinction concerns the openness of a creative task. Charness and Grieco (2019) conducted two laboratory experiments comparing an open creative task (e.g., designing a city or proposing an innovation) and a closed creative task (e.g., solving a mathematical problem or composing a story with predefined words), both under incentivized and flat-payment conditions. Their results revealed that incentives increased creativity in closed tasks, while they were less relevant in open-ended tasks. Importantly, no evidence was found for any negative effects

of payment schemes, suggesting that PfP does not necessarily undermine creativity, even in open settings.

A related line of research differentiates between top-down, i.e., tasks with multiple solutions, and bottom-up, i.e., tasks with one solution, creative thinking. It is shown that performance-dependent pay can have negative effects in bottom-up tasks, but positive effects in top-down tasks (Kachelmeier et al., 2024). This distinction suggests that the cognitive structure of a task, and the degree of solution multiplicity it entails, critically determines whether PfP enhances or inhibits creativity.

Temporal aspects of the task also moderate incentive effects. Kachelmeier et al. (2019) distinguish between one-shot tasks and those allowing for incubation periods, in which participants can reflect or temporarily disengage before producing creative output. Their experiments show that, in one-shot conditions, quantity-based incentives yield higher creative performance than fixed pay. However, after an incubation period of 20 minutes, the differences between incentive conditions narrow. This finding suggests that incubation allows externally motivated effort to transform into intrinsic processing, reducing immediate incentive dependence. Interestingly, neither creativity-contingent nor minimum-rate payment schemes outperformed the quantity-based treatment, implying that simplicity and temporal flexibility of incentives can sometimes be more effective than complex reward designs.

Similarly, rather than differentiating task types, one can differentiate incentive focus, whether it targets input (e.g., time spent, effort) or output (e.g., quality or quantity of ideas). Blaufus et al. (2026) indicate that both types of incentives increase the time spent per task as well as quantity of ideas. Whereas ideas under output-focused incentives tend to be less complex and are developed faster. For incremental incentives, it is not the intensity of the reward that matters most, but rather its mere presence, which appears sufficient to activate performance-related motivation without undermining intrinsic interest. Others (e.g., Kachelmeier et al., 2008; Kachelmeier & Williamson, 2010; Laske et al., 2025) argue that when incentivizing output, focusing on quantity or trying is sufficient for generating creative outcomes.

In contemporary work environments, many creative tasks are carried out in teams, where social processes such as coordination and group cohesion play a decisive role in shaping outcomes (Ai et al., 2023; Charness & Chen, 2020; Pescosolido & Saavedra, 2012). The effectiveness of PfP incentives in these contexts is therefore not only a function of individual motivation but also of collective dynamics and interpersonal interdependencies. A particularly relevant form of PfP in team settings used in the practical field is the tournament incentive. Tournaments add

an additional social and comparative dimension to incentive schemes, potentially amplifying both motivational and behavioral effects. Team contests are among the most effective mechanisms to strengthen team identity in laboratory environments (Eckel & Grossman, 2005). They can influence individual effort, within-team cooperation, and between-team competition simultaneously, thereby potentially affecting creative performance.

Empirical findings describe how tournament incentives affect creative team performance. Bradler et al. (2019) show that introducing bonuses in a competitive environment effectively improves both routine and creative performance, without evidence of motivational crowding-out. In contrast, Erat and Gneezy (2016) report that while piece-rate and competitive incentives increase overall effort, they do not enhance creativity in a rebus-puzzle task, and that competitive incentives may even be counterproductive for women, echoing the gender-specific findings of Niederle and Vesterlund (2007).

An empirical field experiment on creative teams shows that tournament incentives increased team performance without reducing participants' willingness to engage in similar tasks again, indicating no evidence of crowding-out of intrinsic motivation (Englmaier et al., 2024b). In a complementary study, Englmaier et al. (2024a) found that PfP incentives are most effective for experienced teams, whereas team size has no significant influence on outcomes. The authors highlight the need for future research to systematically vary inter-team incentive structures, as different forms of team-based PfP may yield distinct effects depending on team composition, task type, and prior collaboration experience, a research gap that is directly addressed in the second half of this dissertation.

The intensity of tournament incentives plays a critical role in determining their effect on creativity. Ariely et al. (2009) famously document that very high incentives can lead to choking under pressure in different tasks, a phenomenon in which individuals' performance deteriorates when stakes become excessively high. This effect has been linked to over-motivation and heightened arousal, which impair cognitive flexibility. Field evidence by Gross (2020) from logo design competitions further illustrates the nuanced relationship between competition and creativity. In winner-takes-all settings, competition initially enhances originality, but excessive competition intensity leads to a decline in creative investment. The most favorable outcomes are observed under balanced competition, where incentives are strong enough to motivate effort but not so extreme as to induce stress or strategic conservatism. However, not all evidence points toward detrimental effects of strong incentives. Kim (2025) emphasizes that the relative

level of incentive intensity, rather than its absolute value, determines its impact. In her experimental work, high intensity rank-order tournaments significantly outperformed both low intensity tournaments and fixed-pay conditions for individuals performing a creative task.

To sum up, research on PfP incentives and creativity shows mixed results that differ from their generally positive effects on routine tasks. While SDT and CET suggest that monetary rewards may crowd out intrinsic motivation, driving creativity (Amabile, 1988; Deci & Ryan, 1985), many empirical studies do not find strong evidence for this negative effect. More recent work points to a synergistic view, indicating that certain types of extrinsic incentives, such as creativity-contingent, informational, or meaningful rewards, can complement intrinsic motivation and support creative performance (Amabile & Pratt, 2016; Eisenberger & Shanock, 2003; Fischer et al., 2019). Differences in findings across empirical studies can partly be explained by variations in task design, such as whether tasks are open or closed (Charness & Grieco, 2019), rely on top-down or bottom-up thinking (Kachelmeier et al., 2024), or are conducted as one-shot tasks or include incubation periods (Kachelmeier et al., 2019). Although most empirical studies examine individuals, the role of teams is increasingly recognized as relevant, particularly in competitive settings, where tournaments may strengthen group cohesion and coordination, potentially improving creative performance (Charness & Chen, 2020; Englmaier et al., 2024b). While incentive intensity appears to influence creative output, current evidence on its direction and strength remains inconclusive and will be examined further in the third and fourth part of this dissertation. Finally, most studies focus on individual creative performance rather than teams engaging in creative tasks, pointing towards a need for further research conducted with this dissertation.

Having established the fundamental concepts and existing empirical findings on creativity research, mobile work, and PfP incentives in the context of creative performance, the remainder of this dissertation presents four essays, each examining a distinct aspect of the work environment and its influence on team creative performance through the lens of economic experiments. To conclude this thesis, a short summary of key findings as well as broader ideas for future research in the field of creative teams, i.e., moving towards convergent thinking, the usage of AI, and a field experiment, are presented.

PART II: CREATIVE PERFORMANCE: A COMPARISON OF INDIVIDUALS AND TEAMS – AN EXPERIMENTAL STUDY

A shorter version of this essay has been published as: Grund, C., Harbring, C., & Klinkenberg, L. (2026). Creative Performance: A Comparison of Individuals and Teams – an Experimental Study. *Journal of Business Economics.*

Abstract: There is an increased relevance of non-routine creative work. The use of virtual teams within organizations comes along with communication and coordination challenges for the case of creative work in particular. We focus on the creative performance of newly formed virtual dyadic teams and investigate performance differences compared to individuals working virtually on a creative task. We conducted an online lab experiment, accounting for a work-from-home environment, comprising two working phases one week apart, extending the common focus on one-shot interactions. We make use of Unusual Uses Tasks (UUT) to assess creativity of dyadic teams and individuals. Given that this task is typically used at the individual level, our study contributes to the field of creative experimental tasks by implementing the UUT in a team setting with direct interaction and a virtual workplace. Our findings show that dyadic teams exhibit significantly higher creative performance than individuals across the two virtual phases hinting for the meaningfulness of the UUT approach on the group level. Additionally, we check for mechanisms exploring the performance differences between individuals and teams working on this creative task with regard to heterogeneity in performance, the role of errors next to productive performance as well as the role of trust and sympathy within teams.

1 Introduction

An increasing demand for non-routine and creative tasks is emerging in the workplace, as these are essential drivers of innovation and, consequently, a source of competitive advantage (Autor et al., 2003; Englmaier et al., 2024a). Although there is the notion that innovation is often driven forward by teams (West & Sacramento, 2023), we have little insights into the effectiveness of teams as compared to individuals in working on creative task in a fully virtual work environment.

There has been a notable increase in the adoption of technologies facilitating virtual communication and remote work in response to evolving organizational structures and task characteristics (Barrero et al., 2021; Brenan, 2020). This shift has resulted in more teams collaborating in

virtual settings, using digital communication tools. Due to the advancements in technology and the work-setting adaptations enforced by the COVID-19 pandemic since 2020, employees have become more familiarized with these tools and have adapted to the new demands and opportunities of flexible work models. There are already some studies analyzing the effectiveness of virtual teams (Brucks & Levav, 2022; Dutcher & Rodet, 2021; Kerr & Murthy, 2004). Previous research provides mixed findings concerning the impact of virtual work on performance and creativity. This applies to both individuals (e.g., Dutcher, 2012; Stockolos et al., 2002) and teams (e.g., Alsharo et al., 2017; Giambatista & Bhappu, 2010; Kennedy et al., 2010). Existing research on creative teams primarily explored the effect of team member diversity on creative output (e.g., Dutcher & Rodet, 2022; Hoever et al., 2012; Kurtzberg, 2005), incentives for creativity (e.g., Charness & Grieco, 2019; Erat & Gneezy, 2016), and the optimal communication methods and workplace settings for such teams (e.g., Brucks & Levav, 2022; Grözinger et al., 2020). Focusing on virtual teams, especially the stage of team formation is found to come with challenges in communication and team performance (e.g., De Guinea et al., 2012; Lurey & Raisinghani, 2001). Current knowledge about the initial formation phase of a virtual team and its influence on performance during the first interactions is insufficient (Reiter-Palmon et al., 2021). Additionally, there has been little research addressing the underlying, more fundamental question of whether it is at all beneficial to assign creative tasks to a team rather than to one or more individuals in a virtual work setting. This leads us to our research question:

Do newly formed teams show higher creative performance than individuals when working from home?

Besides, we aim to analyze the mechanisms through which teams perform successfully on a creative task. We explore the role of sympathy and trust as crucial components of interaction next to ability.

To answer our research question, an online experiment is conducted. In this experiment the creative performance of individuals and dyadic teams working on a non-routine creative task is compared.[6] The experiment consists of two experimental phases, as the study is especially interested in the creative performance of teams when those teams are newly formed. This formation of a team comes with the most problems regarding communication and building of trust, which, in turn, can affect creative performance (Abi Saad & Agogué, 2023; De Guinea et al., 2012).

[6] In this essay, a creative task refers to a task that requires divergent thinking, which is connected to the creation of many ideas and solutions for a given problem (see Part I chapter 1 for further detail). A creative team is defined as two people working in direct exchange of thoughts and ideas together on such tasks.

The remaining of this essay is structured as follows: Section 2 presents the relevant definitions and theories for this work, as well as an overview of current research in the fields of virtual work and creativity. Building up on those current findings and theoretical approaches, the hypotheses are derived. Section 3 describes the experimental design and section 4 presents the results, starting with descriptive results, hypotheses testing and finally further results regarding the potential underlying mechanisms, driving our results. In section 5 the results are discussed, implications for future research and practice given, and the contribution of this essay presented.

2 Definitions, current research and theoretical background

In the following section, the concept of WFH is introduced, with a specific focus on virtual teams. Thereafter the role of the workplace for creative performance of teams is highlighted, followed by hypotheses of the study.

2.1 Work from home and virtual teams

The amount of literature covering the topic of virtual work has increased over the past decades, as a response to technological inventions enabling digital work, as well as the restriction of the COVID-19 pandemic. A variety of definitions concerning virtual work came into existence. The here presented study focuses on WFH, also known as home office (Kurland & Bailey, 1999; Lim & Teo, 2000). Several studies analyze the possible effects of WFH on performance, satisfaction, and well-being, or consider specific conditions that need to be fulfilled to yield beneficial impacts for companies and employees (e.g., Barrero et al., 2021; Bloom et al., 2015; Lewandowski et al., 2022; Mas & Pallais, 2017; Vij et al., 2023).

Virtual teams are of special interest in the field of virtual work. Such teams are defined as assembly of individuals, often dispersed across various geographical locations and distinct time zones, who unite to collaborate on a shared project (Garro-Abarca et al., 2021; Horwitz et al., 2006; Peñarroja et al., 2013). A critical component for the success of virtual teams is existing trust between team members, as trust increases quality of communication and interaction, especially in a virtual setting (Chang et al., 2011; De Guinea et al., 2012). But it can be challenging to build trust, as virtual communication tools, e.g., chats and video conferences, may feel less human (Kennedy et al., 2010), thus decreasing social presence and exchange through poor communication channels available in contrast to FtF interaction (Daft & Lengel, 1986; Short et al., 1976).

Another challenge regarding effective team collaboration can be seen in the initial stage of a team, as it is reported that virtual teams often face problems, especially in communication, at the beginning of their virtual work, which later on diminish (e.g., De Guinea et al., 2012; Lurey & Raisinghani, 2001). However, the literature mainly focuses on existing virtual teams or cross-sectional in contrast to longitudinal observations of virtual teams. This gap in current research is addressed in the present study.

2.2 Creative teams and the workplace

The emerging field of creativity research covers a variety of topics, ranging from cultivating and identifying creativity in early education to the facilitation of creativity-enhancing leadership behaviors and workplaces. Creativity, as a driver of innovation, plays a pivotal role in gaining competitive advantage in various domains of the business world (e.g., Ko & Butler, 2007; West & Sacramento, 2023).

The concept of creativity itself is often understood as a process that combines elements that had not been combined before (Feinstein, 2011). The creative process includes the recognition of a creative problem or the need for innovation, generation of ideas and potential solutions, elaboration on those and finally the execution and implementation (Guilford, 1950; Lubart, 2001; Mumford, 2003; Treffinger et al., 2006). Distinguishing between creativity and innovation can be somewhat difficult, as these concepts are frequently conflated. One approach to differentiate these two constructs involves defining innovation as the result of harnessing creativity, coupled with the successful implementation of those creative ideas (Ford, 1996). In this study we do not focus on implementation, but on the idea generation as part of the creative process as an important precondition of generating successful innovations.[7] We define creativity, in line with the definition of Amabile, as the production of novel and useful ideas by an individual or a small group of individuals working together (Amabile, 1988). In the following, creative performance refers to this creativity definition, focusing specifically on the idea-generation stage of the creative process.

One important differentiation regarding creativity and creative idea generation is the distinction between convergent and divergent thinking. Convergent thinking describes the process of find-

[7] Before the process of idea generation starts, the problem solver in charge must first construct the creative problem by detecting and selecting the problem representation, thus, providing structure to ambiguous problems requiring creative solutions, whereas it functions as predictor of divergent thinking (Reiter-Palmon, 2018). This is not considered in our research, as the introduced experimental task anticipates the necessity of this step, so that the process of problem construction is not measurable or observable in this study.

ing the single best solution to a problem and elaborating on this solution as effectively as possible.[8] Divergent thinking, in contrast, focuses on generating as many new ideas as possible (Guilford, 1950) and is in the focus of the presented study.[9]

Existing research focuses, among other things, on potential influential factors that could affect creativity of employees. One such factor of interest in the world of economic research is the potential effect of incentives, tournaments, and rewards (Bradler et al., 2019; Charness & Grieco, 2019; Gross, 2020; Englmaier et al., 2024b; Erat & Gneezy, 2016). A strand of research analyzes different influences, like for example the workplace, on the creative performance of individuals. For example, it is found that less monitoring and creative coworkers can increase the creativity of individuals (Zhou, 2003).

The research field covering creativity in teams, its antecedents and influential factors, is not as much explored as the field of individual-level creativity (Mumford, 2003; Reiter-Palmon et al., 2021). The literature on brainstorming, not always referring to creativity, shows that performance can vary significantly between individuals, nominal groups (group output as the sum of individual outputs), and interactive groups. While nominal groups often outperform real groups in idea *quantity* due to the absence of coordination losses (Isaksen & Gaulin, 2005), this metric overlooks the goal of generating high-*quality* ideas (Rietzschel et al., 2006). Under certain conditions – such as high cognitive motivation, emotional support, and structured group environments – interactive groups can surpass individuals in the originality and feasibility of generated ideas (Nijstad & Stroebe, 2006; Paulus & Brown, 2007). To address the well-documented productivity loss of brainstorming groups compared to individuals (Paulus & Dzindolet, 1993; Paulus & Kenworthy, 2022), electronic brainstorming has been proposed as a strategy to reduce process losses such as production blocking and social loafing. By allowing participants to contribute ideas simultaneously, this method has been shown to close or even reverse the performance gap between nominal and interactive groups in some cases, but mixed findings exist (e.g., Dennis & Williams, 2005; Kohn et al., 2011; Nelson & Guegan, 2022; Paulus & Dzindolet, 2008; Ziegler et al., 2000). However, these studies focus on text-based electronic collaboration tools and rarely account for modern forms of virtual communication, such as videoconferencing, which introduce additional verbal, visual, and social cues. Unlike platforms that only allow simultaneous idea entry and screen sharing, videoconferencing introduces verbal, visual,

[8] See Harvey and Berry (2023) for a study about usefulness and elaboration in the creative process.
[9] The notion of divergent thinking as an indicator of creativity and the accordingly developed tests have prevailed in (experimental) research (Runco & Acar, 2012) as divergent thinking, i.e., idea generation, is crucial for the subsequent process of creativity (Amabile & Pratt, 2016).

and social cues that influence creative interaction over time. Moreover, most existing studies abstain from examining the role of repeated interaction within stable groups, which is a core dimension of many real-world teams.

One focus of current research lays on the composition of a creative team and on the role of diversity (e.g., Dutcher & Rodet, 2022; Hoever et al., 2012; Kurtzberg, 2005). Whereas the underlying hypothesis is that diverse team members can complement each other during the creative process, following the additive model of creative thinking (Feinstein, 2011; Kurtzberg & Amabile, 2001). In several studies, it is found that not every aspect of diversity (e.g., socio-demographic, experiential or characteristic diversity) has a significant or even beneficial impact on the creative performance of teams (e.g., Abra, 1994; Dutcher & Rodet, 2022; Thatcher & Brown, 2010).

Another strand of literature studies the relation between working in presence or WFH and creativity (e.g., Dutcher, 2012; Hunter, 2019). When research focuses on the effect of a virtual workplace, this is often done by focusing on virtual teams, as teams automatically come with a need for communication, which is the major difference between working in presence and WFH. Results on the role of CMC, like chats or videoconferences, and FtF communication are mixed regarding their influence on creative team performance (e.g., Brucks & Levav, 2022; De Guinea et al., 2012; Dutcher & Rodet, 2023; Grözinger et al., 2020; Kerr & Murthy, 2004). Especially with focus on CMC tools, mixed results whether videoconferencing tools inhibit creative performance exist (e.g., Brucks & Levav, 2022; Grözinger et al., 2020). In an extended literature review regarding virtual creative teams, Reiter-Palmon et al. (2021) highlight the role of required communication as the crucial difference between individuals and teams working in the creative field, which is significantly affected by the virtual environment through media richness and building trust. Here, the formation phase of a virtual team is considered to be crucial for building functioning social processes in such teams, whereas those are understudied in the field of creativity research. The here presented study contributes to the less researched field of understanding creative performance of virtual teams in comparison to individuals.

It has to be noted that existing studies were conducted before the increased use of CMC in work and private life, as is the case nowadays. Therefore, changes in the handling of CMC in teamwork as well as general performance in a WFH setting can be expected. Additionally, research focuses for the most part on one-shot interactions in teams, thus limiting our knowledge of the role of CMC on creative teams during repeated interaction and particularly in the formation phase of a team (Reiter-Palmon et al., 2021). Moreover, research on the general comparison

between creative individuals and teams is limited.[10] Our study contributes to this limited research on the direct comparison of individuals and teams working on a creative task in a repeated WFH setting.

2.3 Hypotheses

Dutcher's (2012) study highlights that individuals demonstrated increased productivity on creative tasks when working remotely and not in the lab. This finding can lead organizations to consider virtual work, particularly when a creative task is involved. However, the impact of virtual work on team-level creativity is less straightforward. Brucks and Levav (2022) discover a possible downside to virtual work within teams, as they find a negative effect of screen focus during virtual team collaboration, leading to less stimuli, on a creative task. The assumption that this holds for all virtual teams has a significant weakness, as the experiment was still conducted in the lab and not in a WFH environment. Therefore, it is questionable whether this result will also hold in a WFH situation, where no direct monitoring and more natural surrounding exist, such that there exist similar, if not more, stimuli accessible for members of such located teams. If the more stimulating and natural work environment of WFH situations does, in fact, foster creativity within teams, the possibility of extending individual-level benefits of virtual work towards teams emerges. However, for this transfer to be effective, it is important to consider not only the workplace itself but also the collaboration-specific dynamics unique to teams.

Another aspect to be considered is whether it is at all beneficial to let two or more individuals work on a creative task, as opposed to having them work separately. It is obvious that the creative output in a team may not simply be the sum of their members' performances. There are concerns that such teams might experience efficiency losses due to reduced concentration, communication quality, coordination difficulties and goal orientation in a virtual work environment (e.g., Girotra et al., 2010; Janis, 1972; Mullen et al., 1991; Osborn, 1953). In contrast, having one highly creative team member in a dyadic team can yield similar benefits to two highly creative individuals (Kurtzberg & Amabile, 2001).[11] Furthermore, Kurtzberg and Amabile (2001) suggest that the potential drawbacks highlighted in the brainstorming literature are more

[10] To the best of our knowledge, only one study regards this comparison with focus on the creation of creative capital in teams and individuals (Dutcher & Rodet, 2023).

[11] Similar, Kurtzberg found in a class experiment that one highly creative individual in a dyadic team could lead this team to reach the same output as two creative individuals working independently (Kurtzberg, 1998). In another study examining R&D teams, no definitive answer regarding the relation of team member creativity and team creativity are found. Team creativity is either the simple average of its members creativity or the weighted average, where the creative potential of the most creative team member carries more weight (Pirola-Merlo & Mann, 2004).

of a matter in large teams. This approach aligns with the studies mentioned earlier, emphasizing the beneficial influence of highly creative individuals on a creative team's outcome in comparison to independently working individuals.

Additionally, the positive effect is supported by the combinatory model and the additive model of creative thinking, which indicates that two team members complement each other during the creative process as they can leverage their unique knowledge and perspectives leading to a creative output which can never be reached by a single individual, due to the missing exchange (Feinstein, 2011; Kurtzberg, 1998; Kurtzberg & Amabile, 2001).

The synthesis of initial research findings and theoretical models suggests that the potential negative effects on creative performance during idea generation are more likely to manifest in larger teams or when less rich communication media, like chats, are employed. Creative performance, regarding quantity and quality, in teams under conditions of rich communication, like video conferences, should be higher than those of individuals, as teams are able to build their creative process on supplementary exchange of thoughts and ideas. This leads to the following hypotheses:

H1: Teams show higher quantitative creative performance than individuals.

H2: Teams show higher qualitative creative performance than individuals.

One explanation for the increased creative performance in teams, compared to individuals, is direct communication, allowing individuals to profit from the experiences and backgrounds of their team partner (Dutcher & Rodet, 2022). One important component of high-quality communication is trust between the communicating partners, whereas trust and communication influence each other and increase the flow of knowledge (Gerrard & Lockett, 2018; McEvily et al., 2003).[12]

3 Experimental design

To answer our research question, an online experiment using the lab of RWTH Aachen University has been conducted. The experiment has been implemented via the software oTree (D. Chen et al., 2016). Our experimental design consists of two phases which take place one week apart from each other to account for repeated work and compare the performance of individuals

[12] West (1990) identifies four factors of team climate to be relevant for innovation: vision, participative safety, task orientation, and support for innovation. Participative safety, which includes trust, contributes to creating an environment where team members feel comfortable sharing ideas and taking risks, which are essential for creative work. On one hand, communication increases trust (Boies et al., 2015; Hakanen & Soudunsaari, 2012), while on the other hand, trust is able to increase the (performance influencing) quality of such interactions, which is particularly pertinent in the context of creative teams (Barczak et al., 2010; Y. Chen et al., 2021).

and newly formed teams over time (Reiter-Palmon et al., 2019). As we are interested in the creative performance in a WFH setting, the experiment is conducted via the video communication platform *Zoom* and the participants have to confirm at the beginning of the experiment that they actually participated from their homes. In existing research, it is argued that CMC inhibits creative idea generation, but this was often still measured in the lab and not in a real home office setting (e.g., Brucks & Levav, 2022; Dutcher & Rodet, 2021), so we fill this research gap by implementing such a situation.

We implement two treatments, which differ in the existence of a co-worker. We compare individuals working on the task with dyadic teams of two participants who are matched randomly and have to work together in the first and second phase.[13] In the control treatment the tasks have to be solved individually in both phases.

To measure creativity via divergent thinking, we implement the so-called UUT which is frequently used in experimental research, mostly as an individual task (Guilford, 1967; Plucker et al., 2011; Torrance, 1966, 1974).[14] As its usage is not that frequent as a measure for creativity on a team-level, using it in our design enables us to not only contribute to the field of creative teams, but also to the field of divergent thinking tasks and their possible applications in experimental research. We have selected the UUT for several reasons. First, it holds strong practical relevance, particularly in collaborative settings, as it simulates key aspects of team-based ideation processes such as brainstorming – an activity central to creative performance in organizations. This makes it a suitable task for examining creativity in teams. Second, the UUT is well-established not only in psychological literature but also increasingly recognized in economic research (e.g., Bradler et al., 2019; Dutcher & Rodet, 2022; Rodet, 2021), which underscores its interdisciplinary importance and provides a foundation for building upon existing methodological standards. Third, it offers a clear and structured measurement framework, allowing for the separate assessment of creativity dimensions such as fluency, flexibility, and originality.

[13] We are aware of the fact that a team is sometimes restricted to a group of at least 3 persons. We use a broader definition which includes also dyadic teams (see e.g., Memmert et al, 2015).

[14] We have decided to use the UUT as experimental task as it brings numerous advantages when researching creativity in a virtual team context and comparing this to individual performance. First, the UUT is a very frequently used task on an individual-level and scores high on reliability and validity (Carson et al., 2005; Silvia, 2011). Moreover, the UUT is, compared to other tasks of the TTCT-Figural for example, relatively easy to implement in an online setting where different people solve the task together. Additionally, an objective evaluation method (see the section below) of the results exists, such that our results do not underly the danger of subjective evaluations. For a discussion of the implementation, usefulness and significance of the UUT see: Dutcher & Rodet (2021, 2023); Reiter-Palmon et al. (2019); Silvia (2011); and Weisberg (2006).

Furthermore, its standardized format and widespread use across studies enhance the comparability of results – both within and across individual and team contexts, and in varying experimental modalities.

In the UUT, individuals have to generate as many creative and useful ideas as possible for everyday objects. A creative use is a use for which the object was not designed for. The task is often used to assess quantity and quality of creative ideas and functions as an indicator for general creative performance (Runco & Acar, 2012), as it closely mimics work which does not fall into the categories of mundane or dull tasks (Dutcher, 2012) and resembles work which is expected in creative and innovation-oriented careers (Woodman et al., 1993). Additionally, the UUT is seen as a task reflecting the starting point for a creative process, as it focuses on the generation of diverse ideas and solutions which then need more refinement via convergent thinking and idea elaboration later on in the creative process (Reiter-Palmon et al., 2019).[15]

In each phase, participants have to work on 4 UUTs, plus one as general individual ability check in the first phase. They have 8 minutes for each UUT and are automatically forwarded to the next task or questionnaire. We use the following items in the experiment: brick, hanger, sheet of paper, cord, umbrella, knife, tablecloth, cap of a plastic bottle, as well as a tin can as ability check. Additionally to the items, the participants are given a short instruction what the intentional use of those items are (e.g., tin can: container for food) and that this use is not counted as valid answer as it is not creative.[16] The experimental screen, which is the same for all UUTs expect for the specific item and the it's intended use, can be found in Figure AI. 1 in Appendix I. When working in a dyadic team, they have to come to an agreement who is typing the answers. The other participant writes the team number in the textbox.

Figure 2 shows the experimental design, whereby written text in red only applies to the team treatment. Before the experiment starts, the instructions are presented on the screen and were read out loud by the experimenter (see AI. 12AI. 12 in Appendix I). In those instructions we explain the task, the evaluation and payment scheme and highlight that the participants should be creative when solving the UUTs (Harrington, 1975; Nusbaum et al., 2014). The participants are informed that they will only get paid if they finish both phases. By introducing a piece-rate scheme as an incentive, we keep the evaluation pressure constant over time, which was found to be relevant to maintain engagement (Prokofieva et al., 2024).

[15] See Part I section 1.2 for a more in-depth discussion of the UUT.
[16] In a pre-test we have checked for the difficulty of the used items and divided them such that the items are equally distributed in both phases, concerning their difficulty.

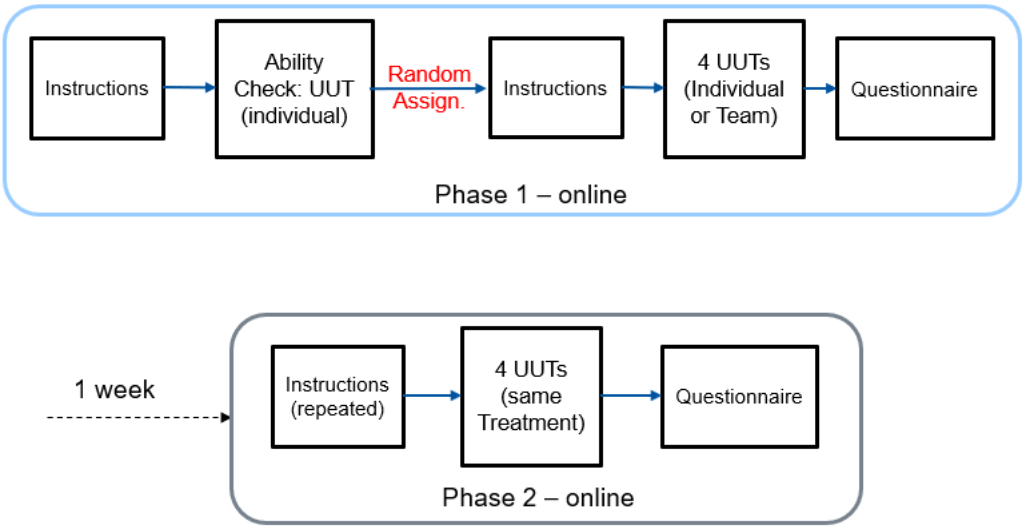


Figure 2: Experimental procedure

First, the participants have to work individually on a single UUT during the first phase to check their general creative ability and their comprehension of the task, which also enables them to get to know the task beforehand. The results of this task are also payment relevant. Afterwards, the teams in the team treatment are formed and additional instructions are presented before starting with the next UUTs. Those teams do not change in the second phase. At the end, participants have to fill in a questionnaire in both phases (AI. 13 in Appendix I). In the first phase demographic data is gathered and the participants of the team treatment get additional questions regarding sympathy and trust between themselves and the team partner. Sympathy is measured using a single item. Trust is measured with a question whether one trusts the team member as well with three items from the General Social Survey (GSS). Additionally, after phase 2, questions on personnel characteristics (BFI-2-XS, Soto & John, 2017) as well as further controls are implemented.

To evaluate the answers of the participants and their performance, we use an objective evaluation method from Bradler, Neckermann and Warnke (2019). While we refer to our measurement approach as "objective," this should be understood in a relative sense – namely, in comparison to fully subjective self-assessments or peer ratings that are also common in creativity research. Our method provides quantifiable indicators of creativity dimensions (e.g., fluency, flexibility, originality) by counting and categorizing responses, which helps standardize evaluation across participants.[17] However, we fully acknowledge that certain aspects – such as determining the novelty or distinctiveness of responses – still require human judgment, which is in line with the concept of creative ideas, which contain uncommon and unexpected ideas with varying value

[17] See AI. 14 in Appendix I for the list of categories used.

depending on the individual judging the ideas. Participants receive one performance point per valid answer (i.e. quantity), as well as additional points for sub aspects of quality: two points for rare answers, four points for very rare answers, as well as one point per covered category. Quality is a main indicator of creativity. Table 2 (Part I section 1.4) summarizes the different measures of creative performance used.

The participants receive 8 Cents per performance point as well as an additional fixed amount of €4 if they also participate in the second phase. Responses that are far away from implementation and do not have any reasonable use, as well as nonsense and redundant answers are excluded, in the following referred to as errors, following the recommendation from Reiter-Palmon et al. (2019).[18] The evaluations have been conducted manually by several student assistants.[19]

While earlier research often relies on subjective ratings to assess creativity, more recent studies increasingly emphasize the importance of multidimensional and more objective creativity metrics (e.g., Long & Wang, 2022). Our decision to apply objective measures is based on both theoretical and methodological considerations. A key criticism of traditional uniqueness scoring, as outlined by Silvia et al. (2008), is its tendency to confound fluency with originality, the ambiguity of rarity judgments and a systematic disadvantage of large sample regarding the likelihood of receiving only few rare answers. Our method mitigates the first two concerns by relying on a fixed originality threshold and clearly defined coding schemes, while the third issue is not applicable due to our comparatively small sample size. Moreover, subjective creativity assessments are known to suffer from limitations in both reliability and validity, even when implemented with trained raters (Long & Wang, 2022). Although such methods remain valuable in certain contexts, they pose particular challenges in studies with more complex group dynamics and longer task durations – as in our case. Given that our research investigates not only individual but also interactive group-level processes, it is crucial to disentangle the different dimensions of creativity to better capture these dynamics. Additionally, recent findings indicate that brainstorming and creativity are now often conflated by participants (Di Mascio et al., 2016), potentially reducing the effectiveness of standard brainstorming rules instead of explicitly motivating quantity and quality of answers. Some scholars have proposed emphasizing only quantity in task instructions (e.g., Paulus et al., 2011), yet this approach is better suited for

[18] Errors are not part of the generated performance points, because they are not subtracted from the performance, as they are counted separately and thus are assigned 0 points to.
[19] Objective measures sometimes risk to underly subjective decisions when defining the rating tables and counting words in the UUTs (Reiter-Palmon et al., 2019). Therefore, all decisions has been documented and double-checked, and a guideline has been developed beforehand.

short, individual-based creativity tasks. In group settings, an exclusive focus on quantity is shown to exacerbate production blocking (Diehl & Stroebe, 1991). Therefore, we opt to emphasize both quantity and quality. Our scoring system reinforces this dual focus by allocating points for qualitative dimensions, thereby shifting emphasis toward idea quality. Finally, subjective judgments of creativity require clear criteria to ensure consistency between raters and the understanding of what is defined as creative, especially when participants are involved in peer assessment or self-evaluation (Birney et al., 2016). We address this by explicitly defining the criteria for originality and flexibility in the participant instructions and for the raters. Although some scholars argue that originality scores are redundant due to high correlations with quantity, i.e., fluency (Hocevar, 1979; Silvia et al., 2008), recent work by Forthmann et al. (2020) cast this claim into doubt. Acar et al. (2023) further show that by calculating originality as a ratio of unique to total ideas, our measure reduces this dependency and offers a more nuanced view of creative performance.

4 Results

In the following chapters the results are presented, starting with descriptive results, and followed by hypotheses testing. Later on, potential mechanisms and further results are presented.

4.1 Descriptive results

The data has been collected between February and March 2023. The final sample comprises 106 participants who have engaged in both phases, resulting in a participation rate of 45% post dropout.[20] The team treatment consists of 48 individuals, forming 24 dyadic teams with 25 males and 23 females. The individual treatment includes 58 participants, 29 males, 28 females and one participant chose not to disclose their gender.

In general, the sample is characterized by a high level of education. The average age of the whole sample is 26 years, ranging from a minimum of 20 years to a maximum of 41 years. Demographic characteristics for the 37 individuals, 19 males and 18 females, who dropped out are limited, as personnel data were collected at the end of the first phase such that individuals who dropped out before are not covered.

[20] In total, 235 participants registered for the experiment, but only 141 showed up. Additionally, 37 dropped out, either because they or their team partner did not attend the second phase. Considering the potential of self-selection, a comparison was made between the results pertaining to performance, socio-demographic characteristics, sympathy, and trust in both treatments for participants who attended both phases and those who dropped out. In both treatments, no significant differences were found. Thus, no evidence for an underlying self-selection effect exists.

The first phase lasted approximately 62 minutes, while the second phase lasted around 47 minutes. For the team treatment, the average payment is €28.38 in phase one and €27.18 in phase two. For the individual treatment, participants receive an average payment of €15 in phase one and €13.92 in phase two.[21] The average ability of the participants in the UUT, measured with a single UUT at the beginning of the first phase, is 31.5 points, ranging from a minimum of 0 points to a maximum of 89 points, whereas on average 13.6 unusual uses are generated. The ability does not differ decisively between the two treatments.[22]

For a comprehensive overview of the various performance measures in both phases per treatment, refer to Table 3. On a relational level, the quantitative performance of dyadic teams is between 1.4 to 2.2 times higher than the performance of individuals and between 1.1 and 1.7 times higher for qualitative performance, depending on the specific UUT (see Figure AI. 2 in Appendix I).

Table 3: Mean performance and SD for different measures of creativity of individuals and teams over both phases

Measure	Phases	Treatments	Mean	SD
Quantitative	1	Individual	56.34	23.62
		Team	87.58	23.17
	2	Individual	58.40	27.69
		Team	91.29	27.52
Qualitative	1	Individual	61.29	30.63
		Team	92.08	28.57
	2	Individual	65.53	34.01
		Team	98.04	37.23
Errors	1	Individual	12.93	16.42
		Team	8.71	4.06
	2	Individual	11.55	9.85
		Team	9.71	2.91

Regarding the control variables of satisfaction, joy, and effort, participants of the team treatment report a higher level of enjoyment during the experiment (mean of 3.98 vs. 3.5, one-sided p= 0,054), while finding the experimental tasks significantly less demanding (mean of 2.23 vs 2.84, one-sided p=0.009) on average. Although team members are slightly more satisfied on average

[21] There are large differences in payment in both treatments, whereby the minimum payout of the team treatment after both phases amounts to €26.48 and the maximum payout reaches €81.52. In the individual treatment, the minimal payout is €8.48, while the maximum is €55.52. Noteworthy, the payments do not show that two individuals perform better than a team, because of redundancies in the answers given by two independent individuals Also see section 4.3 for further results on this notion.

[22] This is important to note, as therefore the following significant differences between treatments are not due to different distribution of ability over the two treatments.

(mean of 4.56 vs. 4.46) with their working environment, e.g., working from home, this difference was not statistically significant.

4.2 Hypothesis testing

The first and second hypothesis posit that the creative performance of dyadic teams should be significantly higher than the performance of individuals working alone on the same tasks. Performance is assessed based on the quantity and quality of the unusual uses created, whereas for the first phase, the ability-checking task does not count into the measure of performance, as it is conducted at an individual level, independent of treatment. Figure AI. 3 in Appendix I displays the average qualitative performance of both treatments over the two phases, with dotted lines indicating the average qualitative ability of the participants. The lines for both treatments follow a similar pattern, indicating that the different objectives chosen for the UUT have varying difficulties. These differences in difficulty are perceived similar in both treatments.

Figure 3 below shows the average performance levels regarding quantity, quality and errors made by the treatments, together with the test results.[23] The team treatment differs significantly from the individual treatment, producing more unusual uses ($p<0.001$, $r=0.51$). This also hold true for the qualitative performance of the first phase ($p<0.001$, $r=0.45$). Regarding the second phase, the pattern persists for both quantitative and qualitative performance ($p<0.001$, $r=0.44$; $p<0.001$, $r=0.42$). Concerning the inverse performance measure of errors made during the experiment, the treatments do not differ significantly from each other during the two phases or the aggregated measures ($p=0.567$; $p=0.927$; $p=0.515$).

We further test for the robustness of these results, when controlling for the measured ability of participants. Table 4 depicts the treatment effects on quantitative and qualitative creative performance, as well as errors made, controlling for the performance in the ability task, conducting OLS regression analysis.[24] A highly significant effect of the team condition on the quantitative and qualitative performance of both phases is observed, controlling for creative ability. In contrast, there are no treatment differences regarding the numbers of errors made (models 7-9). With the results now presented, we can accept the first and second hypotheses.

[23] The significance of the difference of the creative performance between the two treatments was evaluated using Mann-Whitney U-Tests.

[24] The analysis controls for the aggregated ability of two participants forming a dyadic team, as the general ability to perform a UUT can influence the creative output at team level. The inclusion of other control variables, such as gender composition, the Big5 or satisfaction with the experimental place, reduced the adjusted R^2. No control variable exhibited a consistent significant effect on the performance measures for the two phases.

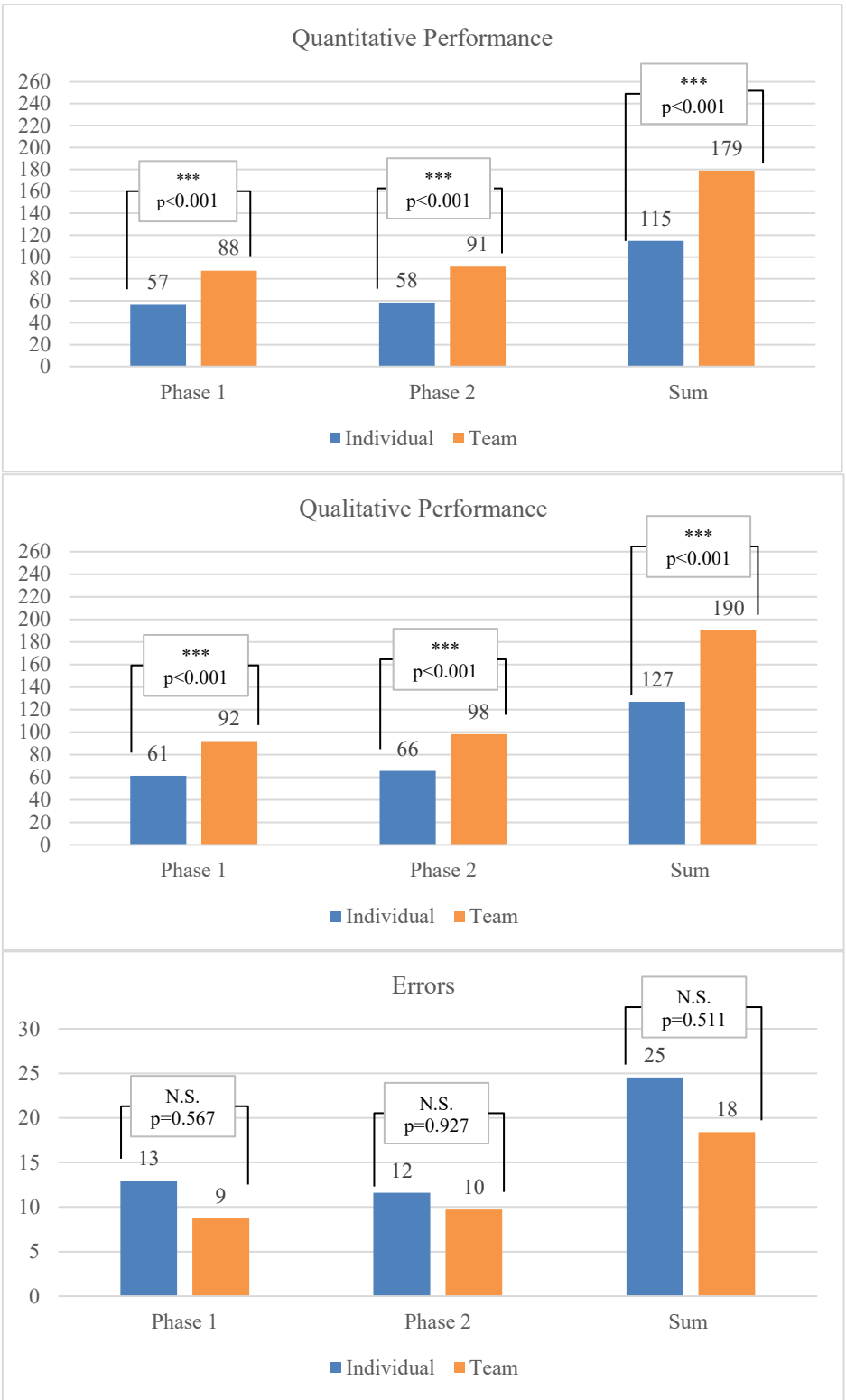


Figure 3: Average performance results of the two treatments with test results of Mann-Whitney U-Tests

Table 4: OLS regression on treatment effects for different performance measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	Quan. Phase 1	Quan. Phase 2	Quan. Sum	Qual. Phase 1	Qual. Phase 2	Qual Sum	Errors Phase 1	Errors Phase 2	Errors Sum
Team	30.81***	32.50***	63.31***	28.54***	30.12***	58.67***	-1.03	-0.98	-2.07
	(4.870)	(6.124)	(10.044)	(6.944)	(8.289)	(13.000)	(1.513)	(1.361)	(2.011)
Quan. Ability	2.606***	2.366***	4.972***						
	(0.357)	(0.457)	(0.742)						
Qual. Ability				1.588***	1.683***	3.27***			
				(0.299)	(0.326)	(0.526)			
Error Ability							3.12***	0.849***	3.97***
							(0.510)	(0.153)	(0.511)
Constant	20.40***	25.76***	46.16***	32.88***	35.42***	68.29***	5.45***	9.52***	14.97***
	(5.625)	(6.769)	(11.426)	(4.839)	(5.671)	(9.104)	(1.366)	(1.228)	(2.044)
Obs.	82	82	82	82	82	82	82	82	82
Adj. R-squared	0.596	0.427	0.557	0.444	0.376	0.482	0.687	0.127	0.588

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

4.3 Potential mechanisms exploring the difference between individuals and teams

As it becomes evident that dyadic teams show significantly higher creative performance in the UUT within an online setting of newly formed teams, the subsequent question arises, which factors influence their creative performance. In the following, further results and checks concerning influential factors such as sympathy and trust, team composition, the ratio of quantity and quality, as well as a comparison with simulated teams are presented.[25]

Correlations

Spearman's rank correlations are calculated to check for correlations between preferences, socio-demographic factors and team performance (Table AI. 8 in Appendix I). Only six, partly weak, significant correlations show. The preference to work in an online workplace setting correlates weakly significantly negative with quantitative performance of phase one, as well as with both measures of quality. Additionally, strong positive correlations are found between the preference to work as a team on errors made in the first phase and of age on the errors made in the second phase. We conduct several regression analyses checking for potential direct or interaction effects of those control variables. No such effects are found.

Interestingly, creative ability correlates significantly and positively with the trait of being open towards new experiences ($r=0.268$, $p=0.008$), aligning with literature suggesting that open individuals tend to be more creative (Carson et al., 2005). In turn, openness correlates significantly and positively with perceived joy during the experiment ($r=0.207$, $p<0.028$). The effort needed to fulfil the experimental tasks correlates significantly and positively with the trait of neuroticism ($r=0.218$, $p=0.027$), being employed ($r=0.321$, $p=0.002$) and having a workplace conducive to remote work ($r=0.263$, $p=0.011$), while significantly negatively correlating with joy ($r=-0.316$, $p=0.001$) and satisfaction with the experimental workplace ($r=-0.262$, $r=0.010$). Notably, the preference for teamwork correlates significantly and negatively with age ($r=-0.275$, $p=0.006$) and preference for forms of remote work ($r=-0.188$, $p=0.069$).

Sympathy & Trust

Furthermore, we investigate the role of developed sympathy and trust over the two experimental phases on creative performance of teams. Trust and building of trust are of special importance

[25] As the research interest in the diversity of teams increased in the recent years (e.g., Dutcher & Rodet, 2022; Hoever et al., 2012), we conducted an analysis of the influence of different aspects of team diversity: gender, age and creative ability. No significant results for any performance measure in the two phases were found.

for virtual teams as the development of trust can be inhibited by the missing social presence of team members (Chang et al., 2011; Daft & Lengel, 1986; Garro-Abarca et al., 2021; Peñarroja et al., 2013). This can especially be problematic for creative virtual teams, which heavily rely on team identity and knowledge flow, which is influenced by the level of trust in those teams (Gerrard & Lockett, 2018; Jaravel et al., 2018; Naotunna & Zhou, 2022; Nemiro, 2002). Trust and sympathy, i.e., liking each other, as related construct (Lu, 2015), can therefore function as performance enhancing components in virtual teams. Table 5 shows the average level of sympathy and trust over the two experimental phases in dyadic teams. A slight increase over the two phases in both measures is visible.

Table 5: Average levels of sympathy and trust in dyadic teams over the two phases with SD in brackets

	N	Sympathy		Trust	
		Phase 1	Phase 2	Phase 1	Phase 2
Teams	24	4.44	4.56	3.54	3.69
		(0.86)	(0.40)	(0.75)	(0.66)

Table AI. 9 in Appendix I shows Pearson correlations of performance results of dyadic teams, now incorporating consideration of sympathy and trust. The perceived sympathy for the other team member after the first phase correlates significantly and positively with quantitative performance of this phase (r=0.423, p=0.040), as well as of the second phase (r=0.433, p=0.034). Perceived trust after the first phase also showed a significant positive correlation with the quantitative performance of the second phase (r=0.345, p=0.099). Regarding the qualitative performance and the errors made, only weak correlations exist. Conversely, the measures of sympathy and trust after the second phase show no significant correlations with all three performance measures but exhibit a strong and significant correlations with each other. Given the identified correlations, we can hint towards a beneficial connection between sympathy and the creative performance of the first phase. Regarding the relation between performance and trust, only a positive relation between trust after the first and performance in the second phase was found. Thus, an indication for the importance of time in the context of trust development and creative performance exists.

We conduct a regression analysis (Table 6) focusing on the role of sympathy and trust in creative virtual teams.[26] Note however, that the number of observations is rather small, such that the results can only be interpreted as first indications. During the first phase, a significant positive

[26] No regression for the sum of the performance measures was made, as sympathy and trust were measured twice, at the end of each phase.

but weak influence of sympathy on the qualitative performance is found (model (3): β=17.93, p=0.057), whereas trust does not significantly influence the creative performance of the first phase. This scenario shifts for the second phase (model (4)). Sympathy shows significantly negative impact on quantity (model (2): β=-29.72, p=0.015) and quality of the second phase (model (4): β=-41.67, p=0.048). Consequently, increased sympathy during the second phase tends to lower the quantity and quality of answers given, potentially hinting towards losses in concentration if high levels of sympathy between team members exist.

Table 6: OLS regression on the effect of Sympathy and Trust on different performance measures

VARIABLES	(1) Quantitative Performance Phase 1	(2) Quantitative Performance Phase 2	(3) Qualitative Performance Phase 1	(4) Qualitative Performance Phase 2	(5) Errors Phase 1	(6) Errors Phase 2
Sympathy	11.61	-29.72**	17.93*	-41.67**	-2.800	-2.263
	(8.397)	(11.11)	(8.893)	(19.79)	(1.721)	(1.363)
Trust	0.252	23.55**	-9.041	27.91**	2.773	3.261***
	(10.26)	(9.963)	(12.43)	(12.21)	(2.152)	(0.987)
Quantitative Ability	1.277*	0.837				
	(0.643)	(0.741)				
Qualitative Ability			0.481	-0.184		
			(0.667)	(0.914)		
Errors in Ability					0.226	-0.691**
					(0.708)	(0.328)
Constant	17.29	128.3**	36.62	188.6**	11.00***	8.959*
	(23.43)	(48.10)	(23.84)	(83.64)	(3.460)	(5.164)
Observations	24	24	24	24	24	24
Adj. R-squared	0.183	0.115	0.039	0.102	0.029	0.236

Sympathy and Trust were measured at the end of each phase. The measures corresponding to the phases are used.
Robust standard errors in parentheses
*** p<0. 01, ** p<0.05, * p<0.1

Trust, in turn, shows a significant positive influence on all three measures of performance in the second phase (models (2), (4), and (6), respectively). On the one hand, trust positively influences the quantity (β=23.55, p=0.028) and quality (β=27.91, p=0.033) of creative answers. On the other hand, trust significantly increases number of errors made (β=3.261, p=0.004) too.

Performance of low- and high-performing teams

Considering the possibility that the enhanced performance of dyadic teams may be attributed to specific highly performing teams rather than an overall improvement when working collaboratively on a creative task, we categorize teams into low- and high-performing groups, using a median-split, and comparing those to individuals performing the tasks. When focusing on absolute performance, the 16 bottom places regarding qualitative performance, all come from individuals and not from dyadic teams. The corresponding graphs for the quantitative and qualitative performance of these three groups can be found Figure AI. 4 and AI. 5 in Appendix I.

It is observed that low-performing teams and individuals exhibited nearly identical performance at the beginning of the experiment, without the ability-checking task. However, starting from the second task in the first phase, the low-performing teams constantly outpace the individuals on average, a trend that persists even at the outset of the second phase. Using factorial ANOVA, the three groups differ significantly from each other regarding average quantitative ($F(2,81)=29.12$, $p<0.001$; $F(2,81)=17.54$, $p<0.001$) and qualitative performance ($F(2,81)=20.83$, $p<0.001$; $F(2,81)=9.07$, $p<0.001$). Fisher-Hayter post-hoc tests reveal that the difference between the low-performing teams and the individuals stays significant on a 5%-level for quantitative and qualitative performance.

Diversity in teams

As research interested in diversity of teams increased in recent years (e.g., Dutcher & Rodet, 2022; Hoever, et al., 2012), we conduct an analysis of the influence of four different aspects of team diversity: gender, age, individual creative ability and personnel characteristics. When examining differences in gender composition and the age gap between two members of a dyadic team, no significant results for any performance measure in the two phases were found.

Regarding the degree of heterogeneity in creative ability, team heterogeneity was calculated as the difference in ability between two team members, with teams being divided into heterogenous and homogeneous teams via a median split. Theory and research anticipate an increase in creative team performance with a rise in the team's heterogeneity level, suggesting that heterogeneity in context of different backgrounds and abilities, i.e., one member is relatively more able, should have a positive impact on the creative performance of teams, potentially driven by the high-ability member (Dutcher & Rodet, 2022; Kurtzberg & Amabile, 2001). Mann-Whitney U-Tests do not prove evidence for any significant difference between homogeneous or heterogeneous teams regarding quantitative and qualitative performance as well as errors made in both phases. Thus, no evidence for a beneficial effect of one relatively high ability member being paired with a low ability member is found.

Another source of heterogeneity in a team is the characteristics of the team members, which are measured through the Big5 personality traits. To compare teams while considering the team members' characteristics, the difference between the traits of two members is calculated. Subsequently, teams are subdivided into no, low and high diverse teams. ANOVA analysis indicates a significant difference in the average total performance of the first ($F(2,21)=2.86$ $p<0.01$) and second phase ($F(2,21)=4.35$ $p<0.05$) when considering the personality trait of conscientiousness. Tukey-HSD post-hoc tests reveal a significant difference between teams that are

rated as homogeneous and those that are highly heterogeneous. Specifically, the homogeneous teams in their characteristic of conscientiousness show better performance in both phases (MD=50.292, HSD=3.081 at $p<0.1$; MD=79.236 HSD=4.03 at $p<0.05$). No significant results for the other personality traits are found.

Ratio of quality per quantity

Another potential constraint to consider when assessing and comparing the creative performance of individuals and dyadic teams is the possibility that the significantly higher creative performance of teams is solely attributed to a higher quantity rather than quality (Forthmann et al., 2020). When assessing creativity in an objective way, as done in this study, particularly the results of qualitative performance suggest a heightened level of creativity in teams. Hence, it becomes necessary to compute a quality index, considering the ratio between quality and quantity of produced ideas to exclude the possible confounding effect of quantity on quality. The qualitative index is computed as follows, with the qualitative performance comprising originality and flexibility:

$$\text{Qual}_{\text{quantity}} = \frac{Quality}{Quantity}$$

$$\text{Qual}_{\text{originality}} = \frac{rare + very\ rare\ answers}{Quantity}$$

$$\text{Qual}_{\text{flexibility}} = \frac{number\ of\ covered\ categories}{Quantity}$$

The average ratio of quality per quantity for teams is 0.14 points (SD=0.06) for the first and 0.19 points (SD=0.06) for the second phase, whereas the average of individuals is 0.13 points (SD=0.07) and 0.15 points (SD=0.07). For the adjustment of originality against quantity, Mann-Whitney U-Tests reveal that teams exhibit a significant difference in comparison to individuals, with teams demonstrating higher relative performance in the first ($z=1.447$, one-sided $p=0.074$) and second phase ($z=2.856$, one-sided $p=0.004$). Covering the component of flexibility and its relation to quantity, significant differences in performance of the first and second phase are found. However, in this case, the individuals show higher performance ($z=4.145$, $p<0.001$; $z=3.117$, $p=0.002$).[27]

[27] The tests beforehand are all calculated using averages. In contrast, measures of the quality index are computed for each team and individual and averages are calculated afterwards. Thus, the results could be driven by outliers. To account for this change, we also calculate the median of the quality index and conducted the same tests with it, with no changes in results.

Teams and Simulated Teams

We do not find that dyadic teams perform twice as good as individuals in every measure of performance and UUT. This might lead to the conclusion that the implementation of such teams, consisting of two individuals and thus doubling wage payments, might not be financially justified. Consequently, the idea emerges that if the solutions generated by two individuals working separately on the same task are collected and aggregated afterwards, it will result in a higher overall number of creative ideas than those generated by dyadic teams. To test this expectation, a simulation using the existing experimental data is required. For this, the creative output of the 58 participants in the individual treatment needs to be combined pairwise, resulting in a total of 1653 simulated pairs.[28] This new simulated aggregate of creative performance of simulated pairs can then be compared to the performance of teams, showing whether dyadic teams still perform better than separately working pairs.[29]

We, therefore, conduct one example simulation in which we randomly match two participants of the individual treatment, forming a simulated team. In total, data of 29 simulated pairs was generated.[30] Table 7 shows averages and SDs of the three performance measures of the two experimental phases for teams and simulated teams.

Table 7: Average performance of Teams and Simulated Teams with SD in brackets

	N	Quantitative Performance			Qualitative Performance			Errors		
		Phase 1	Phase 2	Sum	Phase 1	Phase 2	Sum	Phase 1	Phase 2	Sum
Team	24	87.58	91.29	178.88	92.08	98.04	190.13	8.71	9.71	18.42
		(23.17)	(27.52)	(47.02)	(28.57)	(37.23)	(55.87)	(4.06)	(2.91)	(5.72)
Simulated	29	101.10	101.72	202.83	100.03	93.00	193.03	27.45	22.03	49.48
Team		(15.70)	(26.29)	(37.13)	(28.56)	(28.31)	(47.44)	(21.67)	(9.34)	(27.41)

When testing for differences between dyadic teams and our example simulation, no significant differences are found regarding quality (p=0.532 for phase 1 and p=0.714 for phase 2) and

[28] The 1653 pairs result from 58*57 possible combinations of participants, divided by two to form dyadic teams.

[29] Unfortunately, such a simulation encounters a significant challenge, as in the UUT several identical answers were provided by nearly every individual and team. For instance, for the item "tablecloth" using it as a curtain or as a bag, and the corresponding synonyms, was given over 60 times. Given the richness of the German language in synonyms, this would require manual correction which is i) prone to errors and ii) enormously time consuming, as we do not have a suitable algorithm at hand. Assuming that the manual verification of given answers of one UUT of a simulated pair would take on average 4 minutes, this would lead to a total time expenditure of 32 minutes for both experimental phases per pair. This translates to 881.6 hours needed to complete the manual correction for all simulated pairs, not accounting for the required double-check.

[30] One dataset for each simulated team was generated by merging the answers of the two individuals forming a simulated team into one dataset. During the process, the given answers of a simulated team were manually controlled and corrected for similar answers. Thereafter, the final output of the simulated team is calculated by using the known dimensions of creativity.

quantity of answers of the second phase (p=0.136). Also, the two groups differ significantly from each other in terms of quantity in the first phase, such that the simulated teams produce significantly more valid ideas (p=0.034). Moreover, the two groups differ significantly from each other with regard to errors made (see Figure 4). Simulated teams made significantly more errors over the two phases of the experiment (p<0.001), thus, being less efficient in their answers as they might produce more answers that are valid in the end but at cost of more invalid answers and without benefits for originality and flexibility of ideas. Therefore, when deciding between a team or two separately working individuals, these results hint that one has to note that the answers produced by individuals are more likely to contain errors and thus can be less efficient in terms of hit rate.

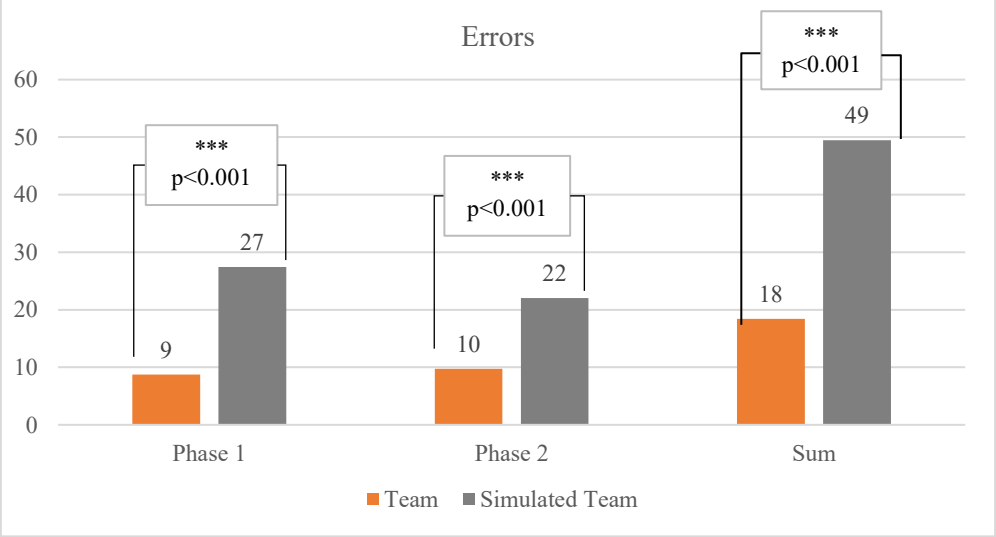


Figure 4: Average errors made of Teams and Simulated Teams with test results of Mann-Whitney U-Tests

Furthermore, we calculate the quality index for simulated teams, again considering the average as well as the median, whereas the average of the simulated team is 0.07 points (SD=0.03) for the first and 0.009 points (SD=0.04) for the second phase. The median performance of the teams are 0.14 points and 0.18 points, whereas those for the simulated team amounted to 0.07 points and 0.08 points. Highly significant differences (p<0.001) show for the average and median for both phases, such that the quality index is significantly higher for real than simulated teams. Thus, highlighting the notion made that simulated teams produce higher quantity, but the efficiency of answers in terms of flexibility and originality, being strong indicators for creativity, is less.

To check for the representativeness of the example simulation conducted, we compare the characteristics of teams and simulated teams. Table 8 shows averages, SDs, minima, and maxima

of several characteristics of the two groups as well as of the ability measures. Figure AI. 6 in Appendix I shows the gender composition of teams in the two groups.

Table 8: Characteristics of Teams and Simulated Teams

	Variable on group level	Average	Std. Dev.	Min	Max
Teams	Age	25.96	2.42	23	31.5
Simulated Teams	Age	27.01	4.15	22	38.5
Teams	Satisfaction Place	4.56	0.47	3	5
Simulated Teams	Satisfaction Place	4.47	0.42	3.5	5
Teams	Extraversion	3.54	0.42	2.84	4.33
Simulated Teams	Extraversion	3.28	0.41	2.67	4.17
Teams	Openness	3.90	0.56	2.33	5
Simulated Teams	Openness	3.68	0.39	2.84	4.50
Teams	Neuroticism	2.83	0.49	2	3.67
Simulated Teams	Neuroticism	2.83	0.61	1.67	4.84
Teams	Agreeableness	3.65	0.53	2.33	4.50
Simulated Teams	Agreeableness	3.63	0.54	2.50	4.84
Teams	Conscientiousness	3.33	0.56	2	4.17
Simulated Teams	Conscientiousness	3.52	0.59	1.84	4.50
Teams	Quantitative Ability	13.98	6.02	0	26
Simulated Teams	Quantitative Ability	13.79	4.56	3	22
Teams	Qualitative Ability	18.19	9.12	0	38.5
Simulated Teams	Qualitative Ability	17.90	8.17	5.5	34.5
Teams	Errors in Ability	1.38	1.37	0	4.5
Simulated Teams	Errors in Ability	2.40	3.05	0	15

Testing for significant differences using Mann-Whitney U-Tests and ANOVA reveal significant differences regarding the personality traits of extraversion ($p=0.040$) and openness ($p=0.062$), with both traits being more pronounced in teams. To check whether those factors influence the higher levels of efficiency for teams, we conduct regression analysis and control for those two factors, as well as included interaction effects. No significant direct nor interaction effects are found (Table AI. 10 and AI. 11 in Appendix I).

Furthermore, we find significant differences regarding the age of two members of the same team, whereas the age difference is higher in simulated teams (Figure AI. 7 in Appendix I). Checking for a possible effect of age difference on creative performance via regression analysis, no significant direct nor interaction effects on any performance measure are found.

Those results suggest that while separately working individuals may generate a greater quantity of ideas, dyadic teams may excel in terms of quality, emphasizing different aspects of creative

performance. Thus, the performance in focus of interest is important for the decision between several individuals working separately or a team working on a creative task.

5 Discussion and conclusion

In the presented study, we explore potential differences between the creative performance of individuals and dyadic teams in a repeated virtual work setting, as well as identifying factors influencing creative performance of such teams.

Several analyses reveal a significant difference between individuals and dyadic teams, with teams demonstrating significantly higher quantitative and qualitative creative performance, thus confirming the first and second research hypotheses. As the implemented objective evaluation method regarding the performance during the UUTs comes with the need to correct for the level of quantity when assessing qualitative performance (Forthmann et al., 2020), an index of quality is calculated to account for this correction. The overall quality index reveals significantly higher creative performance of dyadic teams. Doing so with regard to originality (rare answers) and flexibility (covered categories), a significantly higher creative performance of teams for originality, but not for flexibility showed. This unexpected result for flexibility can be explained by the fixed number of possible categories, thus an upper bound for flexibility, existing. This constraint does not apply to originality. This notion leads to the consequence that the result of the quality index considering the identified categories in relation to the number of ideas generated is deemed inconclusive, as this might be due to the constrained number of categories.

In line with the findings above favoring teams in comparison to individuals, testing for significant differences in creative performance between low-performing teams, identified by a median-split, and individuals reveals that even low-performing teams outperform individuals from the second task on, focusing on average creative performance.

Furthermore, teams show higher efficiency, i.e., higher quality per quantity and less errors, compared to an example simulation of the results of two individuals which were later on combined. Thus, when deciding whether a creative task should be solved by dyadic teams in direct contact or two separately working individuals, the specific performance goal, i.e., quantity, quality, or efficiency, needs to be considered. Active communication and exchange in contrast to only merging the output of two individuals, seem to enhance some components of creative performance of teams working on a specific creative task.

Additionally, significant and positive relations between trust and creative performance of teams are found. Teams can establish trust relatively quickly (e.g., Meyerson et al., 1996). The one-week interval between the first and second phases appears to be a sufficient timeframe for trust to exert a significant impact on creative performance of dyadic teams. This highlights the dynamic nature of trust development within teams and its varying influence at different stages of collaboration.

The same beneficial influence for the phenomenon of sympathy as for trust was expected. However, this only holds for the first experimental phase. It needs to be noted that the experiment has been conducted online, whereby the individuals worked in so-called break-out rooms, where only the team members themselves had access to, such that no observation through the experimenter was possible. Considering this, the possibility exists that two participants who felt a high degree of sympathy towards each other tended to lose focus on the task (Methot et al., 2016; Pillemer & Rothbard, 2018) and started chatting over other, task-unrelated, topics. Such losses of task focus may even be encouraged by the UUT itself, as it triggers divergent thinking in any possible field, leading to a conversation full of diverse topics and therefore may increase longer conversations about a specific topic which the two members find interesting, drifting away from the task.

Focusing on the effects of diversity regarding team members' characteristics, results hint towards a negative effect of heterogeneity for the trait of conscientiousness, referring to the work attitude of an individual. Thus, potential conflict between team members induced by different conceptions of work behavior might affect creative performance, of those teams negatively (Herring, 2009).

In the following, limitations and implications for future research as well as the practical field are discussed before highlighting the contribution of the presented study.

5.1 Limitations

Some limitations of this essay need to be pointed out. First of all, while divergent thinking is a crucial aspect of creativity, the study's findings should be interpreted within the context of idea generation tasks, and the generalizability to situations requiring convergent thinking may be limited. But divergent thinking, as the process of generating ideas, is foundational to creativity across all domains, serving as a starting point of innovation. As such, assessing divergent thinking provides a robust initial step in research on creative processes in dyadic teams.

Secondly, the study exclusively focusses on dyadic teams, therefore the findings may not necessarily generalize to teams of different sizes. Team dynamics can vary significantly with changes in team size. The impact of factors such as trust, sympathy, and overall team performance may differ in larger teams due to increased complexity and communication challenges. Another possible influence of increasing team size is that the benefit of one additional team member declines the larger the existing team is, as coordination costs increase (Anagnostopoulos et al., 2012). Next to increasing team size, the limited time frame of two working phases with one week apart needs to be considered. Even though this essay moves beyond one-shot observations, it does not capture any long-term effects of trust and team interaction but the early formation phase of teams.

Furthermore, the relatively small sample size of 48 participants in the team treatment, leading to 24 dyadic teams, needs to be considered when discussing the results concerning diversity, sympathy, and trust in those teams. This small sample size can result in some biases, whether they are upward or downward is not clear, but may be needed to get further researched on as diversity is of increasing relevance at the workplace.

Lastly, in a broader context, the here assessed creative performance of dyadic teams in comparison to individuals, is done in a western cultural context. This is somewhat important, as creativity is defined and perceived differently between regions and cultures and WFH comes with the advantage of forming teams with members that are regionally, and thus culturally, dispersed. For example, for a more eastern part of this world, creativity is more of a spiritual, natural, non-linear, and less outcome focused issue. In contrast to those two perceptions of creativity, the Anglo-Saxon world has a more distant, methodological, persistent, serious, outcome focused perspective on creativity, which can happen on demand (Cabra & Guerrero, 2022). This last perception of creativity may be even underlined by the way creativity is measured in the presented study, as individuals were asked to be creative, with that engaging in creativity on demand, while creativity is methodological and outcome focused measured via divergent thinking. This observation leads to the limitation that the results on creative performance should be transferred with caution to a non-western context and further research could consider nationality of team members as additional control variable, as different cultures think of creativity in different ways.

5.2 Implications for further research

The findings of this study open several promising avenues for future research, regarding extensions to experimental design, a deeper understanding of team dynamics, and exploration of alternative work environment configurations.

Regarding experimental design, a number of refinements could strengthen the validity and generalizability of the present findings. While the UUT proved well-suited for the goals of this study, future research could consider implementing tasks set in a more applied real-life context, such as generating ideas for restructuring a campus area (e.g., C. X. Chen et al., 2012) or addressing specific university-related problems (e.g., Rietzschel et al., 2006). Furthermore, incorporating multiple divergent thinking tasks, as recommended by Silvia et al. (2008), would enhance the reliability of creativity measurement across varied cognitive contexts. Equally important is the inclusion of convergent thinking measures alongside divergent ones, for example by using the *Remote Association Test* (Mednick, 1962). In this test participants identify a fourth word connecting three given words, for example, glass connects broken, clear, and eye. Integrating such a measure would allow future studies to assess both dimensions of creative thinking within a single experiment and to compare individual and team performance across both.

A second direction concerns the social and interpersonal dynamics within creative teams. The trust and sympathy effects identified in this study may be specific to dyadic teams, where the absence of a third member removes the moderating and disciplining influence that larger group structures can exert. Future research should therefore examine whether these effects replicate in larger teams and over longer interaction periods. It remains an open question, for instance, whether trust and sympathy stabilize after a second interaction phase or continue to develop, and whether individuals close the performance gap with dyadic teams over extended time frames. Beyond team size, controlling for ethnic and professional diversity would help illuminate the role of compositional factors in creative team performance. Future research should ultimately aim to identify the optimal team size for creative tasks, explicitly accounting for the trade-off between coordination costs, communication costs, and creative output. Building on this, an investigation of whether systematic differences in communication patterns exist between high- and low-performing creative teams would be of practical value, as communication styles are trainable and could serve as a lever for improving team creative performance.

A third and directly consequential direction concerns the work environment itself. As this study demonstrates that dyadic teams outperform individuals in the UUT in WFH setting, the question

naturally arises whether alternative workplace configurations could amplify this advantage further. A follow-up experiment directly comparing dyadic team performance in online versus in-presence settings would address this question. Beyond this binary distinction, the growing prevalence of hybrid work arrangements, combining remote and in-presence work, warrants dedicated investigation (Barrero et al., 2021), and is directly addressed in Part III of this dissertation. Finally, while this study focused on task instructions in a controlled environment as an important first step, real-world group creativity involves considerably more complex social dynamics. Future research should therefore consider additional dimensions of collaboration, including leadership, social roles, and process facilitation. The use of facilitators is well-researched with respect to required competencies (Nelson & McFadzean, 1998), contextual and phase-specific roles (Wardale, 2013), and facilitation techniques (Santanen et al., 2004; Wróbel et al., 2021). Exploring these factors further, including the potential of digital facilitation tools, represents a valuable direction for future research (Thorpe, 2016).

5.3 Practical implications

The results of this study carry some practical implications for employers seeking to optimize creative output. It should be noted, however, that as discussed in the limitations, the present findings are most directly applicable to the early, generative phases of creative work in well-defined problem spaces. With these boundaries in mind, the following practical implications can be derived.

First, the findings suggest that dyadic teams represent an effective unit for non-routine work such as brainstorming or idea development in R&D contexts. Importantly, as this advantage is demonstrated in an online setting, organizations are not restricted to pairing employees who share a physical location. Virtual dyadic teams can be formed across sites and regions, allowing organizations to draw on their full human capital and combine complementary talents, partly regardless of geography.

Additionally, the finding that real dyadic teams outperform simulated teams in terms of quality per quantity of ideas and making significantly fewer errors carries a direct practical recommendation. When the goal is efficient high-quality creative output, organizations should favor assembling real collaborative teams over assigning the same task to individuals working separately. The efficiency gains of genuine teamwork, both in terms of idea quality and answer accuracy, suggest that real collaborative interaction adds value that cannot be replicated by simply pooling individual contributions.

Moreover, the positive effect of trust on the creative performance of dyadic teams has direct managerial relevance. As trust requires time to develop, employers should ensure that virtual dyadic teams are given sufficient interaction time before performance expectations are set. Where possible, structured onboarding activities, such as introductory interactions or collaborative warm-up exercises prior to the actual creative task, should be considered, so that trust can exert a beneficial effect from the outset rather than only emerging in later phases of teamwork.

Finally, the finding that dyadic teams whose members differ in conscientiousness show lower creative performance suggests that individual personality traits should be considered when forming creative teams. Therefore, practitioners should account for trait compatibility alongside ability diversity when assembling dyadic teams for divergent thinking tasks.

5.4 Contribution

The study presented offers several contributions to the field of creativity research. It is motivated by the fundamental yet insufficiently answered question in literature: whether it is at all beneficial to have dyadic teams, rather than individuals, work on divergent thinking tasks. Additionally, the focus extends towards the initial formation phase, a period reported to be particularly challenging for creative collaboration, but which remains understudied in existing research (Reiter-Palmon et al., 2021). To the best of our knowledge, it is the first study to directly address both aspects in a virtual work setting with repeated interaction. Our findings validate the effectiveness of small, collaborative team structures in enhancing divergent thinking, which is an essential component of early-stage innovation, within remote work settings. Crucially, the study does not merely simulate a virtual work context in an experimental lab but enables an actual WFH situation, as participants completed the task from home. This strengthens the validity of findings and their transferability to real-world organizational structures.

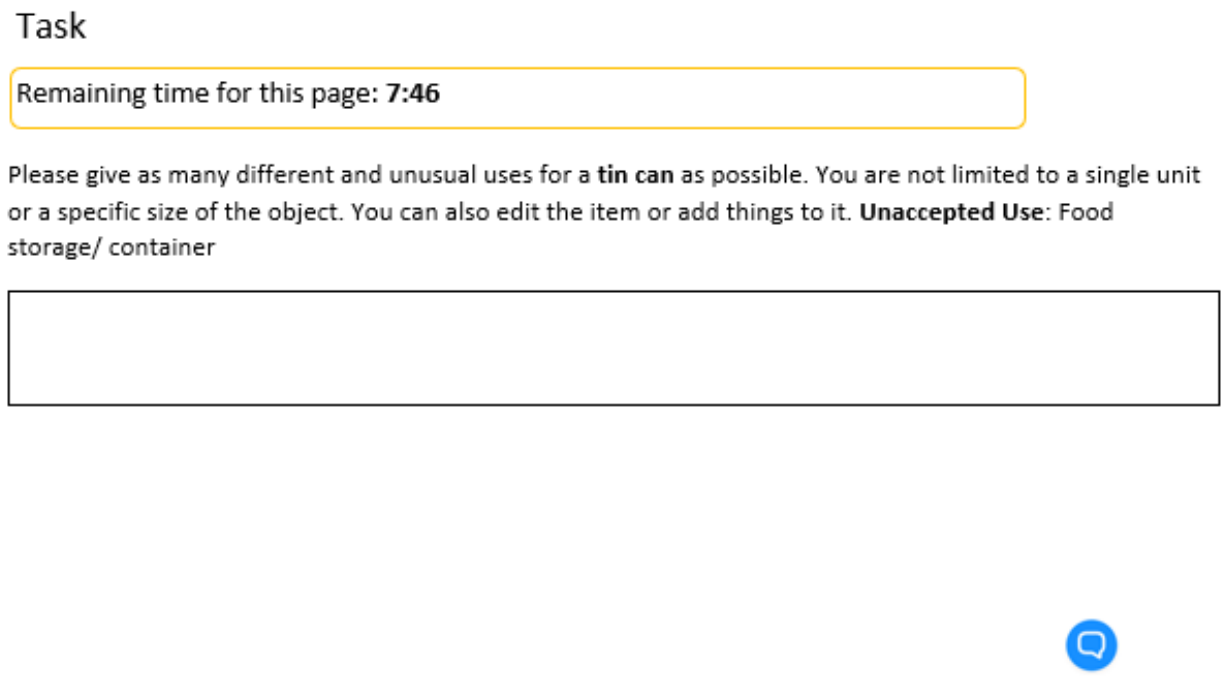
Beyond the individual-team comparison, the study's contributions extend to a broader set of factors shaping creative team performance, including trust, sympathy, and diversity. By examining team performance across repeated interaction phases, the study sheds light on how collaborative creative output develops over time, even within a limited number of work phases. By identifying trust as a time-dependent driver of dyadic team performance, it offers evidence-based guidance for the design of virtual teamwork in modern, flexible work environments.

A further methodological contribution concerns the implementation of the UUT as a creative task applicable to both individuals and teams in a virtual context. Existing experimental re-

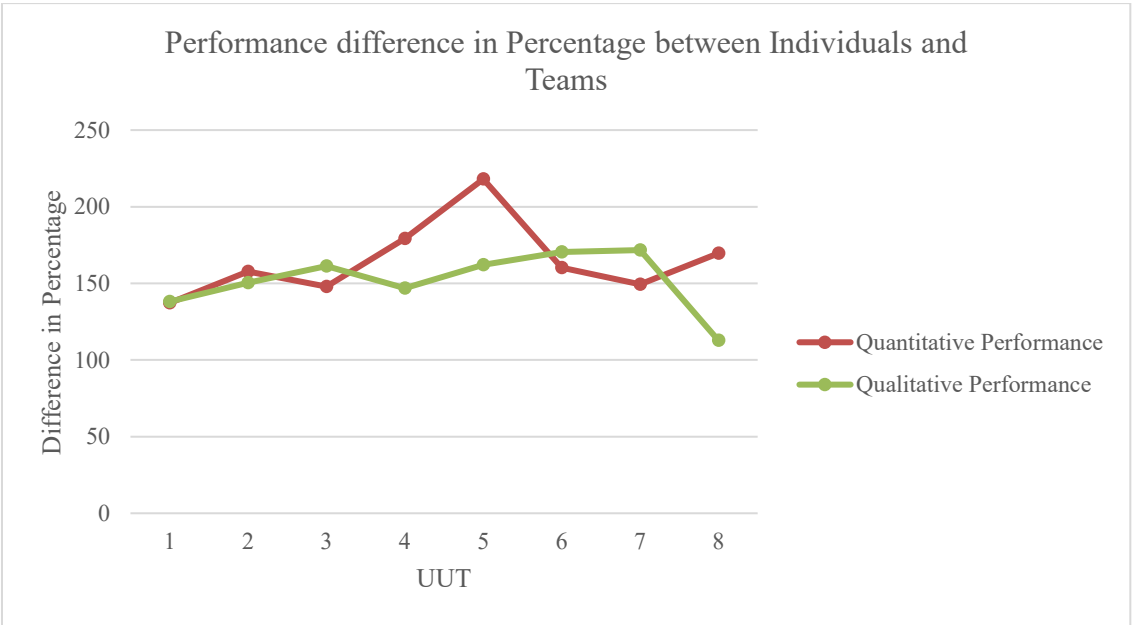
search on creative teams frequently either combines individually generated outputs into a simulated team score (e.g., Dutcher & Rodet, 2022) or restricts individual contribution through task-specific equipment (e.g., Grözinger et al., 2020). In our task variation, team members can contribute an unlimited amount of input and creative ideas in direct exchange, accounting for exchange and group discussions as occurring in real workplace teams. Thereby, contributing to a higher generalizability of research results generated in the lab and enlarging the task pool of researchers interested in creative performance of teams.

APPENDIX I

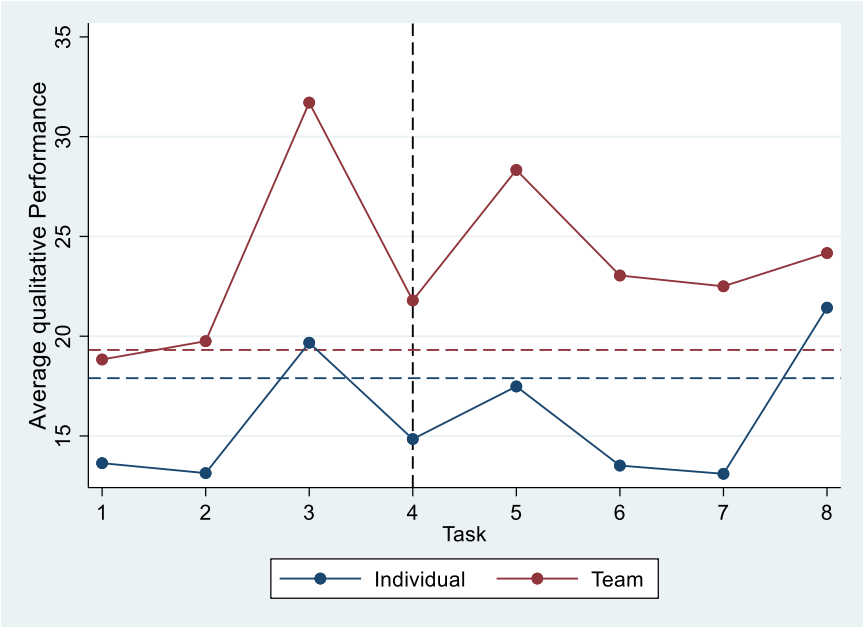
Figures



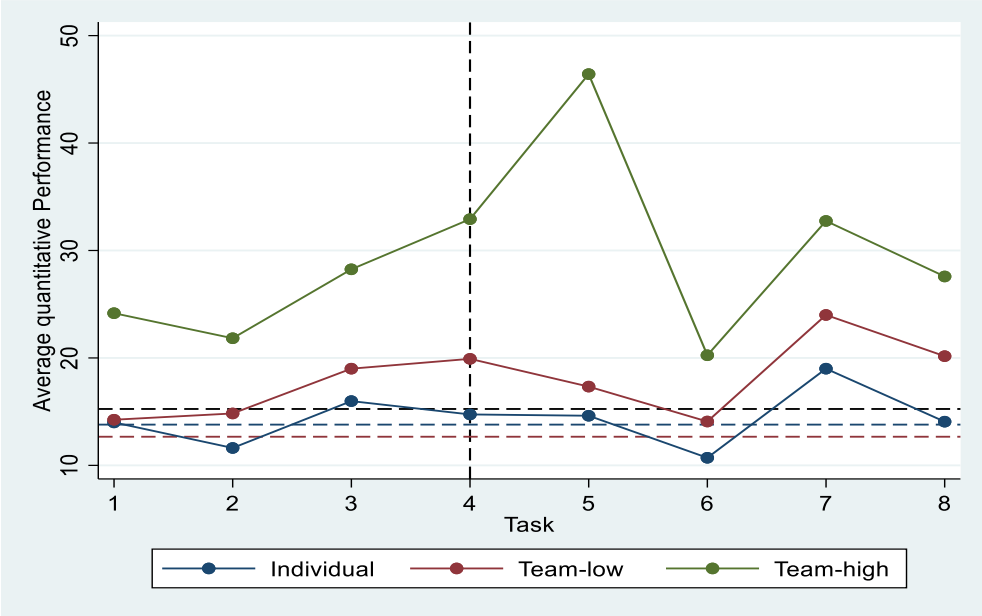
AI. 1: Experimental screen of the ability check task



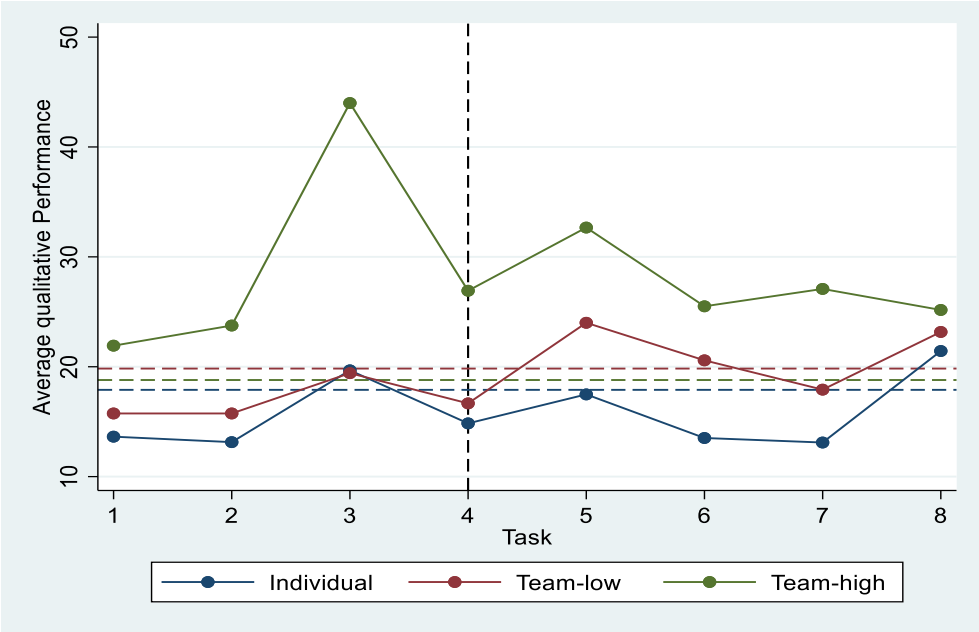
AI. 2: Difference between quantitative and qualitative performance between treatments



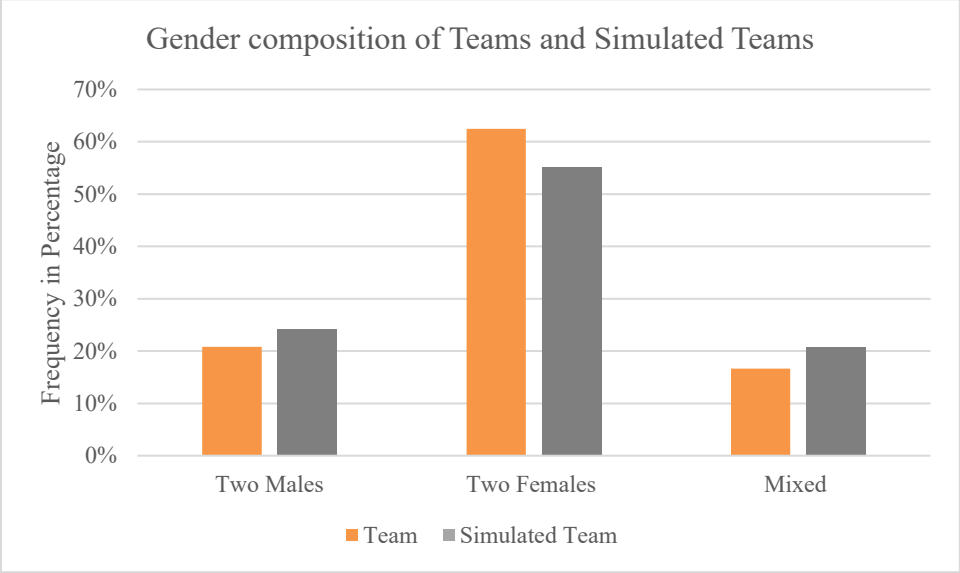
AI. 3: Average qualitative performance of treatments per UUT with average ability over both phases



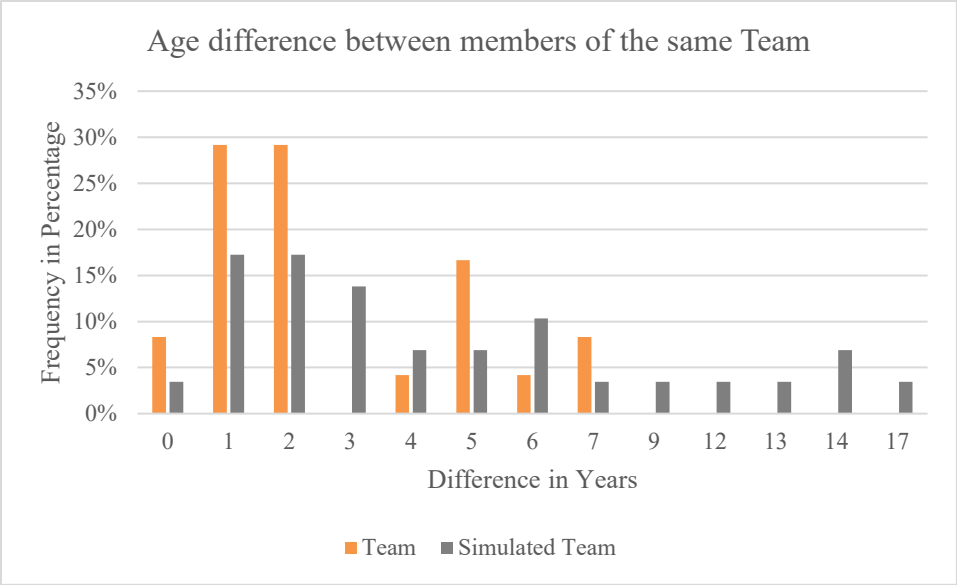
AI. 4: Average quantitative performance of individuals, low- and high-performing teams over both phases



AI. 5: Average qualitative performance of individuals, low- and high-performing teams over both phases



AI. 6: Comparison of gender composition of teams and simulated teams



AI. 7: Comparison of frequency of levels of age difference in years between teams and simulated teams

Tables

AI. 8: Spearman's rank correlation

VARIABLES	Quantity Phase 1	Quantity Phase 2	Quality Phase 1	Quality Phase 2	Errors Phase 1	Errors Phase 2
(1) Age	0.22	0.27	0.08	0.28	0.32	0.68***
(2) Agreeableness	0.27	-0.00	0.10	-0.08	0.28	-0.40*
(3) Conscientiousness	-0.03	0.07	-0.19	-0.04	0.01	0.16
(4) Extraversion	0.10	0.18	0.07	0.21	0.19	0.08
(5) Neuroticism	0.34	0.09	0.22	0.08	0.30	-0.03
(6) Openness	0.10	-0.07	0.00	0.13	0.01	0.07
(7) Satisfaction	-0.19	-0.22	-0.15	-0.15	-0.16	-0.27
(8) Preference WFH	-0.38*	-0.35	-0.37*	-0.40*	-0.11	-0.29
(9) Preference Teamwork	0.12	-0.05	0.06	-0.18	0.59***	-0.06
(10) Joy	0.05	-0.28	0.01	-0.12	0.18	0.10
(11) Effort	-0.03	0.12	0.00	0.05	-0.03	-0.08

Spearman rho = -0.329
*** p<0.01, ** p<0.05 * p<0.1

AI. 9: Pearson correlations of different performance measures, sympathy, and trust

VARIABLES	Quantity Phase 1	Quantity Phase 2	Quality Phase 1	Quality Phase 2	Errors Phase 1	Errors Phase 2	Sympathy Phase 1	Sympathy Phase 2	Trust Phase 1	Trust Phase 2
(1) Sympathy Phase 1	0.423**	0.433**	0.346*	0.335	0.346*	0.335	1.000			
(2) Trust Phase 1	0.341	0.345*	0.184	0.362*	0.184	0.362*	0.792***	0.354*	1.000	
(3) Sympathy Phase 2		-0.133		-0.180		-0.180	0.264	1.000		
(4) Trust Phase 2		0.289		0.267		0.267	0.482**	0.535***	0.645***	1.000

Note: Correlation of Sympathy and Trust of Phase 1 also considered for the performance measures of the second phase, as the initial level of sympathy and trust, developed in the first phase can influence the creative performance of the second phase. *** p<0.01, ** p<0.05, *p<0.1

AI. 10: OLS regression on interaction effect of extraversion on performance ratio of teams and simulated teams

VARIABLES	(1) Quantity Phase 1	(2) Quantity Phase 2	(3) Quality Phase 1	(4) Quality Phase 2	(5) Errors Phase 1	(6) Errors Phase 2
Simulated Teams	30.35	-15.65	-66.81	-62.12	59.44**	33.76**
	(50.26)	(68.94)	(59.68)	(88.35)	(25.93)	(14.67)
Extraversion	-2.721	-13.41	-6.685	-18.19	1.939	0.812
	(11.98)	(15.06)	(13.68)	(21.04)	(2.151)	(1.408)
Simulated Teams x Extraversion	-5.186	6.885	22.27	15.96	-12.26	-6.475
	(14.16)	(19.19)	(17.04)	(24.15)	(7.508)	(4.498)
Constant	96.67**	138.8**	115.7**	162.4**	1.846	6.833
	(43.54)	(55.12)	(49.62)	(78.78)	(7.741)	(5.011)
Observations	53	53	53	53	53	53
Adj. R-squared	0.084	0.004	-0.007	-0.029	0.238	0.431

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

AI. 11: OLS regression on interaction effect of openness on performance ratio of teams and simulated teams

VARIABLES	(1) Quantity Phase 1	(2) Quantity Phase 2	(3) Quality Phase 1	(4) Quality Phase 2	(5) Errors Phase 1	(6) Errors Phase 2
Simulated Teams	2.833	26.91	14.11	41.83	47.75	36.92*
	(38.25)	(60.81)	(47.00)	(73.50)	(39.29)	(19.47)
Openness	-0.749	3.505	-3.697	12.18	0.115	0.472
	(7.797)	(10.25)	(9.650)	(12.22)	(1.242)	(0.643)
Simulated Teams x Openness	3.005	-4.266	-1.896	-12.00	-7.873	-6.650
	(10.05)	(16.02)	(12.54)	(19.37)	(10.11)	(5.071)
Constant	89.96***	77.61**	106.5***	50.50	8.258*	7.865***
	(30.69)	(38.34)	(38.07)	(44.67)	(4.768)	(2.691)
Observations	53	53	53	53	53	53
Adj. R-squared	0.069	-0.019	-0.035	-0.034	0.224	0.434

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

AI. 12: Instructions

The instructions were originally written and used in German.

Instructions - Individual

Welcome, you can now open the link to the experiment, which you received in advance by mail, and click on the first "Next".

Welcome Page

Dear participants,

thank you for taking time to participate in the experiment. In the following, the procedure of the experiment and the task will be explained to you in more detail. Please switch your camera on and the sound to "mute".

Please turn off your cell phones now, listen carefully and do not use any other browser window during the experiment than the one in which the experiment is open. You can now open the link to the experiment, which you received in advance by mail.

Phase 1: The experiment consists of 2 phases. The first phase consists of one task, the second phase consists of 4 tasks that are identically structured. You have 8 minutes for each task. After these 8 minutes, you will be automatically redirected to the next task and time will start again. You will work on the tasks in break-out rooms. In the tasks, you will have to find as many alternative and creative usage options as possible for a given item. You will get an example of such a task explained right away.

Phase 2: This time the experiment consists only of the second phase from last week's session. Nevertheless, we will explain the task and scoring again in the following, so that there are no misunderstandings.

Now please click on "Continue" and confirm on the next pages that you agree to the privacy policy and the conditions of participation.

Privacy and Conditions of Participation

Phase 1: You can start the experiment later on instruction. You will now be reading a sample task, and the scoring scheme will be explained.

Phase 2: You can start the experiment later on instruction. After the processing time has expired, please leave the break-out rooms, switch to "mute" and please fill out one more questionnaire. You will now be reading a sample task, and the scoring scheme will be explained.

Example task - lunch box: Please write down as many different and unusual uses for a **lunch bo**x as possible. You are not limited to a single unit or a certain size of the object. You can also "edit" the items or add things to them. The conventional use of the object does NOT count as a valid answer. For example, for the lunch box, this would be using the box as a storage device/container for food. Please explain briefly their uses if they are not very obvious and do not just tell us the "final products". For example, for the item "newspaper" please do NOT just write "mirror" without explaining what exactly you mean by that. We can**not** evaluate such an answer!

Please enter your answers in the space below the task, separating them with a paragraph.

Evaluation & Payment: We will award points to each of your alternative uses. We will award one point for each correct or feasible and therefore valid answer, **two points if the answer is rare**, and **four points if it is particularly rare**.

Rarity is measured by the frequency with which exactly this answer was mentioned by others. Also, you will receive only one point for each term if more than one of your named terms has the same use. For example, for the three uses "weapon", "self-defense" and "defense" you will receive only one point instead of 3, because these uses are too similar. Also, there is **one point for each category** found for an object. For example, the answers "chair" and "stool" fall into the same category "furniture", while " badminton" falls into the other category "sports" and thus generates an additional point for a new category. For these 3 answers, there would then be 3 points for number, 2 points for the two categories, and possibly points for rarity. So, in summary, you can get points for the number of valid uses, the number of different categories, and the rarity of the proposed uses.

One point corresponds to a reward of 8 cents. Your achieved points from each round will be added up afterwards and the corresponding amount will be paid to you.

Phase 1: If you also show up for the second session, you will receive an additional fixed payment **of €4** at the second session in addition to the performance-based pay.

Phase 2: In addition to the performance-based compensation explained above, the compensation in this session **also** consists of a **fixed payment of €4.**

After the experiment you will be redirected to a payout page. Please note that the evaluation of your performance takes some time and the transfers may take longer. Therefore, payment will be made in two weeks at the earliest.

ATTENTION: If you leave or end the experiment prematurely yourself, you automatically lose any claim to compensation!

If you encounter any problems, please use the chat icon on your screen or the private chat in Zoom.

Phase 1: Now go to the break-out rooms and click on "Next" to start the first phase. Please leave your break-out room again afterwards!

Phase 2: Please click on "Next" now.

Ability Checker (first session only)

If you now click on "Next" again, you will see a page on which your ID is specified. Please use this as your zoom name now.

ID-Page

Phase 1: If there are still questions, please get in touch now via the chat icon! Otherwise, please go to the break-out rooms now and work on the following tasks of the second phase. You will receive all other instructions on your screen. You will be automatically directed to the following tasks after the timer has expired.

Phase 2: If there are still questions, please get in touch now! Otherwise, go to the break-out rooms now and, by clicking "Next", begin the task. You will receive all other instructions on your screen. You will be automatically directed to the following tasks of the experiment after the timer has expired.

Tasks + Questionnaire + Payout page

Phase 1: The experiment is now finished. Please remember to come to the second session in exactly one week at the same time. You will receive an additional payment if you participate again! You can now leave the Zoom room. Have a good day.

Phase 2: The experiment is now finished. Thank you for your participation. You can now leave the Zoom room. Have a good day.

Instructions - Team

Welcome,

you can now open the link to the experiment, which you received in advance by mail, and click on the first "Next".

Welcome- Page

Dear participants,

thank you for taking time to participate in the experiment. In the following, the procedure of the experiment and the task will be explained to you in more detail. Please switch your camera on and the sound to "mute".

Please turn off your cell phones now, listen carefully and do not use any other browser window during the experiment than the one in which the experiment is open.

Phase 1: The experiment consists of 2 phases. The first phase consists of one task, the second phase consists of 4 tasks that are identically structured. You have 8 minutes for each task. After

these 8 minutes, you will be automatically redirected to the next task and time will start again. You will work on the tasks in break-out rooms. In the tasks, you will have to find as many alternative and creative usage options as possible for a given item. You will get an example of such a task explained right away.

Phase 2: This time the experiment consists only of the second phase of last week's session. Nevertheless, the task and scoring will be explained again below so that there are no misunderstandings.

The group composition does not change.

Now please click on "Continue" and confirm on the next pages that you agree to the privacy policy and the conditions of participation.

Privacy + Conditions of Participation

Phase 1: You can start the experiment later on instruction. You will now be reading a sample task, and the scoring scheme will be explained.

Phase 2: You can start the experiment later on instruction. After the processing time has expired, please leave the break-out rooms, switch to "mute" and please fill out one more questionnaire. You will now be reading a sample task, and the scoring scheme will be explained.

Example task - lunch box: Please include as many different and unusual uses for a **lunch bo**x as possible. You are not limited to a single unit or a certain size of the object. You can also "edit" the items or add things to them. The conventional use of the object does NOT count as a valid answer. For example, for the lunch box, this would be using the box as a storage device/container for food. Please briefly explain their uses if they are not very obvious and do not just tell us the "final products". For example, for the item "newspaper" please do NOT just write "mirror" without explaining what exactly you mean by that. We can**not** evaluate such an answer!

Please enter your answers in the space below the task, separating them with a paragraph.

Evaluation & Payment: We will award points to each of your alternative uses. We will award one point for each correct or feasible and therefore valid answer, **two points if the answer is rare**, and **four points if it is particularly rare**.

Rarity is measured by the frequency with which exactly this answer was mentioned by others. Also, you will receive only one point for each term if more than one of your named terms have the same use. For example, for the three uses "weapon", "self-defense" and "defense" you will receive only one point instead of 3, because these uses are too similar. Also, there is **one point for each category** found for an object. For example, the answers "chair" and "stool" fall into the same category "furniture", while " badminton" falls into the other category "sports" and thus generates an additional point for a new category. For these 3 answers, there would then be 3 points for number, 2 points for the two categories, and possibly points for rarity. So, in summary, you can get points for the number of valid uses, the number of different categories, and the rarity of the proposed uses.

One point corresponds to a reward of 8 cents. Your achieved points from each round will be added up afterwards and the corresponding amount will be paid to you.

Phase 1: If you also show up for the second session, you will receive an additional fixed payment **of €4** at the second session in addition to the performance-based pay.

Phase 2: In addition to the performance-based compensation explained above, the compensation in this session **also** consists of a **fixed payment of €4.**

After the experiment you will be redirected to a payout page. Please note that the evaluation of your performance takes some time and the transfers may also take longer. A payout will therefore be made in one week at the earliest.

ATTENTION: If you leave or end the experiment prematurely yourself, you automatically lose any claim to compensation!

If you encounter any problems, please use the chat icon on your screen or the private chat in Zoom.

Phase 1: Now go to the break-out rooms and click on "Next" to start the first phase. Please leave your break-out room again afterwards!

Phase 2: If there are any problems please report immediately via the chat icon on your screen or via the private chat option of Zoom, otherwise you will lose any claim to compensation! If a team member drops out of the experiment, they will automatically lose any claim to compensation and the remaining member will also not be able to continue participating. Please try to avoid this at all costs!

Now click on "Continue"

**Ability Checker in the first session*.*

Phase 1: You will now work with another participant on the second phase in a break-out room. The following tasks are identical to the first task. The other participant with the same ID will be the other team member from you. The ID will be assigned to you in the next step.

Your combined performance now determines your payout from the next 4 tasks.

During the processing, you are welcome to share your screen with your other team member. At the end of the processing time, please fill out another questionnaire **individually**.

Since you are working on the task as a team and will also **be graded as a team from now on**, I ask you to agree on who will enter the answer choices in the text box. The other person **who does not work on the text field**, please enter the **ID** of the team in his or her text field, so that we can clearly assign you.

If there are problems regarding the internet connection or your other team member does not participate in the experiment anymore or does not appear in their break-out room at all, please report **immediately** via the **chat icon** on your screen or via the private chat option of Zoom, otherwise you will lose any claim to compensation! If a team member breaks out of the experiment, they will automatically **lose any claim to compensation** and the remaining member will also not be able to continue participating. Please try to avoid this at all costs!

At the end of the processing time, please leave the break-out rooms, switch to "mute" and fill out one more questionnaire.

Now click on Next.

Phase 1: If you now click on "Next" again, you will see a page on which your **ID** is specified. Please use this as your **zoom name** now.

Phase 2: Please click on "Next" again and change your zoom name to the ID from last week at the latest. You will also find this again in the e-mail.

Grouping & ID

Please move to the break-out rooms now and, once you have coordinated with your other team member, click "Next" to begin the experiment. Please agree on who will write the answer choices in the text box. The other person who does not edit the text box, please enter the team ID in his or her text box so that we can match you clearly.

If your other team member does not appear, please report **immediately** via chat. You will receive all other instructions on your screen. You will be automatically directed to the following tasks of the experiment, after the timer has expired.

After the processing time has expired, please leave the break-out rooms, switch to "mute" and answer the following questions alone or indicate how the statements apply to you.

If there are still questions, please get in touch now via the chat icon! Otherwise, please go to the break-out rooms now and start the next tasks by clicking on "Next". You will receive all other instructions on your screen. You will be automatically directed to the following tasks of the experiment after the timer has expired.

Tasks + Questionnaire + Payout page

Phase 1: The experiment is now finished. Please remember to come to the second session in exactly one week at the same time. You will receive an additional payment if you participate again! You can now leave the Zoom room. Have a good day.

Phase 2: The experiment is now finished. Thank you for your participation. You can now leave the Zoom room. Have a good day.

AI. 13: Questionnaires

The Questionnaires were originally written and used in German.

Scale: 1-5 (unless otherwise specified); Strongly disagree; Disagree somewhat; Neither agree nor disagree; Agree somewhat; Strongly agree

Phase 1:

All:

Which gender do you feel you belong to?

Male / female / diverse / do not want to state

How old are you?

XXX years old

What is your highest educational qualification?

High school diploma / Vocational training / Technical college entrance qualification / University entrance qualification (Abitur) / Bachelor's degree / Master's degree

Only for the team treatment:

Please indicate the extent to which you agree with the statements below regarding teamwork:

My other team member is sympathetic.

I knew the other team member before the experiment.

I would lend my other team member personal items of mine.

I would lend money to my other team member.

I would trust my other team member with my apartment keys.

I trust my team member.

Phase 2:

Only for the team treatment:

Please indicate the extent to which you agree with the statements below regarding teamwork:

My other team member is sympathetic.

I knew the other team member before the experiment.

I would lend my other team member personal items of mine.

I would lend money to my other team member.

I would trust my other team member with my apartment keys.

I trust my team member.

All:

Below you will find a number of characteristics that might apply to you. For example, would you say about yourself that you enjoy spending time with other people? For each of the following statements, please indicate the extent to which you agree.

I am someone who tends to be quiet.

I am someone who is compassionate, has a soft heart.

I am someone who tends to be disorganized.

I am someone who worries a lot.

I am someone who is fascinated by art, music, or literature.

I am someone who is dominant, acts as a leader.

I am someone who is sometimes rude to others.

I am someone who has difficulty getting started on tasks.

I am someone who tends to feel depressed, blue.

I am someone who has little interest in abstract ideas.

I am someone who is full of energy.

I am someone who assumes the best about people.

I am someone who is reliable, can always be counted on.

I am someone who is emotionally stable, not easily upset.

I am someone who is original, comes up with new ideas.

All:

Do you already have experience with mobile work?

Yes/ No/ Self-employed

Are you currently in an employment relationship?

Yes/ No

If yes or self-employed: In which form of work do you work?

Presence/ online/ hybrid

Please indicate the extent to which you agree with the statements below:

I am satisfied with my place of work where I conducted the experiment.

I prefer mobile working to working in presence.

I prefer to work in a team rather than alone.

I enjoyed the task.

I found the task to be exhausting to work on.

AI. 14: Categories

The following categories (originally in German) in which answers were sorted were predefined using pre-test data:

- Animals
- Furniture
- Garden / Plants
- Construction
- Tools
- Weapon
- Container
- Stand / Holder
- Decoration
- Toy Crafting
- Assistive Device (general, rather than things that are not necessarily a tool)
- Cutting / Piercing / Carving
- Writing / Painting / Drawing / Printing
- Transport / Locomotion Fuel / Material
- Sports
- Technology
- Accessories
- Clothing
- Textile
- Shielding from Environmental Influences (sun, privacy screen, etc.)
- Household (cleaning cloth, etc.)
- Fastening
- Instrument
- Crockery

- Cutlery
- Cooking (roasting string, etc.)
- Medicine

PART III: AN EXPERIMENT ON CREATIVITY IN VIRTUAL TEAMS

A shorter version of this essay has been published as: Grund, C., Harbring, C., & Klinkenberg, L. (2025). An Experiment on Creativity in Virtual Teams. *Journal of Economic Behavior & Organization, 231,* 106926.

Abstract: The organization of work and the characteristics of tasks have undergone considerable changes in recent years. The developments include (i) an increased relevance of virtual teams and (ii) a higher demand for non-routine tasks in organizations, including creativity. Existing research on creative teams focuses on one-shot or existing teams, overlooking the importance of the formation phase of teams. This formation phase is particularly relevant for teams working in a virtual workplace setting, where communication and coordination may be constrained by the environment. Next to virtual work, hybrid working models ascend, also for teams. Therefore, we examine the influence of workplace settings and changes in these settings on creative performance of teams. We also investigate whether the individuals' ability to choose their workplace affects creative performance. We answer those questions by conducting a 2-phase experiment with dyadic teams in the lab and online to model a presence and a virtual workplace setting and account for the formation phase of teams. We implemented the "Unusual Uses Task" as non-routine creative task. Our results showed that teams working in presence outperform those working online. Interestingly, working at least one phase in presence induces higher creative performance than entirely working online, underscoring the relevance of hybrid workplace settings. Moreover, no significant effects of self-selection on performance were found.

1 Introduction

The organization of work and characteristics of tasks have considerably changed over recent years. We observe a general increase in technologies enabling virtual communication and other forms of virtual work (Brenan, 2020). This, coupled with the demand for a global workforce (Kharroubi, 2021) and accelerated by Covid-19 pandemic restrictions, results in an increasing demand for virtual team work (Barrero et al., 2021). A second, still ongoing, change in the workplace is the high demand for solving non-routine tasks such as creative tasks (Autor et al.,

2003; Englmaier et al., 2024a). These developments result in (i) an increased relevance of virtual teams and (ii) a higher demand for creative work in organizations as part of non-routine tasks.

It is reported that virtual teams often face challenges, especially in communication and with that in efficiency, at the beginning of their virtual work (De Guinea et al., 2012). However, research mainly focuses on existing virtual teams, or those that work together for a single period. Consequently, issues concerning the formation phase of a team have not explicitly been addressed, but rather implicitly with regard to outcomes related to trust, communication and coordination (Abi Saad & Agogué, 2023; Vuchkovski et al., 2023).

Existing research already investigates the effectiveness of (i) virtual teams and (ii) incentives for creativity in teams (e.g., Bradler et al., 2019; Brucks & Levav, 2022; Englmaier et al., 2024a). Previous studies, which were mainly conducted before the Covid-19 pandemic and the general increase and familiarization of digital work, provide mixed findings concerning the effect of virtual work on performance in general and creative performance of individuals and teams (e.g., Alsharo et al., 2017; Dutcher, 2012; Giambatista & Bhappu, 2010; Kennedy et al., 2010). Several influential factors on the performance of teams have been assessed in earlier research, i.e., socio-demographic diversity; heterogeneity in abilities, and the physical workplace design (e.g., Auer et al., 2024; Dutcher & Rodet; 2022, 2023; Martens, 2011). Though, the effect of specific workplace settings on teams interacting repeatedly over multiple phases has not yet been analyzed.

Particularly, hybrid working models, in terms of switching between working in presence and online, gain in importance and are found to beneficially influence self-reported job performance (Iqbal et al., 2021; Naqshbandi et al., 2023). Therefore, we implement two working phases and also focus on these hybrid working models. This allows us, for example, to investigate whether an initial working phase in presence enhances the subsequent performance of a team that later works online in contrast to a team working fully online.

Thus, our study contributes to (and links) two different strands of research. First, non-routine creative tasks and performance of (newly formed) teams. Second, the influence of different workplace settings (i.e., on-site and virtual) as a consequence of their increased importance in teams.

In organizational reality, employees often select a company offering certain work settings and respective policies based on their preferences. Hence, considering the effect of self-selection seems essential for designing effective workplace settings for teams. Therefore, we also analyze

whether the individual's ability to choose where to work impacts the creative performance of teams.[31]

To conclude, we aim to shed light on the following research questions:

1. *Are teams more creative in presence than they are in an online setting?*
2. *What is the impact of the initial workplace setting for subsequent creative performance of teams?*
3. *Does the ability to self-select into a workplace setting affect the creative performance of a team?*

To approach these research questions and fill the current gap regarding teams working on non-routine tasks in specific and changing workplace settings, we conduct an (online) experiment with dyadic teams working on creative tasks, varying the workplace setting over two experimental phases one week apart, thus accounting for the formation phase of teams and extending the literature of one-shot creative teams (e.g., Charness & Grieco, 2019; Dutcher & Rodet, 2022; Englmaier et al., 2024a; Grözinger et al., 2020). The two experimental phases are either conducted at university or online in six experimental treatments. The implemented non-routine task to measure creative performance of teams is the UUT, in which participants have to list creative uses for a given object, covering the field of divergent thinking (Torrance, 1966).

2 Definitions and previous literature

The following sections provide an overview of the concept of creativity and the behavior of creative teams. Thereafter, the role of workplace, i.e., virtual or in presence, is theoretically and empirically discussed, building upon social presence theory and media-richness theory. Lastly, the opportunity to self-select into a workplace setting and its potential consequences for creative performance is thematized.

2.1 Creativity in teams

Complementarities among team members are a crucial issue for implementing teamwork (e.g., Brown et al., 2004). The argument has been applied to our context in the additive model of creative thinking, which posits that team members complement each other when searching for

[31] For studies allowing individuals to select teamwork or individual work and payment schemes see for example Eriksson et al. (2009) and Kuhn & Villeval (2015).

new ideas (Kurtzberg, 1998; Kurtzberg & Amabile, 2001; Yuan et al., 2022).[32] Similarly, the combinatory theory of creativity suggests that combining individuals to creative teams can foster creativity through active creating, diversity and experience (Feinstein, 2011, 2023).

Creativity, defined by Amabile (1988) as the production of novel and useful ideas by an individual or a group of individuals working together, is essential for innovative solutions. When focusing on creative work in teams, the concept of team-specific human capital (TSHC) is of importance. TSHC refers to the skills and knowledge individuals develop through interacting with one another (Blair, 1999) and a better understanding of tasks (Gerrard & Lockett, 2018). The development of TSHC is particularly crucial during the team formation phase. As TSHC increases with the time and quality of interaction between team members, it increases the effectiveness of coordination, trust and knowledge flow (McEvily et al., 2003; Gerrard & Lockett, 2018). In this sense, TSHC can be seen as a booster of com-elementalities in teams as described above.

Creative thinking has been distinguished between convergent and divergent thinking.[33] Convergent thinking covers generating a single best idea and implementing it, whereas divergent thinking focuses on generating as many new ideas as possible (Guilford, 1950). We contribute to the literature on divergent thinking and apply the UUT to assess divergent thinking as indicator for creative performance (Torrance, 1966). Solving the UUT, individuals have to generate as many creative and useful ideas as possible for everyday objects, such as a tin can. A creative use is a use for which the object was not designed for. The task is predominantly used for exploring the creativity of individuals (Bradler et al., 2019; Dutcher, 2012). Corresponding research on teams is limited to a context where team members work on the task in a one-shot game (e.g., Dutcher & Rodet, 2022, 2023). As an exception, Grund et al. (2026) recently investigated the difference in creative performance between dyadic teams and individuals using the UUT in a WFH setting. Dyadic teams performed significantly better regarding the quality and efficiency of creative ideas than individuals. Thus, we conclude that communication and complementarity in skills may beneficially foster creativity in teams compared to individuals in a WFH setting. Therefore,

[32] Dutcher and Rodet (2023) provide evidence for this concept in an experimental study, demonstrating that highly creative individuals can enhance the future creative performance of individuals with less creative individuals by sharing creative capital through team collaboration.

[33] The concept of idea generation via divergent thinking is viewed as a suitable indicator for creative potential (Runco & Acar, 2012) and as starting point of many creative processes as referred to in our contribution. Different approaches on measuring divergent thinking emerged over recent years (i.e., Reiter-Palmon et al., 2019; Weisberg, 2006) including imagining a future city or storytelling (e.g., Charness & Grieco, 2019). Note that convergent thinking complements this in terms of idea conceptualization and implementation (Mumford, 2003; Reiter-Palmon et al., 2019; Sternberg et al., 2024).

FtF interaction in presence might further enhance team creativity due to direct and rich communication available. We build on the previous studies and apply the UUT to the team context with varying workplace settings. Section 3 explains the experimental design as well as the treatments and measurement of creativity in further detail.

2.2 Workplace setting and creative teams

The importance of WFH as one specific dimension of virtual work has drastically increased over the past decades (Kurland & Bailey, 1999). Certain WFH arrangements can have beneficial effects for both employers (e.g., because of increases in productivity or cost reductions due to of less office space being required), and employees in terms of an increase in satisfaction, commitment and work-life balance (Bloom et al., 2015; Kaduk et al., 2019; Nakrošienė et al., 2019). In practice, the work setting can rotate between WFH and working in presence, leading to a hybrid model of work switching between FtF and CMC such as chats or video conferences (Barrero et al., 2021).

The social presence theory (Short et al., 1976) states that CMC leads to less effective relational communication as in FtF communication. Media-richness theory (Daft & Lengel, 1986) applies this concept to team communication and indicates that the more complex the message to be communicated, the richer the medium used should be. Here, FtF is the richest form of direct communication which can be strengthened through additional materials provided during communication. Rich communication is characterized by the number of communication channels available, the speed of a possible response and the perceived social presence of the partner.

Early experimental results indicated that communication via chats induces weaker collaboration (Brosig & Weimann, 2003), quality of idea generation (Kerr & Murthy, 2004) and less efficient transfer of creative capital between individuals (Dutcher & Rodet, 2023) compared to FtF communication. More recent research focusing on creative teams includes video conferences as an additional way of CMC. In sum, results indicate for better performance of video conferences compared to chats and worse results in comparison to FtF (Brucks & Levav, 2022; Dutcher & Rodet, 2021, 2023; Grözinger et al., 2020). In contrast to these one-shot lab experiments, Coenen and Kok (2014) hint for the particular relevance of the setting within the formation phase in teams in their field study.

We add to the literature by addressing the formation phase during repeated interaction in different workplace settings (working indeed from home vs. in presence). We also consider effects of switching between these settings like hybrid working models.

2.3 Self-selecting into a workplace setting

Self-selection into certain workplace settings is relevant in practice and imply the issues regarding which options are chosen and whether self-selected individuals perform better. Previous contributions on teams study the role of selecting teamwork or individual work (e.g., Kuhn & Villeval, 2015). Our experimental design enables us to take the decision to work in presence or online into account, as enforced work arrangements can have negative effects on employees (Kaduk et al., 2019). Granting autonomy to select a work setting may have positive effects on individuals' motivation as they may choose the setting that fits their preferences best and the autonomy itself may enhance individuals' intrinsic motivation as indicated by SDT (Amabile, 1983; Deci & Ryan, 1987).

To sum up, few studies investigating creative performance of teams and the role of the workplace setting exist. But – to the best of our knowledge – experimental research on the formation phase of a team and their creative performance with respect to different workplaces is missing. Therefore, our experimental study contributes to this literature by uncovering potential effects of the workplace settings and the sequence of those on creative performance of dyadic teams. Additionally, we approach the question of whether the autonomy to select a workplace setting has measurable positive effects on creative performance.

3 Experimental design, treatments and measures

To approach our research questions, an experiment, using the lab of RWTH Aachen University has been conducted. Our experimental design consists of two phases which take place one week apart from each other to account for repeated interactions over a longer time frame and to be able to analyze the formation process of a team over two working phases. In each phase, participants have to solve 4 UUTs together with one teammate. Additionally, subjects individually answer several questionnaires. Team composition does not change over the course of the experiment. Teams have 8 minutes for each UUT and are automatically forwarded to the following task or questionnaire afterwards. As we are interested in the influence of workplace settings, we implement the design in two different settings. First, we choose an online experiment via the video communication platform Zoom to account for the WFH setting.[34] In existing research, it has been argued that CMC inhibits creative idea generation but this is often still measured in the lab and not in a real home office setting (e.g., Brucks & Levav, 2022; Dutcher & Rodet,

[34] The participants had to confirm that they actually participated form their home.

2021); therefore, we aim to alter this in our experiment. Second, to account for work in presence, the experiment is executed in a PC lecture hall. This setting allows several groups to work simultaneously without disturbing each other (see Figure AII. 1 in Appendix AII).

Figure 5 shows the experimental procedure. Before the work in randomly assigned teams begins, the participants get a first questionnaire and the instructions are presented on the screen, as well as read out loud by the experimenter.[35] Next, they work individually on the first UUT which has two functions. First, the participants get to know the task and second, the results of this first task are used as a measure for the creative ability of an individual. The experimental screen of the ability check task can be found in Figure AII. 2 in Appendix II.

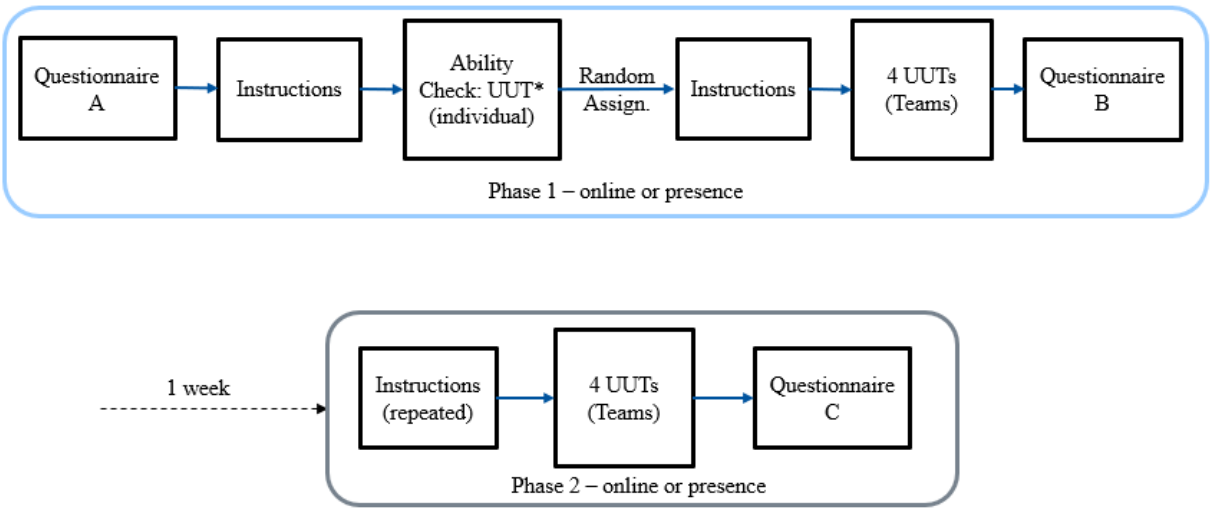


Figure 5: Experimental procedure

Subjects are randomly assigned to dyadic teams. We use the following items for the UUTs in the experiment: brick, hanger, sheet of paper and cord in phase 1, as well as umbrella, knife, tablecloth and cap of a plastic bottle in phase 2, as well as a tin can as ability check. In a pretest we have checked for the difficulty of the used items and divided them such that they are somewhat comparable concerning their difficulty over the four items across the two phases. Note, however, that difficulty differs across items in general so that a performance increase over time cannot be derived individually but only across groups. Additionally, the participants are given a short instruction on what the intentional use of those items is (e.g., tin can: container for food) and that this use is not counted as a valid answer. It is highlighted that they should be

[35] In PR as well as in HO, the instructions were read out loud either in presence or via Zoom and the experimenter was directly available for questions. Thereafter, all communication between subjects and experimenter was conducted via an implemented chat-tool on the screens of the experiment, such that there are no differences regarding the exchange between participants and the experimenter.

creative and the structure of the experiment as well as evaluation of answers and payment scheme are presented to them. The instructions can be found in AII. 15 in Appendix II.

Participants have to answer several questionnaires individually (see AII. 16 in Appendix II). In the beginning of the first phase, before the ability check, they are asked about their demographic data. This is implemented before the actual experiment as we aim to gather demographic information of the participants, even if they would drop out in the later course of the experiment. After the ability check and the 4 UUTs of the first phase the participants receive another questionnaire. With this questionnaire we ask whether participants knew each other before the experiment and for their levels of trust and sympathy towards the other team member, as those factors can play a major role in the process of teamwork and creative work.

The participants also receive a questionnaire after the 4 UUTs of the second phase. They are virtually asked the same questions as at the end of phase 1, as we are interested whether sympathy and trust change over the course of repeated interaction. Afterwards we implement 15 items regarding the Big 5 personality traits (the BFI-2-XS; Soto & John, 2017). Besides, they are asked about their experience with mobile work, their current situation of employment, the level of joy they felt during the experiment, and effort they needed to conduct the experimental tasks.

To evaluate the given answers of the participants and their performance, we use the objective evaluation method of Bradler et al. (2019), assessing three indicators of creativity: validity, flexibility and frequency (Guilford, 1959). The participants are evaluated as a team in the four team tasks per phase and individually in the ability task in phase 1. Section 1.4 in Part I of this thesis shows the different measures of creative performance used for quantity (i.e., validity) and quality (i.e., flexibility and frequency) of answers and discusses the objective evaluation method. Those evaluations are conducted manually by student assistants.[36] The frequency of an answer as well as the validity are determined with given answers of another study (Essay 1) with 154 participants, which was conducted beforehand and follows the same procedure and

[36] Two student assistants have worked separately on the evaluation of answers for one UUT. When they disagreed, a third rater was consulted. As we use the catalogue of another study to evaluate the answers, some answers from participants of the current study were not given in the other study. If this is the case, the assistants categorized the new answer, counted its frequency, and implemented it in the catalogue of the given data collection period. In the next collection period, the pre-existing, unchanged, catalogue was used again. We have opted for human raters instead of using latent semantic analysis (LSA) (e.g., Beaty et al., 2022; Dutcher & Rodet, 2023), as LSA – as it was available for us – might help streamline analysis but may not fully address the specific need to identify the frequency of individual ideas within a category – a key aspect of our analysis. While LSA can estimate how often a category is mentioned, it may not readily distinguish how uncommon or unique individual ideas are within that category, as the stated answers in one category are sometimes again grouped together when indicating similar uses (e.g., sword, saber or dagger in the category weapon).

uses the same UUTs. A rare answer is then equivalent to a usage which was named by approximately 2% of participants and a very rare answer corresponds to 0.6%-1.3% of participants who gave this answer. The evaluation scheme is presented to the participants via the instructions. Participants receive €0.08 for each performance point earned, along with a fixed payment of €4 for their overall participation in the experiment.

As we are interested in the effect of virtual workplace settings on the creative performance of teams, we implement treatments, in which we vary the workplace setting between virtual work from home (HO) and work in presence (PR) and change the sequence of workplaces in the two-phases. We use *HOHO*, *PRPR*, *HOPR*, and *PRHO* as notation for our four treatments, where *PRHO* indicates a setting, in which subjects worked in presence during the first phase and then switched to the WFH setting in phase 2. Additionally, we also want to investigate the effect of being able to self-select whether one wants to work in a virtual or presence workplace setting together with the other team member. The self-selection option is implemented via a question in the invitation mail for the experiments. The participants have to choose in their email whether they want to participate in presence or online and click on the corresponding option, which lead to two additional treatments *HOHO-endo* and *PRPR-endo*. Afterwards all participants, independent of treatment, are forwarded to the registration page where they are shown the optional timeslots, whereas for the endogenous treatments, the slots correspond to their decision made previously. After the registration for a specific timeslot, participants of the exogenous treatments get a follow-up email with the information where the experiment would take place (online or in presence) in both phases. So, they know beforehand whether they would work online or in presence and if the workplace changes for the second phase. Table 9 gives an overview of the workplace settings and their order for the six treatments.

Table 9: Treatments

Treatment	Workplace 1. Phase	Workplace 2. Phase	Self-selection option
HOHO	Virtual	Virtual	No
PRPR	Lab	Lab	No
PRHO	Lab	Virtual	No
HOPR	Virtual	Lab	No
HOHO-endo	Virtual	Virtual	Yes
PRPR-endo	Lab	Lab	Yes

4 Hypotheses

Addressing our research questions, we have already hinted for the relevance of social presence, media richness and TSHC in section 2. Following social presence and media-richness theory, the positive effects of more and richer communication channels and therefore the general quality of communication should be higher in a PR setting than in HO (Boland et al., 2022; Giambatista & Bhappu, 2010; Naotunna & Zhou, 2022; Nemiro, 2002). Next to changes in social presence, the effects of peer-pressure on the level of coordination in a team can differ between HO and PR. The literature on peer-pressure indicates that being observed may positively influence productivity and task focus in the context of team-based pay (Bishop, 2006; Georganas et al., 2015; Kandel & Lazear, 1992). Therefore, such positive effects of observability diminish in WFH compared to working in presence.

Focusing on the formation phase of teams, the development of TSHC might be positively influenced by direct communication in PR as the understanding of the task as well as collective learning can be positively affected by this form of communication leading to increased creative performance (Dutcher & Rodet, 2023). This relation may be explained by the level of coordination and trust in a team interaction (Peñarroja et al., 2013). As the social presence theory (Short et al., 1976) indicates, less perceived presence of team members through CMC in HO and thus social isolation might lead to slower development of trust between team members compared to interaction in presence (Abi Saad & Agogué, 2023; Ficapal-Cusí et al., 2023). This in turn influences the knowledge flow between members (Gerrard & Lockett, 2018; McEvily et al., 2003), which is crucial for the creative output of divergent thinking tasks (Kurtzberg & Amabile, 2001). Similarly, the combinatory model of Feinstein suggests that experience and active creating in teams, through direct interaction and communication, fosters creativity (Feinstein, 2011, 2023). Therefore, communication as in FtF interactions can have a general beneficial impact on creative teams.

Following those arguments, we expect higher TSHC in teams working in presence compared to those working online. Therefore, as TSHC influences the necessary coordination and flow of knowledge between team members working on a divergent thinking task, we expect significantly higher creative performance in the *PRPR* treatment when being compared to *HOHO* (H1).

Moreover, we extend existing research by investigating hybrid working models, simulated by the *HOPR* and *PRHO* treatments. Rather than focusing on comparing the two hybrid treatments

or determining an optimal sequence of hybrid work, our emphasis is on treatments with different initial modalities, allowing us to examine the influence of first interaction on subsequent creative performance in a different workplace setting. First, we examine the performance in phase 2 between *HOPR* and *PRPR*. As argued above, the level of TSHC is higher after having worked in presence in the first phase. Therefore, subjects can build on this and reach higher creative performance in the second phase in *PRPR* than in *HOPR* (H2a). Besides, subjects do not need to readjust in *PRPR* between phases, which also speaks for a higher performance in *PRPR*. Similarly, we can compare *PRHO* and *HOHO*. Although a readjustment is necessary in *PRHO*, we expect that the positive TSHC effect of PR in the first phase dominates. Therefore, the creative performance during the second experimental phase of the *PRHO* treatment should be significantly higher than that of *HOHO* (H2b).

Table 10: Hypotheses

Nr.	Hypothesis	Treatments
	Presence vs. WFH	
1	The creative performance in a team that is exogenously assigned to work in presence is higher than in a team that is exogenously assigned to work at home.	*PRPR* vs. *HOHO*
	Impact of the initial workplace setting on subsequent performance	
2a	After working from home in a first phase teams yield a lower creative performance in presence in a second phase compared to having worked in presence also in the first phase.	*HOPR* vs. *PRPR*
2b	After working in presence in a first phase teams yield a higher creative performance at home in a second phase compared to having worked at home also in the first phase.	*PRHO* vs. *HOHO*
	Impact of self-selecting into a workplace setting	
3	Teams where the team members have assigned themselves to a work setting show higher creative performance than those that were exogenously assigned.	*PRPR-endo* vs. *PRPR* *HOHO*-endo vs. *HOHO*

As the rationale behind implementing the self-selection option is that participants select themselves into the workplace where they expect to perform best – as they knew that payment is performance-based. We expect that participants are familiar with both workplace settings and are, thus, able to self-select according to their preferences and performance expectations. Moreover, the individual need for autonomy is positively influenced by the option to choose the workplace setting (Deci & Ryan, 1987). Thereby, increasing intrinsic motivation which is crucial for performance in general (e.g., Baard et al., 2004; Benz & Frey, 2008; Gagné & Deci, 2005) and creative performance in particular (Amabile, 1983). Therefore, the performance of the endogenously assigned treatments should be significantly higher than of the exogenously

assigned treatments, as participants should follow this rational of utility maximization (H3).[37] Table 10 provides an overview of our hypotheses.[38]

5 Data collection and sample

The data were collected between February and November 2023. In total, we conducted 33 sessions, each consisting of the two phases one week apart. The sample consists of 122 dyadic teams, leading to 244 subjects, who participated in both phases of the experiment. The sample consists of 101 women, is characterized by a mean age of 24.13 years, and 36.98% of participant indicated a bachelor's degree or higher.[39] In total, 459 participants were registered for the experiment from which 147 did not show up. Dropouts, first, include 17 participants who did not finish the first phase. Additionally, 68 subjects did not participate in the second phase, i.e., dropped out after finishing the first phase.[40]

Quantitative and qualitative performance of teams, as described above, serve as dependent variables. The treatments, differing in the workplace setting and self-selection option, are the main independent binary variables. The data gathered through the questionnaires are used as controls and for further analysis. The average qualitative performance in all treatments amounts to 99.75 points (SD=46.46) in the first phase and 138.44 points (SD=65.72) in the second. For the measure of quantitative performance, the average is 86.49 points (SD=29.09) in the first phase and 94.11 points (SD=33.39) in the second. On average over all treatments, the first experimental phase took about 58 minutes and the second about 45 minutes and the average total payment per participant amounts to €37.84, ranging from €33.52 to €41.42 across treatments.

The average individual quantitative ability of a participant, measured with the ability-checking UUT at the beginning of the first phase, amounts to 14.38 points (SD=5.71) and 17.33 points (SD=8.07) for qualitative performance over all treatments (see Table AII. 8 in Appendix AII for the ability of the different treatments). We checked that average points do not differ significantly across treatments using Kruskal-Wallis Tests (p=0.694).

[37] It needs to be remarked that individuals are only able to self-select and decide for a specific option when they know what to expect (Kuhn, 2018). Therefore, the assumption must hold that participants of the self-selection treatments are familiar with WFH and working in presence to be able to self-select into their preferred workplace, which is reasonable, as not only work but studying too is moved towards a digital environment.

[38] The experimental design, treatments, and the corresponding hypothesis of the study are preregistered at *AsPredicted* (AsPredicted #144051).

[39] Table AII. 3 to AII. 7 in Appendix AII show the distribution over the six treatments regarding the gender composition of teams, average age of participants, their educational background, field of profession, and origin.

[40] We address potential selection effects of those dropouts when focusing on regression analysis in section 6.

6 Results

The following chapter presents the results of this study. Firstly, the hypotheses are tested, using Mann-Whitne U-Tests and OLS regression analysis. The section is structured along the three research questions. Thereafter, further results focus on different levels of diversity of teams, i.e., ability and gender composition of teams, and the role of sympathy and trust between team members.

6.1 Hypotheses testing

We conduct pairwise comparisons between treatments, using non-parametric Mann-Whitney U-Tests. The results are calculated and presented on a team level, focusing on the joint team performance across treatments. Table 11 shows the average points achieved and their standard deviation (SD) in the two experimental phases and performance categories of the four exogeneous treatments together with the number (N) of dyadic teams per treatment. We aimed to raise at least 21 dyadic teams per treatment, in line with power calculation (Cohen, 1988: ANOVA with significance of 0.1, 6 treatments and power of 0.8). Qualitative performance is measured as the sum of the different categories generated and the frequency of answers given by the participants per UUT (see also Table 2, Part I section 1.4).

Table 11: Average quantitative and qualitative performance of the exogenous treatments in both phases with standard deviations in parentheses below

	N	Quantitative Performance			Qualitative Performance		
		Phase 1	Phase 2	Sum	Phase 1	Phase 2	Sum
HOHO	24	87.54	91.29	178.83	92.35	98.04	190.96
		(23.08)	(27.52)	(46.87)	(28.79)	(37.23)	(55.87)
PRPR	22	91.86	99.55	191.41	111.05	151.95	263
		(32.12)	(39.24)	(68.79)	(48.27)	(76.88)	(117.31)
PRHO	24	85.33	99	184.33	109.08	153.15	241.54
		(34.36)	(39.07)	(69.75)	(53.56)	(66.72)	(113.29)
HOPR	21	87.43	95.71	183.14	112.67	161.52	274.19
		(28.02)	(32.74)	(58.96)	(46.24)	(62.61)	(104.27)

Figure 6 shows the average qualitative performance and SDs (in parentheses) over the two phases of the experiment, together with the absolute differences in performance between the two phases of the four exogenous treatments. At first sight, this figure points towards a beneficial impact of a phase being conducted in presence compared to *HOHO*.

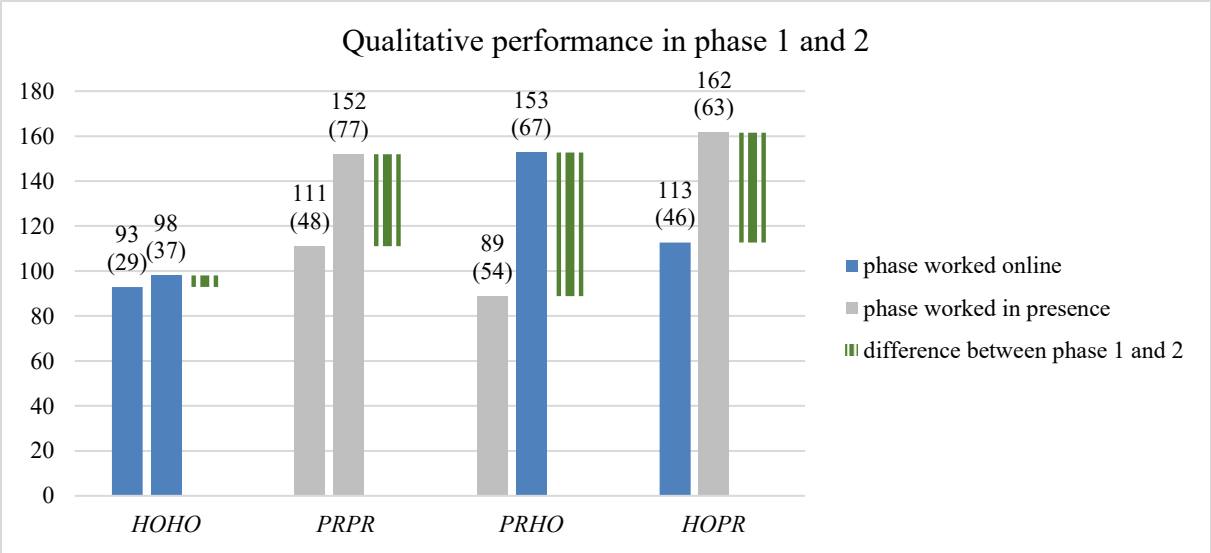


Figure 6: Average qualitative performance and SD in parentheses of the exogenous treatments in phase 1 and 2 and the absolute difference of performance between those two phases.

We find meaningful differences across treatments rather for qualitative (presented in the sections below) than for quantitative performance. Our analysis and hypotheses testing concentrates on the qualitive measure of creative performance as main indicator for creativity and innovative ideas by covering the aspects of flexibility and frequency of ideas. Besides, no difference is found regarding qualitative performance in the first experimental phase, using Mann-Whitney U-Tests for pairwise comparison (p-values ranging between 0.251 and 0.952). Obviously, advantages of the social presence and media richness cannot alone explain differences between working in presence and working from home in our setting. The effects of different workplace settings only materialize when it comes to repeated interaction in the same team. As argued above, TSHC builds up faster in presence during phase 1 and can blossom in phase 2.

Results on pairwise comparisons between treatments using Mann-Whitney U-Tests indicate significant differences between the exogeneous treatments of the experimental study.[41] The following text presents those results and is structured along our research questions.

Are teams more creative in presence than online?

Firstly, the question is addressed, whether working fully in presence outperforms fully working online when concentrating on dyadic teams working on creative tasks. This question is answered by testing hypothesis 1, expecting a beneficial impact of working in presence as a team

[41] Please keep in mind that the treatments do not differ in their creative ability of participants, such that the significant differences in creative performance presented in the following sections are not due to different levels of general ability to solve the UUT. See also the multivariate analysis below.

compared to dyadic teams working online. Table 12 shows the corresponding test results. All following p-values of the non-parametric tests are one-sided as the hypotheses under investigation are directional. Comparing the *HOHO* and *PRPR* treatments, they differ significantly in the measures of creative performance from each other such that the qualitative performance of the second phase is higher in the *PRPR* treatment. This result is in line with hypothesis 1.

Table 12: Results of pairwise comparisons of the qualitative performance using Mann-Whitney U-Test to test for the first hypotheses

H1	Treatments	Phase 1	Phase 2	Sum
	PRPR vs. HOHO	n.s. (one-sided p=0.205)	*** (one-sided p=0.006), r=0.41[42]	** (one-sided p=0.017), r=0.31
	HOPR vs. HOHO	n.s. (one-sided p=0.125)	*** (one-sided p<0.001), r=0.59	*** (one-sided p=0.001), r=0.46
	PRHO vs. PRPR	n.s. (one-sided p=0.358)	n.s. (one-sided p=0.354)	n.s. (one-sided p=0.407)

Further, we compare *HOPR* and *PRHO* with the treatments that do not change the workplace setting to check for the effect of working in presence during the second interaction. We find a significant difference between *HOPR* and *HOHO*, such that the performance of the second phase is significantly higher when it is conducted in presence indicating a beneficial effect of working in presence in a dyadic team, when being compared to teams who never worked in presence. The corresponding Mann-Whitney U-test for phase 1 (Table 12) just miss conventional significance levels. However, multivariate analysis including controls for socio-demographics as well and sympathy and trust show significant differences (see Table AII. 9 in Appendix II), hinting towards a positive impact of the consideration that the second phase is conducted in PR. As *PRHO* and *PRPR* do not differ in their performance of the second phase, it does not exist a negative connection of the second phase being online with creative performance, compared to the second phase in presence, as long as the first interaction is held in presence.

When comparing the absolute differences of qualitative performance between the two phases of the four treatments (Figure 6), it becomes visible that the difference of the *HOHO* treatment is the lowest. This finding also holds when comparing treatment differences pairwise by using Mann-Whitney U-Tests, i.e., the qualitative performance increase from phase 1 to 2 is significantly lower in *HOHO* than the increase in the other three. Therefore, the three treatments containing at least one experimental phase that is worked in presence, show a larger increase, pointing towards potential benefits from interacting at least once in a non-virtual environment.

[42] *r* denotes the effect size $r=\frac{z}{\sqrt{n}}$ of the difference between the two groups. Tested with Mann-Whitney U-Tests.

Therefore, our first guiding question is confirmed for our setting as teams show higher creative performance in presence than online.

What is the impact of the initial workplace setting on subsequent performance of teams?

As stated earlier, hybrid working models where employees switch between working in presence and online increased drastically over recent years. Therefore, the next section investigates the effect of those workplace variations in a team with focus on creative performance in the second experimental phase, thus focusing on hypotheses 2a and 2b. Table 13 shows the results of the Mann-Whitney U-Tests conducted to test the hypothesis.

Table 13: Results of pairwise comparisons of the qualitative performance using Mann-Whitney U-Test to test for the second hypotheses

Hypothesis	Treatments	Phase 2
2a	*HOPR vs. PRPR*	n.s. (one-sided p=0.219)
2b	*PRHO vs. HOHO*	*** (one-sided p=0.001), r=0.43

Hypothesis 2a tests for a difference between the *HOPR* and *PRPR* treatment, as the *HOPR* treatment interacted at first in an online work setting before switching to the same setting as *PRPR*. Because of the difference in the first interaction, the two treatments might differ in their creative performance of the second phase as the level of coordination and TSHC can differ. Though, there is no such significant difference, showing no significant negative or positive effect of working online when the second interaction is in presence, compared to working both phases together in presence.[43] Thus, rejecting hypothesis 2a.

Hypothesis 2b of this study expects higher performance of the *PRHO* treatment in the second experimental phase, when being compared with the *HOHO* treatment, as working the first phase in presence should have a beneficial impact on the creative performance of the second phase. The two treatments differ significantly as expected in the second phase, such that the creative performance of the *PRHO* treatment exceeded the performance of *HOHO*. The results are in line with hypothesis 2b, indicating a beneficial effect of meeting in presence in the first phase.

To further test for the influence of working in presence, as well as for other influential factors on creative performance of dyadic teams, a regression analysis is required.[44] We use a hierarchical OLS estimation to test stepwise for the treatment effects, with *HOHO* being the baseline,

[43] The participants knew before the first phase that they would interact the second phase in presence. This might influence the level of coordination and TSHC.

[44] The conditions to assume normal distributions are checked, with the result that normal distribution of the performance measures needs to be rejected. Therefore, an OLS regression with robust standard errors is applied. Our

and other influential variables. Table 14 shows results of the regression analysis for the four exogenous treatments on qualitative creative performance of the second phase.[45]

Table 14: Multivariate OLS regression analysis on qualitative creative performance of the second phase of the exogeneous treatments.

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Treatment (Base: HOHO)					
PRPR	53.91***	63.96***	39.85***	40.01***	43.18***
	(18.06)	(16.64)	(13.22)	(13.81)	(13.78)
PRHO	54.67***	47.52***	37.62***	36.62***	41.15***
	(15.61)	(15.49)	(11.85)	(11.88)	(11.89)
HOPR	63.48***	70.04***	46.93***	44.97***	45.24***
	(15.62)	(14.05)	(10.83)	(10.93)	(11.59)
Average Ability		3.574***	1.000*	0.978	0.916
		(0.892)	(0.586)	(0.595)	(0.632)
Qualitative Performance in Phase 1			0.931***	0.942***	0.887***
			(0.134)	(0.130)	(0.139)
Sympathy				-0.330	-0.679
				(10.00)	(10.16)
Trust				6.078	5.041
				(5.069)	(5.488)
Difference in Ability					0.593*
					(0.340)
Gender composition (Base: Only Males)					
Mixed					-6.950
					(10.25)
Only Females					-4.731
					(16.78)
Constant	98.04***	29.02	-7.777	-7.806	-8.410
	(7.610)	(18.02)	(15.02)	(15.32)	(18.84)
Observations	91	91	91	91	91
Adj. R-squared	0.120	0.317	0.610	0.606	0.605

Trust and Sympathy are centered around the mean
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The first model shows highly significant treatment effects of *PRHO*, *HOPR* and *PRPR* on the creative performance of the second phase, with *HOHO* being the reference treatment. The highly significant effect of *PRHO*, *HOPR* and *PRPR* in comparison to the *HOHO* treatment remains robust for Model 1 to 5. For the following models we integrate stepwise further variables, such as the average ability of a team (Model 2).

Moving from Model 2 to Model 3, we incorporate the qualitative performance of the previous phase. This approach allows us to account not only for the subjects' ability to solve the task but

results are robust, when using bootstrapping of the standard errors or with the log of performance as the dependent variable.

[45] The regression analysis focusing on the first experimental phase can be found in Appendix II Table AII. 9.

also how well they functioned as a team in the previous phase. The model demonstrates strong treatment effects even after considering previous performance, indicating performance improvements from phase 1 to phase 2.

In Model 4 sympathy and trust are introduced as additional explanatory variable for the creative performance of the second phase. Considering the measures of sympathy and trust between team members, Table AII. 10 in Appendix II reports sympathy and trust towards the interaction partner after the first and second phase at an individual level, separately for the six treatments. A performance enhancing effect of sympathy and trust was expected, as individuals who felt more sympathy and trust during the interaction should have a better level of coordination and communication patterns. When testing this assumption in Model 3, no significant direct influence of sympathy or trust is found. Besides, neither trust nor sympathy are able to explain treatment differences so that we do not find evidence for this particular theoretical argument presented above.

In Model 5 we implement two variables of team composition: the difference in ability between team members and the gender composition of teams, as gender effects and in general heterogeneity can influence the level of cooperation and thus the creative output (Dutcher & Rodet, 2022). No significant effect for the gender composition of a team is found, whereas the difference in ability has a highly significant positive, but rather small, influence on the creative performance, thus showing that heterogeneity in terms of ability might positively affect team coordination when working on a creative task.

We already hinted for the relevance of dropouts in Section 5 above. This is particularly relevant in the HO setting. Table AII. 11 in Appendix II reports mean ability and first phase performance of dropouts compared to our sample. We have checked that our results are robust with respect to a Heckman 2SLS estimation and are not biased due to sample selection (see Table AII. 12 in Appendix II).

We further test for possible interaction effects of sympathy, trust, and gender composition (see Table AII. 13 in Appendix II) in teams with the treatment effects to check for moderation. No significant interaction effects for sympathy nor trust and differences in ability are found. The gender composition of teams does not show a robust significant interaction effect with the treatments.

Additionally, we conduct Poisson-Regression analysis, with frequency (rarity of answers), flexibility (covered categories) and validity (quantity) as dependent variables (see Table AII. 14 in

Appendix II). The models show that the treatment effects are mainly driven by very rare answers (answers which were only given once or twice), thus, being a main indicator fur unusual innovative thoughts occurring in the exogeneous treatments which implement some form of working in presence.

Does the ability to self-select into a workplace setting influences creativity?

Moreover, we test whether implementing the option to choose the workplace setting influences creative performance. The first finding is that there seems to be a low preference for selecting into a fully presence workplace setting, as the *PRPR-endo* treatment only consists of 18 individuals, out of 62 participants in both endogenous treatments, conducted over the course of 10 months. Therefore, the results of the *PRPR-endo* treatment need to be treated with caution. Table 15 shows quantitative and qualitative creative performance of the endogenous treatments over the two phases, as well as for reasons of comparison those of *PRPR* and *HOHO* again. Results also show an increase in both measures of performance from the first to the second experimental phase.

Table 15: Average quantitative and qualitative performance of the endogenous treatments in both phases with standard deviations in parentheses below

	N	Quantitative Performance			Qualitative Performance		
		Phase 1	Phase 2	Sum	Phase 1	Phase 2	Sum
HOHO	24	87.54	91.29	178.83	92.35	98.04	190.96
		(23.08)	(27.52)	(46.87)	(28.79)	(37.23)	(55.87)
HOHO-endo	22	78.50	84.55	163.05	88.82	123.57	212.36
		(25.67)	(27.23)	(50.11)	(37.39)	(54.81)	(87.06)
PRPR-endo	9	91.00	94.89	185.89	117.50	157.56	273.78
		(34.48)	(34.53)	(67.91)	(48.50)	(83.14)	(129.59)
PRPR	22	91.86	99.55	191.41	111.05	151.95	263.00
		(32.12)	(39.24)	(68.79)	(48.27)	(76.88)	(117.31)

Hypothesis 3 suggests that the creative performance of dyadic teams is higher in the endogenous treatments, as the participants should self-select into their preferred work setting, depending on where they expect to perform best, as the performance directly influences payoff. Regarding the two *HOHO* treatments, the hypothesized relation holds for the qualitative performance of the second experimental phase (** (one-sided p=0.046), r=0.24). For the two *PRPR* treatments, no significant differences exist, although qualitative performance is slightly higher in the endogenous treatment. Although we find an indication for a positive effect of self-selection in the online treatment, we can hardly confirm the third hypothesis.

Lastly, we can reconfirm the results of hypothesis 1 by comparing the *HOHO-endo* treatment with the *PRPR-endo* treatment. A weak significant difference for the qualitative creative performance of the second phase is found in favor of the *PRPR-endo* treatment (* (one-sided p=0.089), r=0.25).

6.2 Further results

In the following, further results, not directly linked to the research questions of this study, are presented. Especially, the focus lies on potential factors affecting the communication and coordination process of teams and thereby creative performance.

Another factor of interest in the literature of creativity and teams is the role of heterogeneity between team members, whereby heterogeneity refers to socio-demographic aspects, as well as personnel characteristics as abilities. Therefore, we aim to investigate the influence of heterogeneity between members, concerning their ability differences and socio-demographic team composition, whereas the gender composition of teams is not thematized again, as already discussed during the main results of this study.

Ability

Ability is found to influence creative performance positively. Additionally, it is also already shown that the difference in ability, i.e., low and high ability participants are matched, can influence creative performance positively (Table 14 above). In the following, this general effect is considered, depending on the workplace in which the individuals interact with each other. As the virtual or present workplace comes with changes in communication, this might affect the role of differences in ability and ability in general. Therefore, Table 16 shows a regression analysis on qualitative creative performance during the two work phases, including the two measures of ability and an interaction term between ability and the place where each phase was conducted.

The results considering the first phase show that average ability is a strong and consistent predictor of performance, with a positive and highly significant effect across both specifications. Similarly, difference in ability is weakly positive in Model 1 but loses statistical significance once the interaction is introduced in Model 2. Importantly, the interaction terms with the workplace condition suggest that the moderating role of WFH in Phase 1 is limited. The interaction between HO in Phase 1 and average ability is negative but not statistically significant, indicating no robust evidence that remote work attenuates the effect of average ability at this stage. Likewise, the interaction with differences in ability is small and non-significant. Overall, in Phase

1, workplace conditions appear to have little impact on how team composition translates into performance.

Table 16: OLS regression on qualitative creative performance and interaction effects between measures of ability and workplace

	Phase 1		Phase 2	
VARIABLES	Model 1	Model 2	Model 3	Model 4
Workplace (Base: PR)				
HO in Phase 1	8.179	-15.40		
	(20.95)	(12.91)		
HO in Phase 2			20.69	-12.36
			(32.09)	(18.11)
Average Ability	2.803***	2.386***	5.834***	3.109***
	(0.837)	(0.551)	(1.664)	(0.887)
Difference in Ability	0.655*	0.502	0.564	2.079***
	(0.350)	(0.612)	(0.461)	(0.797)
HO in Phase 1 x Average Ability	-0.972			
	(1.145)			
HO in Phase 1 x Difference in Ability		0.330		
		(0.732)		
HO in Phase 2 x Average Ability			-3.523*	
			(1.931)	
HO in Phase 2 x Difference in Ability				-1.831**
				(0.918)
Constant	47.83***	57.84***	48.55**	70.55***
	(15.15)	(12.04)	(24.15)	(16.12)
Observations	122	122	122	122
Adj. R-squared	0.254	0.248	0.317	0.304

Difference in Ability is the absolute numerical difference between the measured ability of the two team members
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In contrast, the second phase reveals stronger interaction effects. Again, average ability is positively and significantly associated with performance under the baseline (PR). However, the interaction term HO in Phase 2 with average ability is negative and statistically significant, showing that the advantage of higher mean ability is substantially weakened when teams work remotely. Similarly, difference in ability has no significant main effect in Model 3. Once interactions are considered in Model 4, results reveal a significant interaction of the difference in ability of a team with working under HO. In other words, both higher average ability and greater heterogeneity enhance team performance in presence, but these benefits diminish considerably under the WFH condition.

Taken together, the results suggest that the impact of workplace conditions on the relationship between team composition and performance is phase and place specific. In the first phase, there is little evidence that remote work moderates the effects of average ability or ability differences; both factors exert their expected positive influence largely independent of workplace setting.

By contrast, in the second phase, the moderating role of remote work becomes pronounced: the positive effects of both higher average ability and greater ability differences are significantly reduced when collaboration takes place online. These findings highlight that while team composition consistently matters for performance, the extent to which its benefits materialize depends critically on both the workplace setting and the stage of the process.

Moreover, when focusing on the different ability levels of teams, the question arises whether all types of homogenous teams (i.e., both team members are low or both are high performers), are dominated by heterogenous teams. Figure 7 shows qualitative creative performance of the three different groups over both phases. It is shown that heterogenous teams outperform homogenous low and high ability teams over the whole experiment. This difference is statistically significant for the comparison of low-ability homogenous and heterogenous teams ($p<0.001$) but not for comparing high-ability homogenous teams with heterogenous ones ($p=0.183$) using Kruskal-Wallis Tests followed by Dunn's Test. Thus, high-ability individuals do not appear to be held back by lower-ability teammates; on the contrary, they seem able to lift their lower-ability partner's performance while still benefiting from the collaboration, consistent with previous findings (Dutcher & Rodet, 2022).

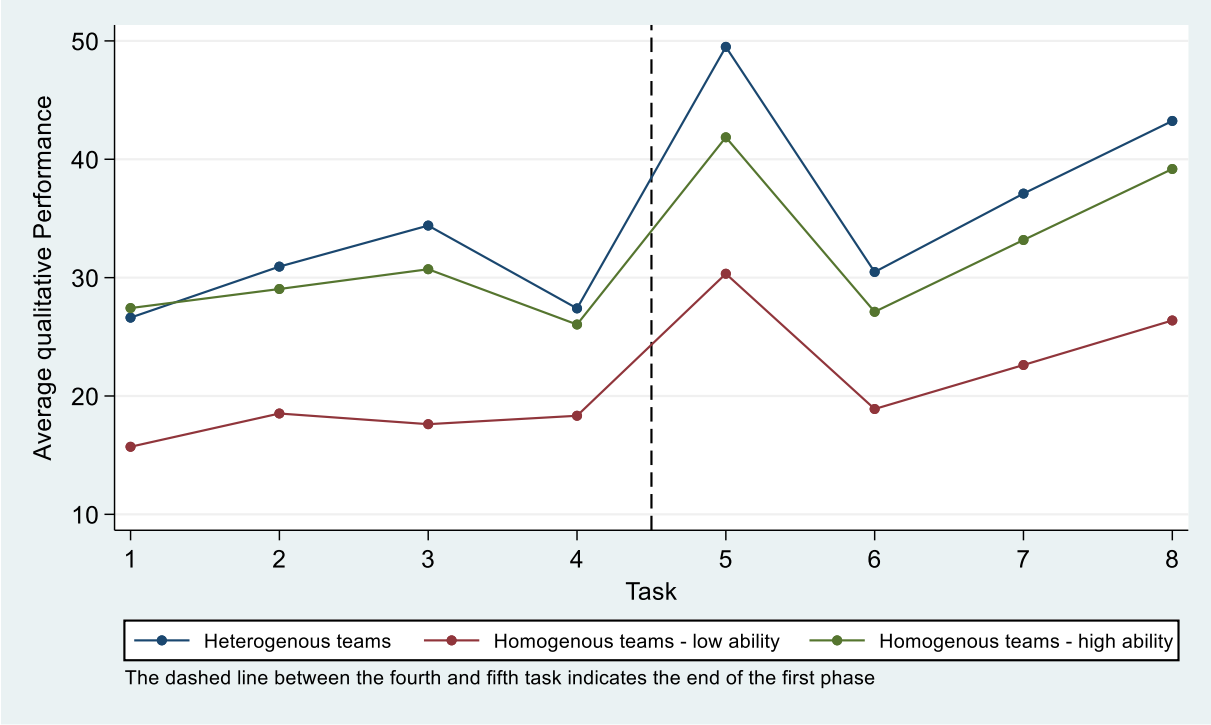


Figure 7: Average qualitative performance of different ability-paired teams over all tasks

Socio-demographic characteristics

In addition to differences in team members' abilities, potential socio-demographic characteristics should also be considered. Prior research suggests that heterogeneity within teams can positively affect the creative process by broadening the range of perspectives team members can use to complement one another (e.g., Dutcher & Rodet, 2022).

Table 17 shows the results of a regression analysis, regressing various diversity measures simultaneously on the qualitative performance measure from the first and second experimental phase. The analysis controls for individual ability on the UUT, as this variable consistently showed a significant effect on the outcome measures. Because the data of *HOHO* were collected as part of a previous study (Essay 1), not all socio-demographic items (e.g., education and profession) were included at that time. Consequently, Models (3) and (4) exclude observations from the *HOHO* treatment. Apart from the two ability measures, already discussed above, no other indicators of team heterogeneity display stable or significant effects.

Table 17: OLS Regression analysis on qualitative creative performance of phases 1 and 2 including different measures of heterogeneity

	(1)	(2)	(3)	(4)
VARIABLES	Phase 1	Phase 2	Phase 1	Phase 2
Average Ability	1.528***	2.826***	1.603**	3.179***
	(0.566)	(0.740)	(0.757)	(0.871)
Difference in Ability	22.52***	27.15***	20.78**	30.66***
	(7.968)	(10.31)	(10.34)	(11.33)
Difference in Age	-14.16*	0.118	0.238	-0.799
	(8.001)	(10.60)	(0.854)	(0.859)
Difference in Origin	11.57	27.18**	-9.799	-14.98
	(7.619)	(11.07)	(9.046)	(11.55)
Difference in Education			11.41	20.49*
			(9.951)	(12.02)
Difference in Profession			18.40**	18.63
			(9.192)	(11.27)
Constant	64.03***	62.01***	56.51***	81.79***
	(11.55)	(16.90)	(13.15)	(18.67)
Observations	122	122	98	98
Adj. R-squared	0.208	0.238	0.260	0.376

Robust Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Note: Models (3) and (4) do not include the *HOHO* treatment, as the data from this treatment stem from an earlier wave, where the items of profession and origin were not yet implemented in the questionnaire.

We compute age heterogeneity of a team, by calculating the age difference and conducted a median-split dividing half of the total sample in homogenous and heterogenous teams. Differences in age do not exhibit any meaningful effect on qualitative creative performance. The estimated coefficients are small and only reach significance in Model (1). Since this model includes the fully virtual treatment and Model (3) does not, the negative significant effect of age

heterogeneity in the first phase seems to be driven by the *HOHO* treatment. A similar pattern emerges for differences in origin: while Model (2) shows a significant positive effect for teams differing in origin (n=22), this result disappears in model (4). These findings suggest that certain salient differences between team members may play a more pronounced role when interaction occurs entirely online.

For heterogeneity in education, a significant difference emerges only for qualitative performance of the second phase, where heterogeneous teams (n=62) perform significantly better. The opposite pattern is observed for differences in professions (n=52), showing a significant effect in the first phase whereas for the second phase, heterogeneous professions slightly miss statistical significance (p=0.102). The positive direction of the coefficients for both variables supports the idea that team members from different fields may benefit more from complementing each other.

7 Discussion and conclusion

Our study contributes to the existing literature on virtual teams and creative teams by examining the impact of workplace settings on creative performance. Our results indicate that working together as a team on a creative task in presence beneficially influences performance compared to teams working entirely online. It becomes apparent that the positive effects of working together in presence spill over to subsequent online work phases compared to a pure online setting. This is particularly important considering hybrid working models: A team's initial working phase in presence provides a favorable basis for working online later. This is of special importance for teams engaged in ongoing creative tasks rather than one-shot brainstorming sessions. Over the course of the two experimental phases, solving creative tasks can be optimized by team members, leading to an established habit on how to work on the UUT, fostering the accumulation of TSHC and its favorable influence on performance (Gerrard & Lockett, 2018) over time. The results are robust to situations in which subjects can endogenously choose the workplace setting instead of being allocated by the experimenter, indicating that the intrinsic value of interaction in presence holds regardless of the method of workplace assignment (Cheng & Gong, 2018; Kuhn & Villeval, 2015).

Interestingly, when focusing on the initial workplace setting's influence on subsequent performance, no significant differences between teams that began in presence and those that transitioned from online to presence are found. There does not seem to be a harmful nor a beneficial effect of working online in the first phase as long as the following phase takes place in presence. Though, we find significant positive effects of getting to know each other in person in the first

(formation) phase of the team on subsequent performance online. We conclude that a beneficial effect of interacting in presence first can spill over to the subsequent work phase and place.

As the factors of sympathy and trust, as channel for TSHC, do not yield significant results in the regression analysis, we posit that other mechanisms driving the difference between treatments have to be relevant. A possible explanation for those results can be found in the pre-experimental awareness of participants whether they get to interact in presence at least once. Therefore, if an individual expects that he or she will have to interact in presence with the other member, reciprocal behavior and social norms may have an influence here. For example, participants probably will be more inclined for communication within their team when sitting directly next to the other team member, whereas in the online setting few turned off the camera, muted the microphone and needed active intervention of the experimenter to start communicating. Furthermore, enacting unfriendly or uncooperative during the first phase might be sanctioned by the team member through direct negative reciprocity in the next phase, which can be more unpleasant in presence. Consequently, FtF interaction, or even the anticipation thereof, can enhance the level of coordination between team members through reciprocal behavior and social norms. This could induce higher intensity of task focus, knowledge flow, and thus an increase in performance.

Moreover, anticipating FtF interaction might lead participants to develop a different perception of the team and their team affiliation, leading to higher pre-engagement and team commitment (Sheng et al., 2010). This can, in turn, have beneficial effects on team performance, as evidenced in Table AII. 9 in Appendix II. Here, we observe differences between fully online work and a hybrid model that begins with online work, potentially facilitated by the developed TSHC through shared knowledge and active communication in FtF interactions (Dutcher & Rodet, 2023; Liu et al., 2011; Neininger et al., 2010).

Those findings strengthen the notion that working in presence might not only provide the best technology for team interaction through FtF communication, but it also increases TSHC leading to high creative performance in subsequent interaction.

We conjecture that group performance is higher if individuals are allowed to self-select into a workplace setting following their preferences. Nonetheless, the results show no robust significant differences in performance between the exogeneous and endogenous treatments within the same workplace settings. Notably, a majority of participants opted for the online setting, likely due to its comfort and flexibility. As theory about self-selection states, one can only decide

rationally between options if one is familiar with all alternatives. We, therefore, asked participants about their experience of working online. Data indicates that participants selecting the online experiment are rather familiar with virtual work (65%) compared to those preferring the setting in presence (40%) (being in line with findings in the field from Nguyen (2021) and Labrado Antolín et al. (2024)). Thus, the interplay of familiarization with WFH and convenience when working online may have influenced the selection of the online setting which resulted in lower payoffs.

Moreover, we analyze the role of different ability-paired teams. Results show that the positive effects of both higher average ability and greater ability differences are significantly reduced when repeated collaboration during the second phase takes place online. Thus, hinting again for the relevance of time when studying the formation process of teams. Additionally, heterogenous teams in terms of their abilities are found to outperform low ability homogenous teams significantly, but not high ability ones. Thus, results point in the direction that a high ability individual is not dragged down by a low ability team member, and the aggregated team performance can be described as weighted average, where the creative potential of the most creative team member carries more weight (Pirola-Merlo & Mann, 2004). This is not found for other dimensions of heterogeneity, i.e., age, origin, education, profession or gender composition.

In the following, limitations of the study and implications for future research and the practical field are presented before highlighting the contribution of this study.

7.1 Limitations

Some limitations of this experimental study need to be pointed out. First of all, as already mentioned, the results concerning the *PRPR-endo* treatment need to be treated with caution, as the sample size of this treatment is relatively small with only 18 participants, i.e., nine teams. Considering the result of no significant difference between creative performance of treatments with and without the option to self-select into a workplace setting, this might only hold for a finite experimental design. When regarding actual workforce and the decision process behind job applications, this is strongly coupled with a self-selection process, not only to sectors (Roy, 1951) but also to workplaces (Danzer, 2019). Studies investigating involuntary working from home found negative consequences like higher turnover and lower job satisfaction (Kaduk et al., 2019). Furthermore, here presented self-selection treatments only include a selection of a preferred workplace, providing two options and a limited selection regarding the point of time of participation. No other possible specificities of self-selection are considered, e.g., task, or the decision to engage in individual or teamwork.

Moreover, the experimental study and its results do only cover dyadic teams, and its findings may not be generalizable to larger teams. Team dynamics can vary significantly with changes in size and therefore effects of, for example, sympathy and trust, but also measures of diversity might not be significant anymore or even gain significance. For example, the importance of trust might increase in a larger team as direct observation of others decreases (Hacker et al., 2019), opening the possibility for free-riding behavior. The creative performance of larger teams may also vary, as with a larger team size communication might be more complex and costlier, possibly decreasing performance. Nonetheless, in a meta study Byron and colleagues (2023) do not find a significant effect of team size on a team's creative performance.

Additionally, building up on the discussion on how to measure creativity, some researchers criticize the measure of divergent thinking as indicator for creativity (e.g., Runco & Acar, 2012), as idea generation is only one part of the creative process, which then needs to be followed by idea conceptualization and implementation (Mumford, 2003; Reiter-Palmon et al., 2019; Sternberg et al., 2024). While divergent thinking is a crucial aspect of creativity, the here presented results should be transferred to equal contexts, focusing on idea generation, as the generalizability to situations demanding convergent thinking skills of teams might be limited.

Moreover, the study aims to assess the creative performance of a newly formed team and its development. The analyzed time frame is limited to two phases, such as additional effects like an additional performance increase, increased importance of sympathy and trust, and the general alignment of TSHC build up in presence and online could appear when extending the timeframe. It is also possible that no beneficial effects of an additional time period exist, as fatigue concerning the somewhat repetitive task might increase. Furthermore, the chosen time frame per session is rather small, thus, not accounting for a whole working day. Nonetheless, most people engaged in divergent thinking tasks, like brainstorming, at an organization do not spend the entire working day performing these tasks.

Lastly, no measurement of communication habits and patterns between team members, which could be a possible element for explaining treatment effects, is implemented. Such measures could assess if high performing teams differ in their type of communication and how online communication differs to communication in presence.

7.2 Implications for further research

In the section above, some limitations are outlined, which lead towards implications and avenues for further research, focusing on creative performance of teams.

Beyond extending the time horizon or increasing the team size, further studies should focus on the implementation of a measure for convergent thinking to broaden the generalizability of the present results across multiple facets of creativity. A suitable task is the *Remote Association Test* in which participants receive three words and have to identify the common associative link between them (Mednick, 1962). Such a test can be added after the UUT, allowing researchers to capture both divergent and convergent thinking, thereby enriching the scope of creativity assessed and offering a more comprehensive understanding of how workplace environments shape creative output.

Follow-up studies could also implement audio-recording software during team interactions online but also in presence. Such data could provide a richer understanding of the actual communication process. Using a machine learning algorithm, such recordings can be analyzed to determine the number of words, sentences or minutes of direct communication, similarities in style of articulation and patterns of idea formulation (Beaty et al., 2022; Dutcher & Rodet, 2023), off-task conversations, and dominance or leadership behavior of one team member (Englmaier et al., 2024b).

Next to verbal communication, it could also be insightful to study virtual non-verbal communications. For instance, one can assess whether participants turn their cameras on during the online phases, show their real background, use a filter, or even use an avatar for communication with the other member. Those aspects might also influence the quality of communication and with that TSHC and creative performance (Garau et al., 2003; Rodeghero et al., 2021).

Building on the idea of analyzing the actual social interaction between team members, future work could also focus more on the gender composition of a team and whether this influences the verbal and non-verbal communication patterns (Hardt et al., 2025).

Given that this study approaches creativity from an economic perspective, i.e., performance, another crucial dimension in economic efficiency-related research is the incentives set for individuals for engaging in a task and producing a desired output. In the present study a piece-rate incentive coupled to a fixed base payment is implemented. However, the broader literature on monetary incentives and creativity is divided regarding whether such incentives hinder or enhance creative performance (e.g., Amabile & Pratt, 2016; Charness & Grieco, 2019; Gross, 2020). Future research could therefore expand the current design by varying not only the workplace conditions but also the incentive mechanisms. Introducing, for example, between-team tournament incentives, purely fixed payments, or mixed schemes could help disentangle how

different reward structures interact with team dynamics and work environments to shape creativity. Admittedly, such an expanded design would require a considerably larger sample to ensure adequate statistical power.

7.3 Practical implications

The results of the presented essay contain valuable insights for managing creative processes in organizations. Induced by technological progress, and further motivated by the Covid-19 Pandemic, a drastic increase of virtual work occurred. Yet, this trend could shift back in the next years, as companies and CEOS consider a return-to-office policy. According to KPMG's survey of 1,300 CEOs (2023), 64% expect a full return to office work within three years, and 87% are considering financial incentives or promotions to encourage employees' return.

However, this rather pessimistic view on virtual work by CEOs is not entirely supported by the findings of the presented study, at least for creative teamwork, which is an elementary aspect of innovation and competitive advantage. Our results indicate that hybrid workplace settings can function as good as working fully together in presence, when focusing solely on the aspect of work performance. As virtual work can come with benefits for the employee, i.e., increases in job satisfaction and work-life balance (e.g., Bloom et al., 2015; Kaduk et al., 2019; Nakrošienė et al., 2019), organizations may benefit from adopting hybrid or flexible in-person models for teams engaged in creative tasks.

As companies reconsider their workplace policies, they should pay particular attention to the impact of work arrangements on team creativity. The study suggests that some degree of in-person interaction remains important for effective creative collaboration. This can occur in an initial phase, like a typical onboarding day, or a subsequent stage within the team formation process, as long as in-person interaction is communicated in advance. To support the effective formation of a team, incorporating at least one in-person collaboration session during the onboarding period may yield valuable benefits for team dynamics. Such in-person interaction can enhance team performance in later stages, regardless of whether future collaboration occurs remotely or in person.

7.4 Contribution

We provide an important forward to a deeper understanding regarding (switches in) workplace settings of teams and their performance in context of creative tasks. While creativity in newly formed teams has been partly studied, and remote work has received increasing attention in recent years, the intersection of these two domains remains underexplored. By examining how

creative performance unfolds in different interaction environments, we provide evidence that links these strands and speaks directly to ongoing discussions about the effectiveness of virtual and hybrid work arrangements.

A central contribution lies in our experimental design, which allows us to move beyond one-shot observations of team creativity. We implement a two-phase experiment with dyadic teams working on creative tasks one week apart, enabling us to capture the formation phase of teams rather than observing only a single, isolated interaction. Despite its relevance for understanding how groups develop creative potential, this phase has received little attention in experimental studies of workplace environments.

By systematically varying whether teams worked in presence or virtually in each phase, we could identify not only the effect of the workplace setting itself but also how teams evolve when their interaction environment changes over time, thus enabling us to shed light on hybrid working models. This dynamic perspective highlights how certain arrangements allow teams to retain or enhance their creative performance. Our results show that in-person interaction tends to foster higher creative output compared to fully virtual interaction, and that experiencing at least one on-site phase can mitigate some of the limitations of virtual work.

Furthermore, by exploring whether teams benefit from having autonomy in choosing their workplace mode, we add evidence on the role of self-selection on creative performance. This addition provides initial evidence on whether perceived autonomy over the work environment enhances creative performance or whether the characteristics of the setting itself remain more decisive.

Overall, by combining a novel experimental design with the examination of dynamic workplace sequences and team formation processes, our study offers a deeper understanding of how virtual, presence, and hybrid arrangements shape creative team performance. These contributions not only fill a gap in existing literature but also lead to our main finding: while virtual collaboration can hinder creativity, hybrid structures that incorporate at least some interaction in presence can substantially improve creative outcomes in small teams.

APPENDIX II

Figures



AII. 1: Experimental setting in persence

Task

Remaining time for this page: **7:46**

Please give as many different and unusual uses for a **tin can** as possible. You are not limited to a single unit or a specific size of the object. You can also edit the item or add things to it. **Unaccepted Use**: Food storage/ container

AII. 2: Experimental screen of a task

Tables

AII. 3: Gender composition of dyadic teams for all treatments

	Gender composition of teams		
	male	mixed	female
HOHO	5	15	4
PRPR	7	12	3
PRHO	9	12	3
HOPR	12	5	4
HOHO-endo	2	14	6
PRPR-endo	3	5	1

AII. 4: Age of the participants divided for all treatments

	Mean	Std. Deviation	Min	Max
HOHO	25.94	2.97	22	33
PRPR	23.00	4.56	19	44
PRHO	22.58	3.70	18	35
HOPR	24.38	3.67	18	39
HOHO-endo	25.27	6.69	19	55
PRPR-endo	22.278	5.17	18	40

AII. 5: Educational level of the participants divided for all treatments

	GCSE	Training	GCE A-levels	Bachelor	Master	PhD	Other
HOHO	0	3	18	23	4	0	0
PRPR	0	1	30	8	3	0	2
PRHO	1	0	34	8	3	1	1
HOPR	0	1	25	12	3	0	1
HOHO-endo	1	1	24	13	4	0	1
PRPR-endo	0	0	13	4	1	0	0

AII. 6: Profession of the participants divided for all treatments

	Social Science	Health	Engi-neering	Art, Music, Design	Teach-ing	Natural science	Linguis-tic and culture	Econom-ics	Public admin-istration
HOHO	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
PRPR	0	0	28	0	0	10	1	5	0
PRHO	0	2	33	0	0	3	0	9	1
HOPR	2	0	25	1	1	5	0	8	0
HOHO-endo	1	1	26	0	1	3	1	9	1
PRPR-endo	2	0	8	0	1	5	0	2	0

Note: The data used for HOHO were collected during another study (Essay 1). The questionnaire used in this previous study did not include a question regarding profession.

AII. 7: Origin of the participants divided for all treatments

	EU	Non-EU
HOHO	N.A.	N.A.
PRPR	41	3
PRHO	38	10
HOPR	34	8
HOHO-endo	38	6
PRPR-endo	17	1

Note: The data used for HOHO were collected during another study (Essay 1). The questionnaire used in this previous study did not include a question regarding origin.

AII. 8: Average creative ability and standard deviation in parentheses measured on the individual level

	Creative Ability	
	Quantitative	**Qualitative**
HOHO	13.33 (7.99)	17.98 (11.67)
PRPR	14.75 (6.36)	16.5 (8.01)
PRHO	13.42 (6.34)	18.25 (11.08)
HOPR	13.71 (6.69)	17.71 (11.46)
HOHO-endo	12.59 (5.97)	14.61 (10.79)
PRPR-endo	13 (5.45)	18.22 (10.67)

AII. 9: Multivariate OLS regression analysis on qualitative creative performance of the first phase of the exogeneous treatments

VARIABLES	Model 1	Model 2	Model 3	Model 4
Treatments (Base: HOHO)				
PRPR	18.13	25.90**	25.17**	26.80**
	(11.95)	(11.48)	(12.03)	(12.12)
PRHO	16.17	10.64	8.917	15.31
	(12.02)	(11.81)	(12.19)	(11.96)
HOPR	19.75*	24.83**	23.71**	20.92*
	(11.78)	(10.61)	(10.99)	(12.28)
Average Ability		2.764***	2.770***	2.479***
		(0.572)	(0.582)	(0.554)
Sympathy			4.047	4.850
			(8.539)	(9.401)
Trust			-0.181	-4.844
			(5.654)	(6.093)
Difference in Ability				0.792***
				(0.304)
Gender composition (Base: Only Males)				
Mixed Team				-10.15
				(11.11)
Only Females				-0.722
				(13.17)
Constant	92.92***	39.53***	40.32***	36.54**
	(5.899)	(12.77)	(13.01)	(14.71)
Observations	91	91	91	91
Adj. R-squared	0.001	0.256	0.241	0.269

Trust and Sympathy are centered around the mean
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

AII. 10: Average Sympathy and Trust in both phases for all treatments

	Sympathy		Trust	
	1	2	1	2
HOHO	4.48	4.56	3.54	3.69
PRPR	4.61	4.48	3.36	3.61
PRHO	4.88	4.63	3.88	3.88
HOPR	4.71	4.69	3.50	4.17
HOHO-endo	4.41	4.39	3.18	3.41
PRPR-endo	4.78	4.67	4.00	4.11

AII. 11: The potential role of selection effects due to Dropouts

	Groups in Phase 2			Dropouts after Phase 1		
	N (Groups)	Mean Ability	Performance Phase 1	N (Groups)	Mean Ability	Performance Phase 1
Whole Sample	122	18.73	106.12	34	19.21	91.12
HOHO	24	17.98	92.35	12	22.17	93.83
PRPR	22	16.5	111.05	4	19.75	89.5
PRHO	24	18.25	109.08	6	14.67	105.5
HOPR	21	17.71	112.67	3	29.33	98.33
HOHO-endo	22	14.61	88.82	8	14.63	78.25
PRPR-endo	9	18.22	117.50	1	15	60

To further test for potential effects of selection through dropping out, we conducted a Heckman 2SLS regression with robust standard errors and included the difference in age between team members as an additional new variable in the selection model (Table AII. 12). Age diversity can influence dynamics such as collaboration, motivation, and continuity, which in turn impacts both team performance and individuals' preferences for team engagement. These effects can arise if different age groups bring distinct perspectives, work styles, and expectations, which can affect cohesion and the overall working atmosphere (e.g., Wegge & Schmidt, 2009; Wegge et al., 2012). In this context, age difference might serve as a predictor of participants' willingness to continue in a team-based experimental setup. Research suggests that greater age diversity can sometimes lead to friction or differences in communication and decision-making preferences, potentially making collaboration less appealing or more challenging for team members (Paoletti et al., 2020). Consequently, age difference could affect participants' initial selection or dropout decisions. This is descriptively found in our study, using Mann-Whitney U-Tests comparing the dropouts with the final sample as indicated above. The Heckman 2SLS model shows that the treatment effects are robust. Besides, evidence for a selection bias cannot be found (insignificant inverse Mills ratio).

AII. 12: Heckman 2SLS selection model

VARIABLES	Qualitative Performance Phase 2	Prob. Participation in Phase 2
Treatment (Base: HOHO)		
PRPR	51.28***	0.617
	(16.08)	(0.381)
PRHO	48.61***	0.449
	(15.12)	(0.352)
HOPR	55.73***	0.741*
	(17.85)	(0.407)
Average Ability	0.781	-0.00924
	(0.671)	(0.0191)
Qualitative Performance in Phase 1	0.925***	0.00243
	(0.132)	(0.00361)
Sympathy	0.873	0.571
	(10.49)	(0.26045)
Trust	5.121	-0.12608
	(6.411)	(0.20986)
Difference in Ability	0.583*	-0.001
	(0.336)	(0.0099)
Gender composition (Base: Only Males)		
Mixed	-6.088	0.411
	(9.776)	(0.300)
Only Females	-3.938	0.243
	(13.10)	(0.4139)
Difference in Age		-0.0659**
		(0.0324)
lambda	37.7660	
	(40.5390)	
Constant	-29.62	0.625
	(28.52)	(0.444)
Observations	91	116

Trust and Sympathy are centered around the mean
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

AII. 13: Multivariate OLS regression analysis of several variables and interaction effects on qualitative creative performance of the second phase of the exogenous treatments

VARIABLES	Model 1	Model 2	Model 3
Treatment (Base: HOHO)			
PRPR	38.97***	40.65***	65.06***
	(14.22)	(14.10)	(17.37)
PRHO	35.99***	36.59***	86.49***
	(11.67)	(11.93)	(17.75)
HOPR	43.88***	45.54***	74.05***
	(11.42)	(11.04)	(14.29)
Average Ability	0.963	0.978	1.630***
	(0.590)	(0.604)	(0.556)
Qualitative Performance in Phase 1	0.979***	0.939***	0.799***
	(0.127)	(0.134)	(0.140)
Sympathy	-28.25		
	(17.60)		
PRPR x Sympathy	24.69		
	(22.30)		
PRHO x Sympathy	62.92***		
	(22.76)		
HOPR x Sympathy	52.19**		
	(25.83)		
Trust		5.181	
		(13.33)	
PRPR x Trust		3.888	
		(19.91)	
PRHO x Trust		3.089	
		(18.37)	
HOPR x Trust		-1.505	
		(14.59)	
Gender composition (Base: Two Males)			
Mixed			23.03*
			(12.08)
Only Females			47.95*
			(24.29)
PRPR x Mixed			-17.03
			(23.42)
PRPR x Only Females			-62.13
			(50.67)
PRHO x Mixed			-67.21***
			(21.03)
PRHO x Only Females			-76.15*
			(44.39)
HOPR x Mixed			-16.43
			(26.00)
HOPR x Only Females			-61.29**
			(26.37)
Constant	-12.13	-7.639	-30.05*
	(14.40)	(15.24)	(17.20)
Observations	91	91	91
Adj. R-squared	0.618	0.597	0.612

Trust and Sympathy centered around the mean,
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

AII. 14: Different measures of creativity and treatments effects of the exogeneous treatments

VARIABLES	Quantity (OLS)	Quality (OLS)	Rare (Poisson)	Very Rare (Poisson)	Flexibility (Poisson)
Treatment (Base: HOHO)					
PRPR	4.489	43.18***	0.175	0.393***	0.0104
	(6.138)	(13.78)	(0.166)	(0.129)	(0.0459)
PRHO	2.991	41.15***	0.0710	0.361***	0.000285
	(5.315)	(11.89)	(0.153)	(0.118)	(0.0529)
HOPR	4.489	45.24***	0.0270	0.406***	0.0374
	(6.138)	(11.59)	(0.159)	(0.109)	(0.0434)
Average Ability	1.094**	0.916	0.0189***	0.00492	0.00164
	(0.517)	(0.632)	(0.00716)	(0.00533)	(0.00251)
Qualitative Performance in Phase 1	0.816***	0.887***	0.00381***	0.00710***	0.00283***
	(0.106)	(0.139)	(0.00126)	(0.000922)	(0.000441)
Sympathy	-0.482	-0.679	-0.157	-0.0488	0.0101
	(5.208)	(10.16)	(0.169)	(0.0941)	(0.0385)
Trust	3.515	5.041	0.0712	0.0709	0.00952
	(3.134)	(5.488)	(0.142)	(0.0579)	(0.0252)
Difference in Ability	0.105	0.593*	0.00243	0.00398	0.000675
	(0.176)	(0.340)	(0.00413)	(0.00279)	(0.00119)
Gender composition (Base: Only Males)					
Mixed	-2.728	-6.950	-0.00934	-0.0867	0.0300
	(4.980)	(10.25)	(0.129)	(0.0851)	(0.0406)
Only Females	2.437	-4.731	0.182	-0.0271	0.0216
	(8.037)	(16.78)	(0.162)	(0.153)	(0.0633)
Constant	4.065	-8.410	0.693***	1.807***	3.486***
	(8.492)	(18.84)	(0.213)	(0.148)	(0.0668)
Observations	91	91	91	91	91

Notes: Trust and Sympathy are centered around the mean. Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The second column shows for reasons of comparison again the model of Table 14 regarding qualitative performance, which consists of rare and very rare answers and the measure of flexibility. Quantity and Quality show results of OLS regression with robust standard errors whereas Rare, Very Rare and Flexibility show Poisson-regression.

AII. 15: Instructions

The instructions were originally written and used in German.

Instructions – Online

Welcome-Page

Dear participants,

thank you for taking time to participate in the experiment. In the following, the procedure of the experiment and the task will be explained to you in more detail. Please switch your camera on and the sound to "mute".

Please turn off your cell phones now, listen carefully and do not use any other browser window during the experiment than the one in which the experiment is open.

Session 1: The experiment consists of 2 phases. The first phase consists of one task, the second phase consists of 4 tasks that are identically structured. You have 8 minutes for each task. After these 8 minutes, you will be automatically redirected to the next task and time will start again. You will work on the tasks in break-out rooms. In the tasks, you will have to find as many alternative and creative usage options as possible for a given item. You will get an example of such a task explained right away.

Session 2: This time the experiment consists only of the second phase of last week's session. Nevertheless, the task and scoring will be explained again below so that there are no misunderstandings.

The group composition does not change.

Now please click on "Next" and confirm on the following pages that you agree to the privacy policy and the conditions of participation.

Privacy + Conditions of Participation

Session 1: Before starting with the instructions, please fill the following short questionnaire.

Questionnaire A (Socio-demographics)

You can start the experiment later on instruction. You will now be reading a sample task and the scoring scheme will be explained.

Example task - lunch box: Please include as many different and unusual uses for a lunch box as possible. You are not limited to a single unit or a certain size of the object. You can also "edit" the items or add things to them. The conventional use of the object does NOT count as a valid answer. For example, for the lunch box, this would be using the box as a storage device/container for food. Please briefly explain their uses if they are not very obvious and do not just tell us the "final products". For example, for the item "newspaper" please do NOT just write "mirror" without explaining what exactly you mean by that. We cannot evaluate such an answer!

Please enter your answers in the space below the task, separating them with a paragraph.

Evaluation & Payment: We will award points to each of your alternative uses. We will award one point for each correct or feasible and therefore valid answer, two points if the answer is rare, and four points if it is particularly rare. Rarity is measured by the frequency with which exactly this answer was mentioned by others. Also, you will receive only one point for each term if more than one of your named terms have the same use. For example, for the three uses "weapon", "self-defense", and "defense" you will receive only one point instead of 3, because these uses are too similar. Also, there is one point for each category found for an object. For example, the answers "chair" and "stool" fall into the same category "furniture", while " badminton" falls into the other category "sports" and thus generates an additional point for a new category. For these 3 answers, there would then be 3 points for number, 2 points for the two categories, and possibly points for rarity. So, in summary, you can get points for the number of valid uses, the number of different categories, and the rarity of the proposed uses.

One point corresponds to a reward of 8 cents. Your achieved points from each round will be added up afterwards and the corresponding amount will be paid to you.

Session 1: If you also show up for the second session, you will receive an additional fixed payment of €4 after the second session in addition to the performance-based pay.

Session 2: In addition to the performance-based compensation explained above, the compensation after this session also consists of a fixed payment of €4.

After the experiment you will be redirected to a payout page. Please note that the evaluation of your performance takes some time and the transfers may also take longer. A payout will therefore be made in one week at the earliest.

ATTENTION: If you leave or end the experiment prematurely yourself, you automatically lose any claim to compensation!

If you encounter any problems, please use the chat icon on your screen.

Session 1: If there are any problems, please indicate immediately via the Chat-Symbol in the lower right corner or via the private chat option in Zoom, otherwise you will lose any claim to remuneration! The use of copy-paste from the Internet or other sources is strictly prohibited and will result in exclusion and loss of remuneration.

Now go to the break-out rooms and click on "Next" to start the first phase. Please leave your break-out room again afterwards!

Session 2: If there are any problems please report immediately via the chat icon on your screen or via the private chat option of Zoom, otherwise you will lose any claim to compensation! If a team member drops out of the experiment, they will automatically lose any claim to compensation and the remaining member will also not be able to continue participating. Please try to avoid this at all costs!

Now click on "Continue"

**Ability Checker (first, individual UUT) in the first session*.*

Session 1: You will now work with another participant on the second phase in a break-out room. The following tasks are identical to the first task. The other participant with the same ID will be the other team member from you. The ID will be assigned to you in the next step.

Your combined performance now determines your payout from the next 4 tasks.

During the processing, you are welcome to share your screen with your other team member. At the end of the processing time, please fill out another questionnaire individually.

Since you are working on the task as a team and will also be graded as a team from now on, I ask you to agree on who will enter the answer choices in the text box. The other person who does not work on the text field, please enter the ID of the team in his or her text field, so that we can clearly assign you.

If there are problems regarding the internet connection or your other team member does not participate in the experiment anymore or does not appear in their break-out room at all, please report immediately via the chat icon on your screen or via the private chat option of Zoom, otherwise you will lose any claim to compensation! If a team member breaks out of the experiment, they will automatically lose any claim to compensation and the remaining member will also not be able to continue participating. Please try to avoid this at all costs!

At the end of the processing time, please leave the break-out rooms, switch to "mute" and fill out one more questionnaire.

Now click on Next.

If you now click on "Next" again, you will see a page on which your ID is specified. Please use this as your zoom name now.

Session 2: Please click on "Next" again and change your zoom name to the ID from last week at the latest. You can also find this again in the e-mail.

Grouping

Please move to the break-out rooms now and, once you have coordinated with your other team member, click "Next" to begin the experiment. Please agree on who will write the answer choices in the text box. The other person who does not edit the text box, please enter the team ID in his or her text box so that we can match you clearly.

If you still have questions, please contact us now using the chat icon! Otherwise, please go to the break-out rooms now and start the next tasks after coordinating with the other team member by clicking on "Next".

If your other team member does not appear, please report immediately via chat. You will receive all other instructions on your screen. You will be automatically directed to the following tasks of the experiment, after the timer has expired.

After the processing time has expired, please leave the break-out rooms, switch to "mute" and answer the following questions alone or indicate how the statements apply to you.

Tasks + Questionnaire B+ Payout page

Session 1: The experiment is now finished. Please remember to come to the second session in exactly one week at the same time. You will only receive payment if you participate again! You can now leave the Zoom room. Have a good day.

Session 2: The experiment is now finished. Thank you for your participation. You can now leave the Zoom room. Have a good day.

Instructions – Presence

Welcome-Page

Dear participants,

thank you for taking time to participate in the experiment. In the following, the procedure of the experiment and the task will be explained to you in more detail. Please click on “Next”.

Please turn off your cell phones now, listen carefully and do not use any other browser window during the experiment than the one in which the experiment is open.

Session 1: The experiment consists of 2 phases. The first phase consists of one task, the second phase consists of 4 tasks that are identically structured. You have 8 minutes for each task. After these 8 minutes, you will be automatically redirected to the next task and time will start again. In the tasks, you will have to find as many alternative and creative usage options as possible for a given item. You will get an example of such a task explained right away.

Session 2: This time the experiment consists only of the second phase of last week's session. Nevertheless, the task and scoring will be explained again below so that there are no misunderstandings.

The group composition does not change.

Now please click on "Next" and confirm on the following pages that you agree to the privacy policy and the conditions of participation.

Privacy + Conditions of Participation

Session 1: Before starting with the instructions, please fill the following short questionnaire.

Questionnaire A (Socio-demographics)

You can start the experiment later on instruction. You will now be reading a sample task and the scoring scheme will be explained.

Example task - lunch box: Please include as many different and unusual uses for a lunch box as possible. You are not limited to a single unit or a certain size of the object. You can also "edit" the items or add things to them. The conventional use of the object does NOT count as a valid answer. For example, for the lunch box, this would be using the box as a storage device/container for food. Please briefly explain their uses if they are not very obvious and do not just tell us the "final products". For example, for the item "newspaper" please do NOT just write "mirror" without explaining what exactly you mean by that. We cannot evaluate such an answer!

Please enter your answers in the space below the task, separating them with a paragraph.

Evaluation & Payment: We will award points to each of your alternative uses. We will award one point for each correct or feasible and therefore valid answer, two points if the answer is rare, and four points if it is particularly rare. Rarity is measured by the frequency with which exactly this answer was mentioned by others. Also, you will receive only one point for each term if more than one of your named terms have the same use. For example, for the three uses "weapon", "self-defense", and "defense" you will receive only one point instead of 3, because these uses are too similar. Also, there is one point for each category found for an object. For example, the answers "chair" and "stool" fall into the same category "furniture", while " badminton" falls into the other category "sports" and thus generates an additional point for a new category. For these 3 answers, there would then be 3 points for number, 2 points for the two

categories, and possibly points for rarity. So, in summary, you can get points for the number of valid uses, the number of different categories, and the rarity of the proposed uses.

One point corresponds to a reward of 8 cents. Your achieved points from each round will be added up afterwards and the corresponding amount will be paid to you.

Session 1: If you also show up for the second session, you will receive an additional fixed payment of €4 after the second session in addition to the performance-based pay.

Session 2: In addition to the performance-based compensation explained above, the compensation after this session also consists of a fixed payment of €4.

After the experiment you will be redirected to a payout page. Please note that the evaluation of your performance takes some time and the transfers may also take longer. A payout will therefore be made in one week at the earliest.

ATTENTION: If you leave or end the experiment prematurely yourself, you automatically lose any claim to compensation!

If you encounter any problems, please use the chat icon on your screen.

Session 1: If there are any problems, please indicate immediately, otherwise you will lose any claim to remuneration! The use of copy-paste from the Internet or other sources is strictly prohibited and will result in exclusion and loss of remuneration.

Now click on "Next" to start to work on the first phase.

Session 2: If there are any problems please report immediately, otherwise you will lose any claim to compensation! If a team member drops out of the experiment, they will automatically lose any claim to compensation and the remaining member will also not be able to continue participating. Please try to avoid this at all costs!

Now click on "Continue”

**Ability Checker (first, individual UUT) in the first session*.*

Session 1: You will now work with another participant on the second phase. The following tasks are identical to the first task. The other participant with the same ID will be the other team member from you. The ID will be assigned to you in the next step. Your team partner is the person sitting closest to you in the same row.

Your combined performance now determines your payout from the next 4 tasks.

At the end of the processing time, please fill out another questionnaire individually.

Since you are working on the task as a team and will also be graded as a team from now on, I ask you to agree on who will enter the answer choices in the text box. The other person who does not work on the text field, please enter the ID of the team in his or her text field, so that we can clearly assign you.

If a team member breaks out of the experiment, they will automatically lose any claim to compensation and the remaining member will also not be able to continue participating. Please try to avoid this at all costs!

At the end of the processing time, the groupwork is over. Afterwards, please fill out one more questionnaire individually.

Now click on Next.

If you now click on "Next" again, you will see a page on which your ID is specified. Please remember this number.

Session 2: Please click on "Next" again and remember your ID from last week at the latest. You can also find this again in the e-mail.

Grouping

Please click “Next” to begin the experiment, once you have coordinated with your other team member. Please agree on who will write the answer choices in the text box. The other person who does not edit the text box, please enter the team ID in his or her text box so that we can match you clearly.

If you still have questions, please contact us now! Otherwise, please sit next to each other now and start the next tasks after coordinating with the other team member by clicking on "Next".

You will receive all other instructions on your screen. You will be automatically directed to the following tasks of the experiment, after the timer has expired.

After the processing time has expired, please separate again, move to your own desks and answer the following questions alone or indicate how the statements apply to you.

Tasks + Questionnaire B + Payout page

Session 1: The experiment is now finished. Please remember to come to the second session in exactly one week at the same time. You will only receive payment if you participate again! Have a good day.

Session 2: The experiment is now finished. Thank you for your participation. Have a good day.

AII. 16: Questionnaires

The Questionnaires were originally written and used in German.

Questionnaire – A

Which gender do you feel you belong to?

Male / Female / Other / do not want to state

How old are you?

XXX years old

What is your highest educational qualification?

High school diploma / Vocational training / Technical college entrance qualification / University entrance qualification (Abitur) / Bachelor's degree / Master's degree

What is your nationality?

EU / Non-EU

Which field of study are you studying/ In which professional field are you working?

Economics; Natural sciences; Language and cultural studies; Teaching; Agricultural and Forestry Sciences; Engineering; Social sciences; Art, Music, Design; Health sciences; Public administration

Questionnaire – B

Scale: 1-5 (unless otherwise specified); Strongly disagree; Disagree somewhat; Neither agree nor disagree; Agree somewhat; Strongly agree

Please indicate the extent to which you agree with the statements below regarding teamwork:

My other team member is sympathetic.

I would lend my other team member personal items of mine.

I would lend money to my other team member.

I would trust my other team member with my apartment keys.

I trust my team member.

Session 1: Below you will find a number of characteristics that might apply to you. For example, would you say about yourself that you enjoy spending time with other people? For each of the following statements, please indicate the extent to which you agree.

I am someone who tends to be quiet.

I am someone who is compassionate, has a soft heart.

I am someone who tends to be disorganized.

I am someone who worries a lot.

I am someone who is fascinated by art, music, or literature.

I am someone who is dominant, acts as a leader.

I am someone who is sometimes rude to others.

I am someone who has difficulty getting started on tasks.

I am someone who tends to feel depressed, blue.

I am someone who has little interest in abstract ideas.

I am someone who is full of energy.

I am someone who assumes the best about people.

I am someone who is reliable, can always be counted on.

I am someone who is emotionally stable, not easily upset.

I am someone who is original, comes up with new ideas.

Do you already have experience with mobile working?

Yes / No

Are you currently in an employment relationship?

Yes / No/ Self-employed

If yes or self-employed: In which form of work do you work?

Presence / Fully remote / Hybrid

Session 1 and Session 2:

I am satisfied with my place of work where I conducted the experiment.

I prefer mobile working to working in presence.

I prefer to work in a team rather than alone.

I enjoyed the task.

I found the task to be exhausting to work on.

PART IV: PAY-FOR-CREATIVITY: FOSTERING CREATIVITY IN TEAMS THROUGH MONETARY INCENTIVES

Abstract: As the relevance of non-routine work and creative teams is on the rise, identifying optimal incentive schemes for creative performance of such teams is of great importance. Existing theoretical considerations and empirical findings disagree on the effects of external monetary incentives or incentive intensity on creativity and lack the connection to creative teams. Therefore, this study systematically explores the effect of PfP incentive systems, i.e., piece-rate pay and between-team tournaments and their intensity on creative performance of dyadic teams. The results of a large-scale laboratory experiment show that there is a beneficial effect of most performance-dependent incentives on creative performance working on a divergent thinking task in comparison to fixed pay. Moreover, it is shown that it is not the intensity of an incentive that matters most for increasing creative performance, but the presence of a tournament between teams. Hence, this study provides valuable insights for organizational practice by demonstrating that monetary incentives can effectively increase quantity and quality of creative performance in dyadic teams working on a divergent thinking task as starting point for innovations.

1 Introduction

Non-routine work and creativity as one specific form have emerged as critical drivers of success in modern organizations (Di Battista et al., 2023). Creativity is the starting point for organizational innovation (Amabile & Pratt, 2016), leveraging problem-solving capabilities, and thus competitive advantage (Anderson et al., 2014; West & Sacramento, 2023). Creative performance is often enforced by innovation tournaments and competitions centered around the most creative idea (Morgan & Wang, 2010; Terwiesch & Ulrich, 2009). As the importance of non-routine work like creativity continues to grow, identifying optimal incentive mechanisms to enhance creative performance has become a pressing challenge. This is particularly relevant in team-based environments, which are increasingly central to creative processes (Kozlowski & Bell, 2013). Incentive schemes and particularly tournaments are a well-studied field in economics. The theorem of a generally beneficial effect of PfP on motivation and output (Bull et al., 1987; Lazear, 2000; Lazear & Rosen, 1981) is present in most organizations focusing on routine tasks. However, the degree of uncertainty associated with producing creative output, whether due to own performance or evaluation methods, leads to questioning the given theoretical and empirical findings from routine tasks in the field of creativity (Bradler et al., 2019). Following from Principal-Agent theory and the potential conflict between sufficiently high incentives and risk aversion linked to variable payments, the effect of PfP on creative performance is not clear (Gerhart & Fang, 2015; Jensen & Meckling, 1976; Kim, 2025). Furthermore, the intensity of such PfP incentives need to be considered due to this interplay between incentive and risk effects.

Despite fundamental existing research in the field of creativity and incentive structures since Amabile's first version of the componential model of creativity (Amabile, 1988), the literature remains divided on the issue of how to best structure incentives for creativity and is underrepresented in the field of creative teams in particular (Malik et al., 2015). Existing research shed some light on the relations between different PfP systems for individuals (e.g., Bradler et al., 2019; Charness & Grieco, 2019), incentivizing teams (C. X. Chen et al., 2012; Englmaier et al., 2024b), or tournaments in teams (Englmaier et al., 2024a) in non-routine contexts, providing ambiguous results. The general debate about the effect of extrinsic incentives on creativity (e.g., Amabile & Pratt, 2016; Deci & Ryan, 1985; Q. Zhang et al., 2024) is complicated by variations in experimental designs, including group composition, tasks, and especially incentive intensity, which is determined by the proportion of salary that is performance-related (Englmaier et al.,

2024b; Kim, 2025). This leads to mixed and sometimes contradictory findings. Therefore, existing research lacks overall interaction between those aspects, i.e., the direct comparison of different incentive systems with varying intensity in the field of creative teams.

This is where this essay adds to the literature by providing a clean laboratory analysis on the effects of different monetary incentives on creativity in teams. This study explores how (absolute and relative) PfP incentive systems and their intensity affect creative performance of teams in comparison to fixed payments. I vary the intensity of PfP by varying payout relevance of performance-based pay versus fixed pay. Moreover, a piece-rate scheme is compared to a between-team tournament setting to account for the pervasive competitive bidding structures. This enables insights about whether and how creative performance of teams can be promoted by monetary incentives. The study bridges gaps in prior research by integrating insights from creativity and incentive literature, providing a comprehensive analysis of PfP systems in the context of creative teams.

To further shed light on the incentive-creativity relation in teams, focusing on incentive intensity, the study evolves around two research questions:

1. *Is the level of creative performance of teams higher under pay-for-performance than pure fixed pay?*
2. *Does a pure pay-for-performance system lead to higher levels of creative performance than a system including also a fixed pay component?*

These aspects are critical not only for advancing theoretical understanding of incentive mechanisms in creative fields, but also for informing organizational practices. By addressing these dimensions, this study aims to provide actionable insights into how organizations can strategically leverage incentives to unlock the creative potential of their workforce.

2 Theoretical background and literature review

In this study, the focus is on the early stage of divergent thinking as part of the creative idea generation process.[46] Creativity is defined as the production of novel and useful ideas by individuals or small groups working together (Amabile, 1988). Furthermore, the focus lies on creativity levels of teams as the combinatory theory (Dutcher & Rodet, 2022; Feinstein, 2011) as well as the additive model of creative thinking (Kurtzberg, 1998; Kurtzberg & Amabile, 2001; Yuan et al., 2022) highlight the creative potential of teams in comparison to individuals due to

[46] The study focuses on divergent thinking as it is the starting point of many innovative processes, whether in product, service, or research agenda development (Amabile & Pratt, 2016).

active exchange of ideas, triggering a new thought process in the other team members. Practical insights confirm these theories (Kozlowski & Bell, 2013). A recent laboratory study revealed higher creative performance of teams compared to individuals under piece-rate incentives solving a divergent thinking task (Grund et al., 2026).[47]

Research on incentive structures is well established in the economic field, mainly for routine tasks. The underlying consensus regarding the theory of PfP systems is that optimal effort levels of individuals can be reached by well-structured absolute (i.e., piece-rate) and relative (i.e., tournament) systems (Bull et al., 1987; Lazear, 2000; Nalbantian & Schotter, 1997). However, the effect of PfP on teams, as opposed to individuals, is more complex, as the dimensions of intra-group behavior, cohesion, identity, and motivation are added (Majerczyk et al., 2019; Wittchen et al., 2013).

Classical economic tournament theory indicates that tournaments between teams can promote performance through increased intragroup cooperation (Abbink et al., 2010; Englmaier et al., 2024a; Nalbantian & Schotter, 1997) and intergroup competition but still risk free-riding behavior, especially under egalitarian sharing rules and hidden action (Alchian & Demsetz, 2009; Gunnthorsdottir & Rappoport, 2006; Majerczyk et al., 2019). In the context of creative teamwork, free-riding is a less common issue as the found superiority of a team in comparison to individuals stems from the exchange of ideas and experiences (Feinstein, 2011). Thus, the approach of the presented study is on the design of incentive systems which further increase the beneficial aspects of creative teams, for example through the implementation of tournaments. The driving mechanism of the success of tournaments is the prize difference between loser and winner (Lazear & Rosen, 1981). Therefore, the intensity of a tournament incentive must be considered when implementing such (Harbring & Irlenbusch, 2003; Krishna & Morgan, 1998).

The competitive nature of tournaments can positively impact intrinsic motivation and group performance by balancing rivalry between groups, ultimately stimulating creativity (Tauer & Harackiewicz, 2004). Additionally, tournament incentives can boost team performance by fostering group cohesion (Baer et al., 2010; C. X. Chen et al., 2012; Frey et al., 2013). Bonus payments, as an attenuated form of pure tournament incentives, show positive effects on individual creative performance, especially when perceived as meaningful (Bradler et al., 2019; Englmaier et al., 2024a; Gross, 2020). Similarly, tournaments as form of monetary incentives

[47] The present study focuses on team creativity and does not address the role of incentives for individuals or the general comparison of teams and individuals. For related work on individual vs. team creativity see for example: Dutcher & Rodet (2022), Gneezy et al. (2022), and Grund et al. (2026), and for incentives in an individual context see for example: Bradler et al. (2019), Byron & Khazanchi (2012), Charness & Grieco (2019), Erat & Gneezy (2016) and Laske et al. (2024).

are found to enhance individual creative performance in closed, but not open-ended tasks (Charness & Grieco, 2014). Monetary components in tournaments can positively impact creativity of teams without inducing a crowding-out effect, but only under carefully structured environments, for example determined by leadership and experience of teams (Englmaier et al., 2024b). However, while tournaments can effectively boost creativity, some studies show that their benefits are not unlimited: Malik et al. (2015) suggest that the effectiveness of monetary incentives can be influenced by individual perceptions of reward importance. Other studies describe that excessively intense competition or overly high rewards may lead to detrimental effects, such as reduced creative effort or lower overall performance (Ariely et al., 2009; Gross, 2020). Erat and Gneezy (2016) even claim that monetary incentives fail to enhance creativity overall. These conflicting findings suggest that the intensity of the incentive is a key determinant for the effectiveness of PfP incentives on creative performance of teams, and it is precisely this factor that is directly manipulated in the present study.

Classic SDT (Ryan & Deci, 2000) originally predicted negative effects of extrinsic incentives on creative output due to a crowding-out effect. Nonetheless, a revised version of the theory, motivated by contradicting research results, posits that extrinsic rewards perceived as supportive can enhance intrinsic motivation and creativity (Amabile & Pratt, 2016; Fischer et al., 2019; Gerhart & Fang, 2015). A tournament might reinforce existing intrinsic motivation if winning the tournament is related to the need for competence and effectiveness which in turn increase intrinsic motivation as described in SDT. In a similar vein, the learned industriousness theory (Eisenberger, 1992) as well as expectancy theory (Jenkins et al., 1998; Vroom, 1964) suggest that rewards reinforce positive behavior and can enhance performance when incentives provide informational value about expectations (Byron & Khazanchi, 2012; Eisenberger & Shanock, 2003).

Empirical research shows mixed results regarding the relation between monetary incentives and creativity, depending on contextual and experimental differences. Some studies highlight positive effects of rewards on creativity, particularly when incentives target both quantity and quality of creative performance (Kachelmeier et al., 2008; Laske et al., 2024). Building on this, an experimental study by Kachelmeier et al. (2024) further suggests that creativity-contingent rewards positively affect divergent thinking tasks with more than one solution. A recent online experimental study on individual creative performance implementing rewards depending on relative performance shows that individual level creativity is highest under strong incentive intensity (Kim, 2025), though neither a winner-takes-all scenario nor a pure piece-rate scheme are considered. A meta-analysis conducted by Kim et al. (2022) further hints towards beneficial

incentive effects for tasks perceived as interesting, with effects robust to task motivation, autonomy, and incentive intensity. Notably, however, incentive intensity is not manipulated in their study, pointing to an important limitation.

Despite these insights, research into incentive mechanisms in the context of creative teams remains scarce. Existing studies often focus on individual performance (e.g., Charness & Grieco, 2019; Kim, 2025, Laske et al., 2024) leaving the dynamics of team-based creative tasks unexplored. This gap is further compounded by inconsistent incentive structures across studies, including tournament prize distribution, which plays a critical role in shaping outcomes. In particular, incentive intensity has rarely been directly manipulated, making it difficult to draw conclusions about an effective incentive design for creative performance of teams.

Standard rank-order tournament theory typically does not assume a winner-takes-all scenario (Lazear & Rosen, 1981). However, this format is common in innovation competitions and may have significant implications for creativity (Ales et al., 2017). Winner-takes-all tournament systems have been widely adopted in innovation contexts due to their cost-effectiveness and their ability to drive radical innovation (Bradler et al., 2019; Morgan & Wang, 2010). If eliminating positive non-monetary compensations, such as reputation effects, these systems focus entirely on winning, thus, maximizing competition and effort among participants. This makes winner-take-all tournaments particularly suitable when high-quality and diverse solutions are needed, especially in scenarios with two or three competitors (Ales et al., 2017; Krishna & Morgan, 1998).

Therefore, a more nuanced understanding of diverse incentive mechanisms and prize structures, tailored to team dynamics and creative contexts, is essential for advancing this field. Consequently, the present study explicitly takes this scenario into account and aims to make a valuable contribution to existing research by providing a clean analysis of different incentive types and intensities, even the extreme but practically relevant ones. In the following, I will derive hypotheses for the specific relationships between absolute and relative PfP incentives and creative performance of teams.

3 Hypotheses

Following Principal-Agent theory, performance-dependent payments like piece-rates can have a positive incentive effect (Lazear, 2000), whereas the optimal combination of variable and fixed payment depends on the risk attitude of the agent. Also, performance-dependent payment systems do not only provide monetary incentives but also serve an informational role: they

signal which aspects of performance, such as quantity and quality, are valued and therefore linked to pay, which is also the case for creativity (Byron & Khazanchi, 2012; Gerhart & Fang, 2015). Unlike flat payments, which lack this signaling function, PfP structures clarify expectations and incentivize effort aligned with creative objectives (Laske et al., 2024). For example, winner-takes-all tournaments with the highest possible prize difference, given a monetary upper bound, or increasing piece-rate payment with diminishing fixed payment components can have different informational value, as they vary the relevance of those performance dimensions by changing their relevance for determining final payout. Compared to PfP systems with base payments, winner-takes-all and pure piece-rate systems inherently sharpen the focus on the payout-relevant performance dimension by setting a clear, singular goal for payoff-maximizing participants: achieving the highest possible level of creativity. Although maximizing the prize difference in tournaments can heighten the risks associated with competitive dynamics, it simultaneously reinforces the signaling function of the incentive, motivating teams to align their efforts toward creative excellence. Moreover, winner-takes-all systems have been shown to foster internal team cooperation and encourage creative exchanges by emphasizing collective goals without the dilution of effort caused by guaranteed base payments (Englmaier et al., 2024a).

Tournaments are shown to not only motivate participants to achieve high levels of creative performance but also to enhance intrinsic motivation through identification with the goal of outperforming competitors, aligning the implementation of an external incentive with the demand for self-determination (Amabile & Pratt, 2016). This effect is particularly pronounced when winner prizes are perceived as meaningful and valuable (Malik et al., 2015). Moreover, field studies showed that monetary tournaments in creative teams foster cooperation, task focus, and productive exchanges among group members, all of which are critical drivers of creative idea generation (Englmaier et al., 2024a, 2024b). Intermediate forms of tournaments, such as those featuring loser prizes or a bonus payment in addition to a fixed salary, have proven to be effective in practice, especially for routine tasks, in comparison to fixed pay. Such mechanisms strike a balance between extrinsic and intrinsic motivators, creating an environment that promotes both individual and team creativity (Bradler et al., 2019).

Based on Principal-Agent Theory and the signaling function of PfP incentives the first two hypotheses of this study state:

> *H1: Pay-for-performance incentive systems lead to a higher level of creative performance in teams than pure fixed payments.*

H2: Creative performance levels of teams increase with the wage's fraction of incentive-based pay.

The first two hypotheses cover the comparison of PfP with pure fixed payment as well as the effect of increasing the fraction of payment which is based on performance. The remaining aspect to clarify is whether absolute and relative PfP systems alter creative performance levels of teams differently. Both induce incentives to increase effort by providing information about the desired output. Theoretical models on the effectiveness of relative payment schemes show that if general risk and high variance in output are present for each participant, relative incentives dominate absolute schemes, as such schemes account for riskiness of tasks as well as the relevance of luck (Lazear & Rosen, 1981).[48] In a chosen-effort experiment with teams, individuals provided higher effort under tournament incentives than absolute target-based schemes and group incentives due to the added dimension of competition (Nalbantian & Schotter, 1997). Winning such a competition can be part of self-determination through the need for competence and effectiveness, thus, affecting intrinsic motivation positively (Amabile & Pratt, 2016).

Additionally, inter-team tournament incentives are not only output-focused but also induce group cohesion and identification in comparison to piece-rates (C. X. Chen et al., 2012). Group cohesion describes the individual attraction to a group, increasing orientation towards the goals of the group (Pescosolido & Saavedra, 2012) and is found to correlate with efficacy and performance (Gächter et al., 2025; Ramzaninezhad et al., 2009). The level of organizational identity can beneficially influence team members' effort, especially in a between-group tournament setting (Ai et al, 2023). Group cohesion is found to be a positive and significant mediator when introducing competition between creative groups (Baer et al., 2014).

Following tournament theory, SDT, and the concept of group cohesion and identity, hypothesis 3 states:

H3: Tournament incentives lead to higher creative performance levels of teams than comparable piece-rate incentive systems.

4 Experimental design

To answer the research questions and test the hypotheses, an experiment in the lab of the RWTH Aachen University has been conducted using oTree (D. Chen et al., 2016).[49]

[48] Performing a creative task is connected to perceived risk as participants are unsure whether the generated output is perceived as creative by others (Dewett, 2006) or how able they are in general to solve creative tasks (Kim, 2025).

[49] The research questions, hypotheses, and experimental design are preregistered at *Aspredicted.org* #214372.

The experiment contains an individual familiarization task at the beginning, a team-task, an incentive or treatment phase, and a questionnaire. In each round the dependent variable, creative performance, is measured via the so-called Droodle-Task. A Droodle is an abstract picture without color and no clear objects shown. The task of participants is to find as many creative titles describing a given Droodle as possible and type these into a text box (Ali et al., 2019; Kahn et al., 2008; Kaivola-Bregenhøj, 2001). Thereby, participants are not restricted in the perspective they take as they are able to rotate the picture on their screen. For example, creative solutions for the Droodle in Figure 1 (Part I, section 1.3) can be "Blender rotors" or "Four elephants sniffing at an orange". For each Droodle, participants have four minutes to enter their creative suggestions for interpreting the picture. Figures AIII. 1 to AIII. 5 in Appendix AIII show the Droodles used and Figure AIII. 6 in Appendix AIII shows an example screen.

A between-subject design is implemented to capture the effects of teams, different external monetary incentives, and their intensity on creative performance. Table 18 gives an overview of the five treatments and the procedure of the experiment. Each treatment is assessed in separate sessions. For each treatment, approximately 30 observations have been targeted, whereas an observation is defined on the team level, thus 30 observations equal to 60 participants. An average payment of €12 is implemented, following the laboratory regulations. Payouts are described in the experimental currency *Taler* and are later on transferred into Euro.

Table 18: Structure and Treatments

Treatment / Timing	Fix	Plow	Tlow	Pmax	Tmax
	Instructions				
Individual	Individually – no Incentive				
Team-Check	Team – no Incentive				
Incentive Phase (3 Droodle)	Team x Fix	Team x Piece with fixed component	Team x Tournament with positive loser prize	Team x Pure Piece	Team x Pure Tournament
	Questionnaire				

As depicted in Table 18, at the beginning, instructions (see AIII. 25 in Appendix III) are presented via screen and also read aloud by the experimenter. The instructions cover an example Droodle, and it is highlighted that participants should not only provide as many solutions as

possible but also be as creative as they could.[50] Furthermore, participants are informed about the fixed payment of €4 for participating as well as working on the first two tasks and that information about additional payment will be provided at later stages. Moreover, the way of evaluating creativity is presented: All answers given for one Droodle are evaluated by a jury of four raters who evaluate the validity and creativity of answers. This corresponds to the method of subjective evaluation of creativity, following the consensual assessment technique (Amabile, 1982).[51] An answer is marked as valid if three of four jurors understand and recognize the answer in the Droodle and if the answer is not too similar to one given before (for example: *big* rotors; *small* rotors). A valid answer receives 5 points. Then, each juror grants between 0 to 3 additional points for creativity, whereas 0 points describe an uncreative and obvious answer, and 3 points are assigned for very creative answers. Therefore, a valid answer can receive 5-17 performance points and quantity and quality of ideas are incentivized (Byron & Khazanchi, 2012; Kachelmeier et al., 2008).[52]

Thereafter, participants get to work on the first Droodle individually. This first task functions as familiarization task as well as incubation period of the creative process (Kachelmeier et al., 2019). This is followed by a team task, establishing teamwork. The teams are formed by two participants (dyadic teams).[53] Working in dyadic teams, participants have to coordinate on who is typing the answers in the text boxes, ensuring direct exchange and teamwork.

Following the team-check, teams proceed to the incentive phase, maintaining their original composition. Five treatments with distinct incentive structures are implemented: (1) *Fix* includes a flat payment of €12 for working on the three Droodles of the incentive phase, which is later equally divided between team members; (2) *Plow* combines a €6 fixed payment with €0.04 per performance point on the team level; (3) *Tlow* features the same base pay through a loser prize of €6 and a winner prize of €18, keeping the expected team payments equal across treatments; (4) *Pmax* provides no such fixed base-salary but a piece-rate of €0.08 per performance point; (5) *Tmax* maximizes tournament incentives through increasing the prize difference by

[50] Highlighting the qualitative component of participants' answers, i.e., creativity, is found to increase quality without lowering quantity of answers (Harrington , 1975; Nusbaum et al., 2014) and gives important informational value (Malik et al., 2015).

[51] The implemented evaluation of creativity follows a subjective evaluation approach which is widespread in creative research. It is assessed as reliable measure and does not limit the creativity of answer towards its rarity, like objective measures might do (Silvia et al., 2008; Silvia, 2011).

[52] Due to the fact that the performance evaluation happens ex-post, no performance feedback is given to the participants during the experiment. Therefore, neither demotivation nor effects of fear of negative feedback are applicable here.

[53] During team formation, two participants sitting next to each other were matched and informed about their new team status.

eliminating the loser prize and increasing the winner prize to €24.[54] In all PfP treatments, one randomly selected Droodle of the incentive phase determines the final payout.

After completing the incentive phase, participants have to answer several questionnaires individually, receiving again €2 each.[55] In the post-experimental questionnaire (see AIII. 26 in Appendix III), self-evaluation of performance, intrinsic motivation (Ryan, 1982; Ryan et al., 1983), effort, pressure, and cognitive flexibility (Martin & Rubin, 1995) are measured. Additionally, locus of control (Rotter, 1966), risk attitude (Beierlein et al., 2014), creative inclination (Burk & Wiese, 2018), the Big5 personality traits (BFI-2-XS) (Soto & John, 2017), and further socio-demographic factors like nationality or field of study are assessed. Regarding teamwork, different aspects of cooperation and coordination as well as equal participation and leadership (Englmaier et al., 2024b), cohesion and identification (C. X. Chen et al., 2012) are measured. Finally, for the tournament-treatments, additional aspects such as preferences, competitive orientation (Burk & Wiese, 2018), and self-evaluated chance of winning the tournament are collected. The chosen control items capture individual and team characteristics that have been shown in previous studies to influence creativity and collaborative performance.

5 Data collection, sample and descriptive statistics

The sample consists of 556 subjects, leading to 278 dyadic teams. On average, one session takes about 49 minutes. The average total payment per participant amounts to €12.01.[56] On the team level, the sample consists of 79 pure female and 104 pure male teams, the remaining 95 teams are mixed with respect to their gender composition. The average age of the sample is 22.70 years. Table AIII. 10 in Appendix AIII shows socio-demographic sample characteristics on the individual level. Kruskal-Wallis tests show that the treatments do not differ significantly from each other in those dimensions.

Several performance-related variables serve as dependent variables, whereas I differentiate between performance in the individual familiarization, team-check, as well as the following incentive phase containing the treatment conditions. Firstly, the number of valid answers given is assessed, determining the quantitative performance measure and indicating fluency. The

[54] The number of direct contestants in the tournament treatments is limited to two, in line with the contribution of Fullerton and McAfee (1999) covering research tournaments.
[55] This payment is implemented in accordance with laboratory regulations to ensure a minimum wage of €6 in the case of losing the *Tmax* treatment.
[56] The payout for different treatments varies depending on the incentive scheme implemented. For the piece-rate treatments, the average payment is €11.77 for *Plow* and €12.14 for *Pmax*, ranging between €9.36 and €15.34 in *Plow* and for *Pmax* between €8.40 and €15.72. The payment ranges of the tournament treatments are predetermined by the treatment design (i.e., the fix loser or winner prize). The goal of €12 on average is reached with the implemented incentive structures.

number of mistakes made or answers which are given twice determines the error rate. The measure indicating creativity is the qualitative performance measure, calculated as sum of the points assigned by the jury to a given answer.[57] Additionally, the measure of originality is calculated ex-post and determines the frequency with which a certain answer is given. Related thereto, an answer is labeled as original if it was only given once. The treatments, differing in the implemented payment schemes, are the main independent binary variables. The data gathered through the questionnaires are used as controls and for further analysis.

Using the subjective evaluation method of creative performance requires checking for inter-rater reliability. Using Cohen's Kappa, the general inter-rater reliability combined is 49.78% and ranges between 38.15% (an answer is rated with 2 points) and 78.14% (an answer is rated as invalid) for the different outcomes, i.e., the specific points between 0 and 3 given for a specific answer or its validity. As creativity and understanding of what is creative is subjective (Amabile, 1982; Mohnen & Ostermaier, 2013; Silvia et al., 2008), no higher inter-rater reliability is to be expected, especially when deciding about nuances, i.e., whether an answer is worth 2 or 3 points.

Table 19 shows the average individual creative performance of participants, differentiated along the performance dimensions of a creative task. The total average individual performance in the familiarization Droodle, i.e., the sum of points achieved per valid answer, amounts to 72.31 points (SD=33.05; Min=5; Max=213) in all treatments. The average performance when introducing teamwork in the team-check task amounts to 126.36 points (SD=42.96). No significant difference regarding the individual performance of participants as well as the team ability is found across treatments, using Kruskal-Wallis Tests (p=0.839; p=0.797). The treatments do not differ in most experiment-related controls, like the level of difficulty, effort, cognitive flexibility or whether one team member is acting as a leader. Significant difference (p=0.093) shows for perceived fun during the team phases, which is weakly significant lower for *Tmax* comparing to the other treatments. Additionally, the level of intrinsic motivation is weakly significant higher in *Fix* than in the other treatments (p=0.079) and individuals in *Plow* report significantly lower levels of group identification (p=0.080) (see Table AIII. 11 in Appendix III).[58]

[57] In the literature of creativity, the here used measure of creative quality is sometimes referred to as originality. In the present study, the measure of originality is an ex-post quantitative measure of uniqueness of ideas (being in line with Essays 1 and 2), whereas here used quality describes the perceived creativity of an idea by the jurors.

[58] As the level of intrinsic motivation is the highest under fixed payment, a potential indication for crowding-out of intrinsic motivation through PfP is found. Using pairwise comparisons, significant differences show for *Fix* and *Plow* (p=0.008), *Fix* and *Tmax* (p=0.036), and *Plow* show weakly significant lower levels of intrinsic motivation than *Pmax* (p=0.09). Nonetheless, following results show positive effects of PfP on performance, such that if crowding-out happens, the beneficial effects of PfP outweigh potential losses in motivation.

Table 19: Average individual performance with standard deviation in parentheses, minimum, and maximum

	Individual familiarization			
	Quantity	**Quality**	**Originality**	**Errors**
Fix				
Average	7.824	32.537	0.759	0.861
SD	3.272	18.386	1.093	1.115
Min	3	4	0	0
Max	18	94	5	6
Plow				
Average	7.768	31.705	0.92	0.812
SD	3.176	17.105	1.041	1.087
Min	2	0	0	0
Max	18	80	5	5
Tlow				
Average	7.864	33.955	0.955	0.973
SD	3.423	19.040	1.104	1.443
Min	3	2	0	0
Max	19	100	5	9
Pmax				
Average	7.955	32.727	1.127	0.636
SD	3.039	15.642	1.118	0.965
Min	1	0	0	0
Max	18	78	5	5
Tmax				
Average	8.026	33.293	1.164	1.19
SD	3.589	17.275	1.521	1.817
Min	1	0	0	0
Max	23	98	8	9

Table AIII. 12 in Appendix III shows Pearson correlations between several controls and creative performance on the individual level. Remarkably, upon others, a significant positive correlation between the level of intrinsic motivation and creative performance is found ($r=0.106$, $p=0.013$). Moreover, a significant positive correlation for female (i.e., gender=1) with creative performance during the team ($r=0.112$, $p=0.008$) and incentive phase ($r=0.074$, $p=0.082$) is found. The same holds for both phases for the level of group cohesion ($r=0.086$, $p=0.043$; $r=0.151$, $p<0.001$) and perceiving the group as advantageous ($r=0.096$, $p=0.024$; $r=0.105$, $p=0.013$). Perceived pressure also shows positive correlations ($r=0.0120$, $p=0.005$; $r=0.109$, $p=0.010$). All in all, indicating positive connections between teamwork, intrinsic motivation and incentivizing creativity, whereas the feeling of pressure does not show negative but positive connections with creative performance of teams.

6 Results

The following section presents the results regarding the three hypotheses claimed. The results are calculated and presented on team level. Table 20 shows the average performance measures during the incentive phase of the respective treatments. The performance of the three Droodles in the incentive phase is aggregated in a performance measure for the whole incentive phase,

as some tasks are perceived as more difficult than others, while this effect is rather independent of treatment and stems from the Droodle itself (see Figure 8). Considering the absolute values, no extreme difference appears when comparing *Fix* and *Plow*. Regarding the other three treatments, results hint for higher creative performance when incentive-based payments are implemented, i.e., higher quantity, quality, and originality. It is noteworthy that the SDs are the highest under *Tmax*, indicating an increase in variability of results when implementing a high intensity tournament between teams. Additionally, the average number of errors made is the highest for *Tmax*, which tends to indicate, at least in part, negative effects due to increased variability when implementing a winner-takes-all scenario.

Table 20: Average performance measures in the incentive phase with standard deviations in parentheses

	N	Quantity	Quality	Originality	Errors
Fix	54	36.833	183.907	5.315	3.407
		(12.887)	(66.345)	(3.928)	(3.224)
Plow	56	38.018	187.518	5.857	3.589
		(13.599)	(64.596)	(3.560)	(4.816)
Tlow	55	40.981	209.509	6.509	3.589
		(10.872)	(56.015)	(3.701)	(4.816)
Pmax	55	41.127	210.473	6.964	3.927
		(10.375)	(53.903)	(3.825)	(4.682)
Tmax	58	46.328	234.121	7.897	4.517
		(14.028)	(67.825)	(4.319)	(4.378)

My analysis and hypotheses testing focuses mainly on the qualitive measure of creative performance as main indicator for creativity and innovative ideas. Figure 8 shows the average qualitative performance per treatment over the course of the experiment. In all treatments there is a significant increase (between 176% and 198% depending on the performance measure) when introducing teamwork for the second Droodle, such that a positive effect of teamwork is found. Moreover, there exists no significant difference regarding individual performance and the ability to solve a Droodle as a team between treatments, hinting towards the relevance of the introduced incentives when explaining the differences between treatments in the incentive phase.[59]

[59] For respective figures on average quantitative performance and originality over the course of the experiment refer to Figure AIII. 7 and Figure AIII. 8 in Appendix III. Similarly, both show no significant difference between treatments in the first two tasks, hinting towards the relevance of incentive systems for explaining treatment differences.

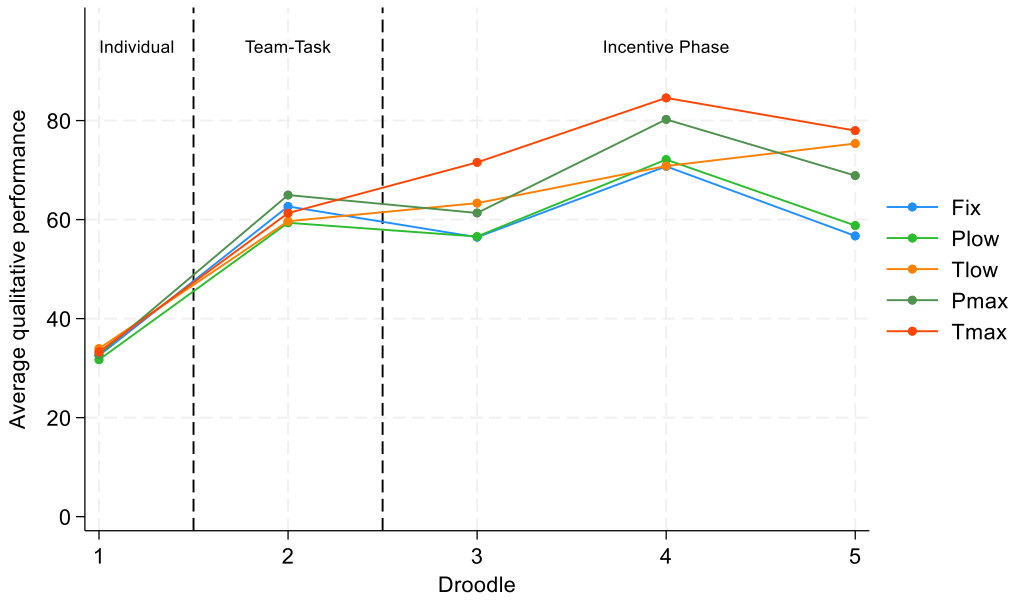


Figure 8: Average qualitative performance per treatment over the course of the experiment.

The hypotheses are tested using non-parametric Mann-Whitney U-Tests for pairwise comparisons between treatments with one-sided p-values, as the hypotheses are directed, followed by OLS estimations. The first hypothesis considers the effectiveness of performance-dependent incentives in comparison to fixed payments. Table 21 shows the results of corresponding Mann-Whitney U-Tests. No constant positive incentive effect is found for *Plow* in comparison to *Fix*. In contrast, *Pmax* and both tournament treatments indicate significantly positive performance effects in comparison to fixed payments. Those effects do not hold for the error rate, especially as the errors made in *Tmax* are higher compared to *Fix*, even though not significant.

Table 21: Results of pairwise comparisons of the different performance measures using Mann-Whitney U-Tests to test for the first hypothesis

H1	Quantity	Quality	Originality	Errors
$Fix < P_{low}$	n.s. (p=0.285)	n.s. (p=0.266)	n.s. (p=0.152)	n.s. (p=0.247)
$Fix < T_{low}$	** (p=0.011)	*** (p=0.004)	*** (p=0.009)	n.s. (p=0.480)
$Fix < P_{max}$	*** (p=0.007)	*** (p=0.003)	*** (p=0.001)	n.s. (p=0.473)
$Fix < T_{max}$	*** (p<0.001)	*** (p=0.001)	*** (p<0.001)	(-) n.s. (p=0.111)

To further understand treatment effects and influential factors like individual and team dynamics on the creative output, multivariate OLS regression analyses with robust standard errors are

conducted.[60] Table 22 shows the results for the qualitative performance measure, as main indicator for creative performance, during the incentive phase. Respective results on quantitative performance can be found in Table AIII. 13 in Appendix III. In the first column, the treatments are implemented with *Fix* being the baseline. No significant incentive effect shows for *Plow*. In the second column, I check for the general team-ability by adding the qualitative performance of the dyadic teams when those are introduced for the first time as control. Corresponding results show a drastic increase in the adjusted R^2 and significant coefficients for the three incentive treatments except for *Plow*.[61] In the following models, the significant positive treatment effects stay robust. Additionally, the coefficient size is constantly the largest for both tournament treatments from model (2) on, hinting towards stronger impact of tournaments on creative performance than piece-rates.

In model (3) the gender composition of a team is added and in model (4) several characteristics on the individual level are introduced, for example perceived difficulty of the tasks or the level of intrinsic motivation.[62] Positive significant effects are found for previous experience with Droodles, perceived pressure, as well as the level of intrinsic motivation. Therefore, the well-known positive influence of intrinsic motivation on creative performance is found. Participants familiar with the task beforehand show higher creative performance, indicating potential learning effects when working on Droodles in general, but keeping the treatment effects significant.

Lastly, in model (5), the characteristics of team dynamics are added. Group cohesion focuses on the group climate and how well members worked together, whereas group identification is about the alignment of task importance and effort provided by oneself and the group. As expected, a feeling of higher group cohesion during the working period indicates significantly higher creative performance. In contrast, a counterintuitive result is found for the level of group identification: A higher level of identification is associated with significantly lower creative performance of dyadic teams, probably indicating a negative effect of groupthink on creative teams which has been shown in previous research (see Janis, 1972; Paulus, 2000). Group cohesion and identification are found to counterbalance each other. The main treatments effects remain robust, though.

[60] The conditions to assume normal distributions are checked, with the result that the assumption of normal distribution of the performance measures needs to be rejected. Therefore, an OLS regression with robust standard errors is applied.

[61] As the team-ability is such a robust significant control variable, also adding explanatory power to regression models, it is used in all following regression models of this essay as well as Essay 4 as control variable, whenever applicable.

[62] These individual perceptions gathered in the questionnaire are averaged over the two team members as the performance analyses are conducted on the team-level.

Table 22: OLS regression on qualitative performance during the incentive phase

VARIABLES	(1)	(2)	(3)	(4)	(5)
Treatments (Base: Fix)					
Plow	3.610	9.236	9.347	12.76*	10.84
	(12.49)	(7.988)	(7.942)	(7.714)	(7.731)
Tlow	25.60**	30.69***	30.94***	32.93***	33.27***
	(11.77)	(8.138)	(8.189)	(8.215)	(8.457)
Pmax	26.57**	22.66***	23.00***	24.25***	23.31***
	(11.59)	(7.752)	(7.696)	(7.549)	(7.720)
Tmax	50.21***	52.52***	52.62***	55.81***	55.53***
	(12.68)	(9.371)	(9.365)	(9.388)	(9.501)
Team-Ability		1.700***	1.717***	1.682***	1.632***
		(0.136)	(0.141)	(0.135)	(0.135)
Gender composition (Base: Only Males)					
Mixed			-0.767	-3.746	-3.607
			(7.063)	(6.934)	(6.836)
Only Females			-5.990	-8.844	-10.41
			(6.934)	(7.325)	(7.513)
Experience Droodle				43.67**	50.53***
				(17.24)	(16.38)
Effort				-1.911	-2.757
				(3.772)	(3.766)
Difficulty				-2.975	-2.968
				(3.866)	(3.910)
Pressure				6.397*	7.650**
				(3.436)	(3.483)
Intrinsic Motivation				14.78***	14.26**
				(5.490)	(6.127)
Group cohesion					18.70**
					(7.631)
Group identification					-11.78*
					(6.082)
Constant	183.9***	77.38***	78.16***	20.43	-8.740
	(9.026)	(9.020)	(9.897)	(29.28)	(36.03)
Observations	278	278	278	278	278
Adj. R-squared	0.068	0.469	0.467	0.485	0.498

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

Focusing on the treatment effect of *Plow,* the coefficient increases when controlling for team-ability and increases severely again in model (4). Table AIII. 14 in Appendix III shows hierarchical OLS regression with each variable being added separately. It shows that after implementing the team-check task, the second half of the increase of the coefficient of *Plow* occurs when introducing intrinsic motivation, thereby underlining a positive effect of intrinsic motivation on creative performance and not necessarily a crowding-out effect through external incentives. The results of the regression analyses confirm the first hypotheses except for *Plow*.

Different potential moderators are tested using interaction effects. No constant statistically significant interaction effects for gender composition, experience, intrinsic motivation, knowing the other team member beforehand, cognitive flexibility, and creative inclination are found (see

Table AIII. 15 in Appendix III). Furthermore, intrinsic motivation, group cohesion and identification, felt pressure and difficulty, and the presence of a team leader are tested as potential mediators, without significant results (see Table AIII. 16 in Appendix III for the indirect effects of the potential mediators).[63]

The second hypothesis refers to the comparison of the incentive intensity by comparing low and high intensity incentives of the same payment scheme with each other. Using Mann-Whitney U-Tests and one-sided p-values, significant differences are found for the measures of quantity, quality, and originality of the two piece-rate treatments, such that participants working under *Pmax* show significantly higher creativity performance. Similar holds for the comparison of the intensity of the two tournament treatments. Additionally, a countervailing significance for the error rate in the tournament treatments shows, such that the number of errors made in the high intensity treatment is significantly larger. In sum, increasing incentive intensity shows a positive significant effect on creative performance, validating the second hypothesis (Table 23).

Table 23: Results of pairwise comparisons of the different performance measures using Mann-Whitney U-Test to test for the second hypothesis

H2	Quantity	Quality	Originality	Errors
Plow < *Pmax*	** (p=0.033)	*** (p=0.009)	** (p=0.034)	(-) n.s. (p=0.218)
Tlow < *Tmax*	** (p=0.018)	** (p=0.046)	* (p=0.059)	(-) * (p=0.083)

Another component of performance, besides the average points achieved, are the SDs indicating variance in the results. Comparing the SDs of the performance categories for the incentive intensity reveals that the two piece-rate treatments differ significantly from each other regarding the SD of the quantitative measure (p=0.048) as well as the number of errors made (p=0.002), such that both are lower for *Pmax*. For the tournament treatments, significant differences are found for quantitative (p=0.061) and qualitative performance (p=0.091), such that higher variability is found for *Tmax*. Those findings indicate a positive tendency for *Pmax* in comparison to *Plow* regarding the steadiness of creative performance induced, vice versa for *Tmax* and *Tlow*.

Lastly, hypothesis 3 aims to test whether comparable piece-rate or tournament incentives differently incentivize creativity in terms of increasing creative performance. Table 24 shows the

[63] Intrinsic motivation is analyzed as moderator but also as potential mediator, as the questionnaire on intrinsic motivation was given to the participants after the experimental phase, such that the treatments itself could influence the intrinsic motivation of participants.

corresponding results, supporting H3. The same effect as for H1 shows, which is that the tournament treatments led to a significantly higher level of creativity focusing on quantity and quality of answers, but this does not hold for the number of errors made. Therefore, the results indicate that a tournament promotes creativity more than a comparable piece-rate scheme. This gets visible too as the coefficient sizes of tournament treatments in the regression analyses above (Table 22) are drastically larger compared to same-intensity piece-rates, indicating higher impact of tournaments on creative team performance.

Table 24: Results of pairwise comparisons of the different performance measures using Mann-Whitney U-Test to test for the third hypothesis

H3	Quantity	Quality	Originality	Errors
Plow < *Tlow*	*	**	n.s.	(-) n.s.
	(p=0.057)	(p=0.018)	(p=0.113)	(p=0.210)
Pmax < *Tmax*	**	*	n.s.	(-) n.s.
	(p=0.020)	(p=0.053)	(p=0.176)	(p=0.127)

Considering the potential difference of the SDs, *Tmax* shows significantly larger deviations than *Pmax* for quantity (p=0.027) and quality (p=0.091). Again, this highlights the possibility of individual reactions affecting performance when competing in winner-takes-all tournaments. No significant differences for the comparison of SDs of low intensity treatments are found.

In summary, I find that PfP systems can increase creative performance relative to a fixed wage, except for low piece-rate compensation paired with a fixed base salary. A positive effect of increasing the incentive intensity is found. However, dominance introducing tournaments in comparison to piece-rate schemes exists. In general, the largest effect sizes, i.e., the coefficients, are found for tournament incentives between teams. Nonetheless, a winner-takes-all tournament comes with comparatively high variations in creative performance.

7 Robustness checks

In the following section, robustness checks to examine the reliability of the main results are presented. By employing alternative performance measures and specifications - including the separation of tournament and incentive intensity effects, the ratio of quality per quantity, relative performance increases, and comparisons across low- and high-ability teams - I assess whether the findings hold under different perspectives.

Tournament or incentive intensity

The previously presented results show that tournaments might dominate piece-rate incentives in the context of creative teams, especially for *Tmax*. To disentangle the effect of incentive size and tournaments further, a regression analysis with two binary variables, indicating whether a

tournament or a high incentive is present, is conducted (see Table 25). The variable *Tournament* includes *Tlow* and *Tmax*, the variable *High Intensity* includes the treatments *Pmax* and *Tmax*, therefore, the interaction effect depicts *Tmax* separately. Thereby, *Plow* (no tournament or high incentive intensity) is automatically set as baseline.

A highly significant positive effect is found when a tournament is present (model (1)). The significant positive effect of a high intensity decreases in size as well as significance (p=0.090) when controlling for the interaction of tournament and high intensity in model (2) (i.e., *Tmax*), while the interaction is not statistically significant. The results suggest that rather a tournament than the fraction of performance-dependent payment positively influences creative teams. Again, no significant interaction effects are found considering team composition regarding gender, experience, intrinsic motivation, or individual characteristics for tournament treatments (see Table AIII. 17 in Appendix III).

Table 25: OLS regression differentiating tournament and incentive effects on qualitative performance during the incentive phase

VARIABLES	(1)	(2)
Tournament	25.44***	21.50**
	(6.422)	(8.692)
High Intensity	18.20***	14.23*
	(6.532)	(8.370)
Team-Check	1.552***	1.555***
	(0.150)	(0.151)
Tournament*High Intensity		7.830
		(12.93)
Constant	93.44***	95.19***
	(9.443)	(9.660)
Observations	224	224
Adj. R-squared	0.406	0.404

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

Further investigating the two tournament treatments separately and especially the high intensity incentive of a winner-takes-all tournament, potential mediators (effort (0.192, p=0.920), pressure (-0.879, p=0.729), self-assessed probability of winning the tournament (6.584, p=0.498)) are investigated. No significant effects are found. Analyzing several interaction effects (i.e., gender composition, probability of winning, risk preference and competitive orientation; see Table AIII. 18 in Appendix III); no significant interaction effects are identified.

Creative performance ratios

The quality of creative answers is determined by general quantity, as the total number of points given by the jurors (i.e., quality) increase with the probability of a valid answer (i.e., quantity)

(Forthmann et al., 2020). Thus, it is necessary to consider the ratios of the qualitative performance measures to the quantity measure. Thereby, I analyze whether the tournament treatments only produce a higher number of answers instead of upholding a stable level of creative quality. Therefore, quality, originality, and the numbers of errors made are divided by the quantity of answers provided. Table AIII. 19 in Appendix III shows the corresponding results using Mann-Whitney U-Tests. Only few significant effects show. Significant differences considering quality per quantity of answers show in positive direction for the comparison of *Fix* with *Tlow* (p=0.034) and with *Pmax* (p=0.067), as well as for *Plow* with *Pmax* (p=0.042) and *Tlow* with *Plow* (p=0.025), all in the expected directions. The same effects as for H1 show for the originality per quantity ratio, as significant differences are found comparing *Fix* with the other treatments except for *Plow*. Although, for errors per quantity no significant differences show, it is important to note that for almost all comparisons, the effect hints towards a higher ratio, thus, more errors per quantity. This is in line with the earlier notion of higher variability in results under PfP systems, even though not significant.

It is important to note that for the other comparisons, no significant effects are found, such that performance-dependent incentive treatments in general do not perform worse in a qualitative sense (i.e., quality and originality) even though providing on average more creative answers. The positive effect on creative performance seems to be driven by an increase in quantity. This does not come at the expense of less qualitatively valuable results.

Therefore, adding the findings from above to the previous results, they indicate that a tournament promotes creativity more than a comparable piece-rate scheme through producing more answers without reducing average quality or originality per answer. As the higher performance of tournaments is mainly driven by quantity, the increase can be interpreted as increasing effort and not increasing the use of creative capabilities.

Increase from Team-task to incentive phase

As Figure 8 above shows, on average, the creative performance of teams increases moving from the team-check task to the incentive phase. This increase is of special interest as it indicates how strongly the incentives implemented might work. A corresponding analysis thus adds to the results of the ratios calculated beforehand, which only focused on the incentive phase. Figure 9 shows the average qualitative performance during the team-check task (blue) and the incentive phase (red), whereas relative increases in average performance when moving from

the team to the incentive phase show for all PfP treatments (see Figure AIII. 9 in Appendix III for quantitative performance).[64]

Table 26 depicts the absolute differences focusing on the main performance measures. Using Kruskal-Wallis Tests followed by Dunn's tests reveals that the increase in both quantity and quality is significantly stronger for the two tournament treatments in comparison to the other three treatments. Thus, a tournament incentive does not only increase quantity but also improve quality compared to previous baseline performance levels with no PfP incentive. Additionally, a significant difference is found between *Fix* and *Plow* for increases in qualitative (p=0.057) and quantitative performance (p=0.047). Thus, if considering relative performance increases rather than absolute average performance, a low piece-rate scheme can outperform fixed payment.

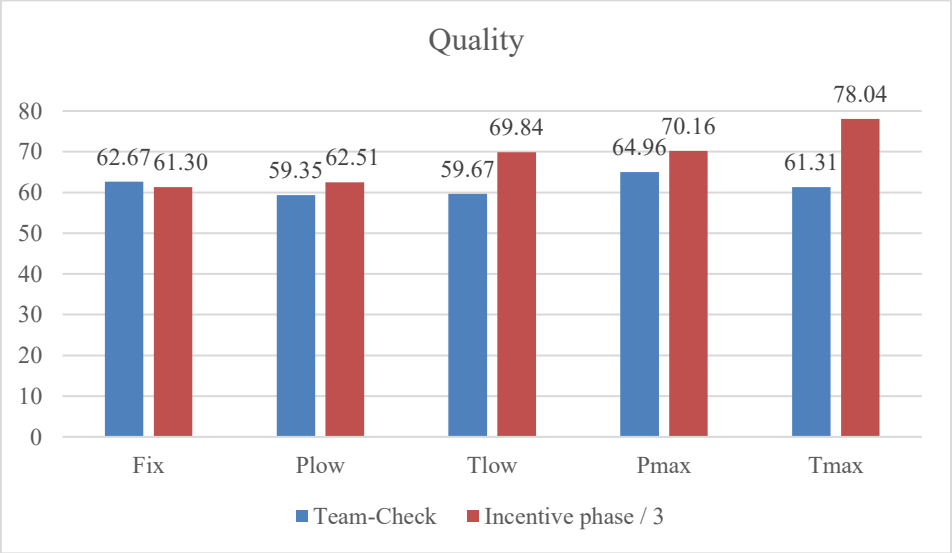


Figure 9: Average qualitative performance of the team-task and the incentive phase averaged over tasks

Table 26: Absolute performance differences between the first team task and the incentive phase averaged over tasks

	Quantity	Quality	Originality	Errors
Fix	-0.648	-1.364	0.420	-0.364
	(2.331)	(13.451)	(1.316)	(1.392)
Plow	0.065	3.149	0.488	0.304
	(2.488)	(17.177)	(1.470)	(1.201)
Tlow	0.776	10.164	0.733	0.194
	(2.601)	(18.180)	(1.590)	(0.881)
Pmax	0.418	5.194	0.703	0.164
	(2.3385)	(19.398)	(1.511)	(1.599)
Tmax	2.489	16.73	0.891	0.437
	(25.497)	(23.646)	(1.391)	(1.332)

[64] The average performance of the incentive phase is divided by the number of Droodles, making the result comparable to the team-check task, which only consisted of one Droodle.

There is no significant difference when comparing the two piece-rate treatments which each other. For the two tournament treatments the comparisons are significant, such that *Tmax* shows significantly higher relative performance increases than *Tlow* for quantity (p=0.008) and quality (p=0.008). Thus, replicating the finding that winner-takes-all tournaments affect creative performance differently than tournaments with a loser prize.

Nonetheless, the same argumentation as above regarding the increase in qualitative performance being driven by quantity may also apply here. Therefore, the same calculation regarding performance increases when implementing the incentive schemes are made for the ratio of quality per quantity (see Figure 10). Significant differences show for the comparison of *Tlow* with fixed payment (p=0.025). Thereby confirming the finding from above that PfP systems drive creative performance especially through effort, i.e., quantity, but holding average quality per answer stable.

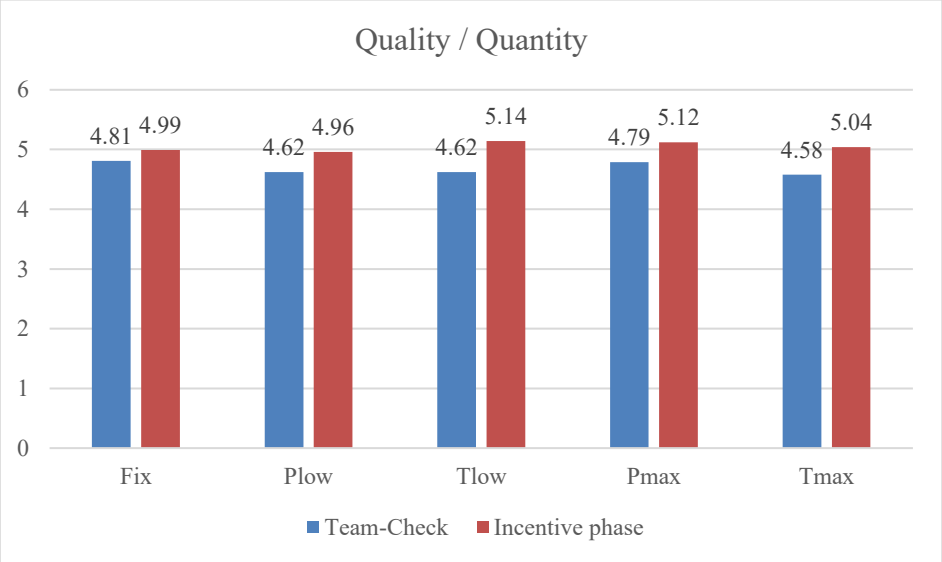


Figure 10: Comparison of the quality per quantity ratio between the team-task and the incentive phase

Top-evaluated ideas

While average idea quality captures the general creative output of a team, practitioners and organizations might be interested in the upper tail of the quality distribution, moving more towards convergent thinking. One step towards this need can be to focus on the highest ranked ideas, i.e., ideas that received either eleven or twelve points from the jury.[65]

[65] The measure of top-ranked ideas gives a qualitative measure of originality, whereas the already established measure of originality is more quantitative, focusing on uniqueness. Please note that the generation of top-ranked ideas is not in focus of the instructions or payment schemes, thus, it is not explicitly demanded. incentivized or in the scope of the present research.

Table 27: Descriptive statistics of teams with top-ranked ideas

	Number of Teams	Top Ideas – Team Phase	Top Ideas per Droodle– Incentive Phase	Top Ideas per Droodle over all teams
Fix	36	17	16.67	0.309
Plow	41	20	24.33	0.434
Tlow	42	11	29.67	0.539
Pmax	46	15	29	0.527
Tmax	42	14	32.67	0.563

Table 27 shows the total number of teams which at least once came up with such a top-evaluated idea, the number of such ideas in the team phase as well as on average per Droodle of the incentive phase respectively for each treatment. Additionally, a relative measure is given, setting the number of top-ranked ideas in context of the sub-sample size per treatment.[66] The latter indicates that both tournament treatments produce a higher number of top ideas, corrected for the number of teams in those treatments. For example, in *Tlow* a team generated on average 0.539 top ideas per Droodle in the incentive phase whereas for *Fix* an average of 0.309 ideas is reached per team working under this condition.

Using non-parametric tests, earlier results of this study partly replicate during the incentive phase with focus on highest evaluated ideas. Treatments with PfP systems show significantly more top-ranked ideas than the fixed pay treatment (p=0.016). No significant difference is found between *Tmax* and *Tlow* (p=0.471), whereas the other comparisons show significant results. Regression analysis confirms this result (see Table AIII. 20 in Appendix III,), showing significant treatment effects except for *Plow* and higher coefficients for both tournament treatments than *Pmax*, indicating stronger effects. Thus, again the notion of tournament incentives per se being more relevant for creative performance than the incentive intensity is underlined.

Low- vs. high-ability teams

An additional question is whether the effects of incentive schemes depend on the ex-ante ability of teams and whether variance in ability affect a team's reaction towards the implemented incentives. To test this, teams are split according to their baseline performance in the team-check task using a median-split. Their performance during the treatment phase is analyzed using OLS regression with interaction effects (see Table AIII. 21 in Appendix III). The results show no systematic significant interaction between teams with a high ability level in comparison to low-ability teams and the treatment conditions, suggesting that incentive effects are not moderated

[66] This correction for sub-sample size is necessary, as sub-samples differ slightly in size, e.g., 54 teams in *Fix* and 58 teams in *Tmax*, and absolute numbers of top-ranked ideas are considered. Assuming equal distribution of creative ability between treatments, the absolute number of top-ranked ideas per treatment increases with the number of teams being able to submit such ideas for the specific treatment.

by initial team ability. Moreover, there is no significant main effect of the median-split itself, neither for quantitative, nor for qualitative measures of performance. Conducting separate regression analyses for low- and high-ability teams, significant treatment effects show especially for low-ability teams (see Table AIII. 22 in Appendix III). Thus, in this experimental setting, PfP incentives show their effectiveness on qualitative creative performance especially for low-ability teams.

Individual level analysis

The previous observations are based on a team level perspective as creative teams are the main focus of the presented essay. Nonetheless, treatments can also have different effects on individual perceptions. Perceiving situations differently can also lead to disagreements and therefore less effective team behavior, especially on a relation-based level (Guenter et al., 2016; Yong et al., 2014). Hence, a regression analysis controlling differences between the perception of team-dynamics, which are assessed by the questionnaires, is conducted (see Table AIII. 23 in Appendix III). Perceiving the contributions of each team member made during teamwork as not equal shows significant negative effects on quantitative (β= -1.269, p=0.027) as well as qualitative performance (β= -8.047, p=0.008), hinting towards a negative effect of conflict in a team on creative performance. Nonetheless treatments do not differ significantly from each other in their perceived levels of contribution (p=0.898). Treatment effects remain unaffected by the additional controls.

Next to differences in perceptions, differences in socio-demographic factors can influence creative performance of teams too, by setting the frame for different experiences and ideas (Dutcher & Rodet, 2022; Hoever et al., 2012; Kurtzberg, 2005). Therefore, in Table AIII. 24 in Appendix III, it is controlled for those differences in, for example, age, nationality, and field of study of both team members. A weakly significant negative effect for different nationalities on qualitative performance (β= -10.219, p=0.075) is found, such that teams with broad differences in nationality show lower qualitative performance.[67] No other significant effects show.

8 Discussion and conclusion

In the present essay it is examined how PfP incentive schemes influence divergent thinking of dyadic teams, a starting point for innovations. The focus lays on two common incentive types, i.e., piece-rates and between-team tournaments, and analyzes how varying the intensity of these incentives, defined as the share of wages tied to performance, affects creative output of teams.

[67] Nationalities are rather broadly defined in this study as German, European and non-European.

This approach allows for a differentiated view of how external monetary incentives shape creativity, including the role of uncertainty and incentive intensity. In addition, the study employs the Droodle-Task to measure teams' divergent thinking, thereby incorporating a creative task that is less known in experimental research.

The results of the laboratory experiment indicate a beneficial effect of tournament incentives and high, but not low intensity piece-rates on creative performance in comparison to fix payment. Therefore, not every PfP incentive leads to higher average creative team performance, contradicting the traditional results from routine tasks. Additionally, increasing incentive intensity, which can be interpreted as increasing the signal sent by incentives but also increasing riskiness, leads to higher levels of creative performance. Furthermore, results reveal that tournament incentives dominate comparable piece-rate payments. This finding is in line with the general finding from routine tasks that tournaments dominate piece-rate incentives in the context of uncertainty, which a creative task is (Lazear & Rosen, 1981). No indication of free-riding behavior in neither piece-rates nor tournament incentives is found, even though egalitarian sharing rules are implemented (Gunnthorsdottir & Rappoport, 2006; Hoffman & Thommes, 2022).

Moreover, it is shown that the intensity of an incentive matters for increasing creative performance, but the presence of a tournament between teams shows even stronger impact on creative performance. This result underlines the notion that between-team competition leads to behavioral changes on the individual level, affecting creative performance. To test those potential changes, different individual components of behavior and perceptions are analyzed, either as aggregate of a team's members or the difference between team members. Positively significant effects are found for previous experience with the task, felt pressure, intrinsic motivation, and group cohesion, whereas a significant negative effect is found for group identification, potentially hinting towards the negative effects of groupthink behavior (Janis, 1972).

With regard to the five implemented incentive schemes, it needs to be noted that a piece-rate incentive coupled to a fix base payment is not found to be effective, but also not harmful, in comparison to pure fixed payment when considering average performance during the incentive phase. Not only is the absolute level of creative performance under such a piece-rate treatment similar to the performance under fixed payment, but also the SDs of the valid answers produced are significantly higher than in a piece-rate scheme with increased intensity. This indicates higher instability and on average lower creativity of results produced under the low-intensity piece-rate scheme in comparison to a high-intensity piece-rate scheme. With the same focus on

variability of results, a winner-takes-all scenario shows fewer stable results than a tournament with a loser prize, especially when focusing on the number of errors made and relatively larger SDs than all other treatments. These findings are in line with Principal-Agent theory which describes a counterbalancing effect between risks associated with variable payments and their positive incentive effect (Jensen & Meckling, 1976).

In general, the preliminary pessimistic concept of SDT that extrinsic rewards decrease creativity does not hold in the presented experimental setting. To the contrary, and more in line with the reworked theory of Amabile and Pratt (2016), extrinsic incentives can foster creativity of teams, probably through their informational signal on the importance of creative performance for payment.

As the calculated ratios of originality (i.e., rarity) and quality per quantity mostly do not show significant differences between treatments, it indicates that the higher performance in the tournament treatment is not due to an increase in creative behavior and thinking, but due to quantity of valid answers which can be attributed to the effort participants put into solving a Droodle. Those calculations might indicate that there is no benefit of PfP incentives when not a general number of good ideas are required, but rather one best idea (i.e., high quality). This thought regarding the generation of top-ranked ideas instead of many good ideas was tested by comparing the fraction of such top ideas per treatment and Droodle. Regression analysis revealed the same positive treatment effects of PfP schemes except for *Plow* in comparison to fixed payment, while the tournament treatments show higher coefficients than piece-rates. Thus, PfP with either a high intensity piece-rate or between-team tournament schemes do not only produce higher average creative performance than fixed payment but also significantly more top-ranked ideas. This is an important research finding, paving the way for studying incentive schemes not only for divergent but also convergent thinking.

In contrast, when focusing on the increase in creative performance induced by different incentives, relative increases appear for all PfP treatments for quantity and quality and, therefore, creative effort, contradicting results of previous research (e.g., Erat & Gneezy, 2016). Furthermore, this highlights that performance under a low-intensity piece-rate exceeds fixed payment too, focusing on the relative performance increase and not the absolute average measure. It is important to note that the participants in the different treatments do not differ significantly from each other in their individual performance nor the ability to work in a team. Therefore, those increases in creative performance during the incentive phase are attributable to the incentive

schemes themselves. Additionally, the incentive effects appear robust controlling for the general ability of a team to solve a creative task.

Further results on individual differences in perceptions show that perceiving individual contributions made as unequal has a negative effect on team performance. Thus, hinting towards the relevance of team climate for a successful creative team. However, no effects of socio-demographic differences between team members are found, except for differences in nationalities. As nationality is rather broadly defined in the present study, differences in such can indicate differences in perception of what is creative (e.g., Cabra & Guerrero, 2022), but also, of course, potential language barriers in team members' communication.

8.1 Limitations

The presented study is hampered by some limitations, which are pointed out in the following. First, no significant mediator or moderator effects could be found, offering no other channel of explanation than treatments directly affecting creative performance. The fact that individual preferences and perceptions have been assessed after the experimental tasks limits their generalizability. One aspect not fully captured in this study is task autonomy, which has been shown to play an influential role for team performance (Kim, 2025). In the absence of additional moderating factors such as leadership or organizational structure (Liu et al., 2022; Pfister & Lukka, 2019), the present design focuses on perceived pressure as an alternative measure, which is negatively correlated with autonomy (Jang et al., 2025; W. Zhang et al., 2017). Given this negative correlation, pressure can serve as a reasonable proxy in the current context. The results of OLS regressions detect perceived pressure as positively significant control variable.

As this study has been conducted in a laboratory setting, the observed effects capture only short-term responses towards incentives. Long-term dynamics, such as potential sorting into particular workplaces (Grund et al., 2025) or incentive schemes (Kim, 2025), are not considered. In practice, self-selection processes may shape how individuals and teams align with their tasks and incentives over time (Kaduk et al., 2019; Kuhn & Villeval, 2015). At the same time, idea generation is a recurring short-term task for many teams not determining their entire working situation. Thus, the results of the present study are transferable towards such workplace contexts.

Additionally, only a team size of two is considered, thereby not allowing implications for larger teams, especially with focus on the potential effect of free-riding in teams. Next to the limited team size, the focus on divergent thinking as component of creativity can limit the use case of

the results towards situations at the beginning of an innovative process, not allowing implications for the stage of idea development and implementation (Reiter-Palmon et al., 2019). At the same time, this focus can also be considered a strength, as the study contributes by providing specific insights into the role of incentives for creativity in the early phase of innovation processes, affecting later results.

Furthermore, the incentive intensities might not be perceived as truly maximized by participants, as all participants received a fixed amount of €4 for participation and the first two Droodles. This was required to ensure that no one left the experiment empty-handed and was communicated up front to prevent varying questions or concerns of participants in the pure piece-rate and winner-takes-all tournament conditions, which could have potentially distorted the results and would make them no longer fully comparable with the other treatments. This previous payment could be interpreted as a fixed base payment or as a loser prize. However, this limitation should be viewed in relative terms: in any experimental or real-world setting, participants are never entirely without an outside option (e.g., unemployment benefits or alternative activities). In this sense, the small, fixed amount serves as an ethical safeguard rather than a substantive distortion of the relative incentive intensities in the treatments and thus does not fundamentally undermine the interpretation of the results.

Lastly, no other incentives, especially non-monetary incentives like recognition and reciprocity (Bradler et al., 2019) are considered in this study, which of course can play a role for creative behavior of workplace teams.

8.2 Implications for further research

Following from the results and limitations, several promising avenues for further research show. Firstly, the team size is limited to two, such that implications for larger teams are hard to make, as for example deviations in the role of team cohesion and communication patterns might occur. This calls for further research replicating this study with larger team sizes to assess whether this is relevant for the link between incentives and creative performance.

Moreover, the present study does not show an effect of participants' sex or the gender composition of a team on creative performance, but such potential effects are strongly discussed in the field of tournament incentives (e.g., Geraldes, 2020; Niederle & Vesterlund, 2011) and creativity (e.g., Baer & Kaufmann, 2008; Vedres & Vásárhelyi, 2023). Yet, no study considers both the incentive and the dimension of creative teams together with the role of gender. Following research should extend the research question towards the influence of the gender composition of a team.

As already discussed earlier, the present design uses felt pressure as control variable, which is negatively correlated to autonomy, yet future research should explicitly integrate perceived autonomy as an additional dimension (Kim, 2025) to check the validity of findings with regard to perceived pressure. Building on the need for autonomy, self-selection options for deciding on a preferred incentive scheme can be implemented.

The presented findings focus on divergent thinking. Analyzing the number of highest ranked ideas reached by teams as additional performance measures during the robustness checks hints towards incentive effects on developing single best ideas, but this is no test for incentive effects on convergent thinking. Nonetheless, the findings that positive treatment effects also apply to this performance dimension can serve as a great starting point for research on subsequent stages of idea development and implementation. Thus, allowing a comprehensive understanding of how incentives influence creativity across the full innovation cycle.

Another dimension of pressure and competition implemented in tournaments is feedback on the performance of other teams (Wooten & Ulrich, 2017). Such feedback could add another dimension, adapting the experiment even further towards relevance for practical working life. As validity and quality of creative answers have been assessed after the experiment was conducted, implementing feedback on the numbers of ideas, valid or not, submitted by a competing team or the general average of ideas submitted by a team could be implemented as additional component in tournament treatments. Depending on the level of feedback, i.e., average, top-performers or live feedback, this might induce differently perceived competition, potentially opening up the path for behavior known as choking under pressure (Ariely et al., 2009).

Next to monetary incentives, non-monetary incentives can play a role in altering behavior and performance in teams (Kube et al., 2012). Focusing on the practical example of large public architecture tenders brough up earlier, it hints towards potential relevance of non-monetary and non-physical rewards like reputation. As those competitions are winer-takes-all tournaments around the most creative idea, only one winner or winning team can exist at the end. Nonetheless, becoming second in a large and prestigious competition can function as signal of creative potential, having signaling effects for further projects. Therefore, such non-monetary incentives as recognition should be considered in future research.

Generally speaking, incentives and their intensity are not a one-size-fits-all solution, as shown by Opitz et al. (2024). The authors conducted an experimental study in which a machine-learning algorithm has been trained to assign the best possible incentive scheme with focus on performance towards individuals with certain personality traits. As their study has been conducted

on the individual level with a routine task, the question arises whether such an assignment of optimal incentives is also possible for teams in a creative context. This would require an enlarged sample size and an adjusted experimental design, focusing more on team characteristics in advance.

8.3 Practical implications

The findings regarding the effectiveness of PfP incentives in comparison to fixed payment, the performance enhancing effect of increasing incentive intensity, and the dominance of tournament schemes over comparable piece-rate schemes carry several practical implications. Most directly, the results are relevant for the incentivization of idea generation in small teams engaged in such tasks on a non-permanent basis, given that the experimental design considered only a limited time horizon. Accordingly, the implications are particularly salient for the early stages of the creative process and are not confined to any specific professional domain.

Beyond their relevance for HR practitioners operating within existing organizational structures, these findings also bear implications for managers and consultants engaged in the (re-)design of corporate compensation strategies and work arrangements that periodically require divergent thinking. Given that PfP incentives are found to yield positive effects on creative performance relative to a team's previous ability, such schemes may be considered not only when a sustained general level of creative output is desired, but also when incremental improvements over prior performance are the objective.

More specifically, when a large volume of ideas with relatively consistent quality is required, a low-intensity tournament structure may represent the most suitable design among those examined, as tournament schemes generally outperformed piece-rates, while high intensity tournaments exhibited greater variance in outcomes. Conversely, when the objective is to elicit top-ranked or breakthrough ideas next to on average high performance, a winner-takes-all tournament format appears most appropriate, as this condition is associated with the highest levels of creative performance, also with respect to highest ranked ideas.

8.4 Contribution

The present study contributes to the growing body of research on non-routine work and, in particular, team-based creative performance, by examining the effects of PfP incentive schemes and their intensities on divergent thinking in a controlled laboratory setting. More precisely, this research advances understanding of influencing divergent thinking outcomes, which refers

to generating several creative ideas, being part of the early process of many innovations. Therefore, being able to incentivize creative performance in this stage can effectively benefit the whole process, enabling innovations and competitive advantages for firms.

The essay makes several distinct contributions to existing literature. First, the study adds to a better understanding of the effects of the two most prominent incentive schemes, i.e., piece-rate and tournaments, on the creative performance of a team, by systematically varying incentive intensity, i.e., the wage fraction that is determined by performance, in both PfP schemes. Rather than comparing two arbitrarily chosen incentive structures, as is common in prior work, this design enables a nuanced assessment of the interplay between incentive and risk effects on creative performance, thereby addressing a gap explicitly identified in the literature (Kim, 2025; Kim et al., 2022). In doing so, the study also contributes to the Principal-Agent literature by providing empirical evidence on how agents respond to performance-contingent pay under conditions of high output uncertainty in a non-routine environment, a context that classical formulations of the theory were not primarily designed to address (Gerhart & Fang, 2015; Jensen & Meckling, 1976). By demonstrating that the structural form of the incentive scheme matters more than its intensity alone, the findings delineate an important boundary condition of standard Principal-Agent predictions.

Second, the study contributes to the underexplored domain of incentivizing creative teams, as opposed to individuals. By focusing on dyadic teams engaged in a divergent thinking task, it integrates insights from tournament theory, team dynamics, and creativity research. In addition, the experimental evidence speaks to the ongoing theoretical debate surrounding SDT and the relationship between extrinsic rewards and intrinsic motivation. By identifying conditions under which PfP incentives enhance rather than crowd out creative performance, the findings align with the revised formulation of SDT (Amabile & Pratt, 2016; Ryan & Deci, 2000) and contribute empirical nuance to a literature that remains divided on this question.

Third, the study contributes to the innovation contest and competition design literature by explicitly incorporating and comparing winner-takes-all tournament structures with loser-prize alternatives. While contest theory has addressed the optimal design of innovation competitions largely from a theoretical standpoint (Ales et al., 2017; Krishna & Morgan, 1998; Terwiesch & Ulrich, 2009), controlled experimental evidence on how different prize structures affect creative output in team settings remains scarce. The present study addresses this gap directly, offering empirical guidance on the conditions under which specific tournament designs are most effective for eliciting creative team performance.

Fourth, the study employs the Droodle-Task as an instrument for assessing divergent thinking at the team level, thereby expanding the methodological repertoire available to researchers in the field of creativity.

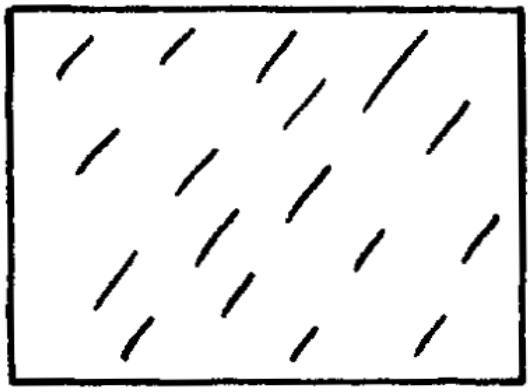
The results demonstrate that tournament incentives, in particular, effectively enhance both the quantity and quality of creative performance in dyadic teams. Furthermore, increasing incentive intensity is found to positively affect creative performance, though the structural choice between absolute and relative incentive systems proves to be of greater consequence, as tournament conditions consistently outperform comparable piece-rate treatments. For practitioners seeking stable and less variable creative output, the findings additionally suggest that a between-team tournament with a loser prize represents a particularly suitable incentive design inducing a high number of creative ideas with high and relative stable levels of quality of results.

APPENDIX III

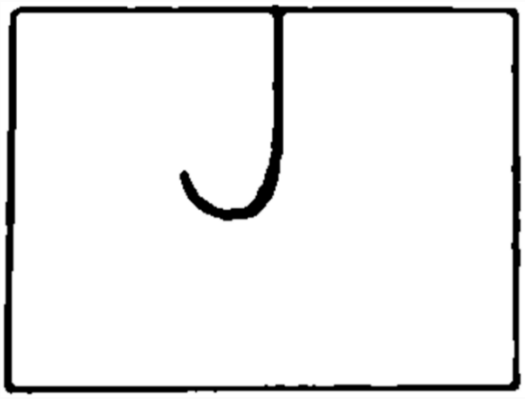
Acknowledgements

I thank the Gesellschaft für experimentelle Wirtschaftsforschung (GfeW) for funding the treatments *Tmax* and *Pmax* with the Rheinhard-Selten scholarship.

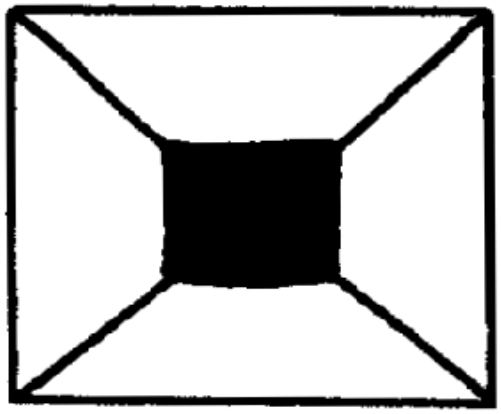
Figures



AIII. 1: Droodle used as ability-check



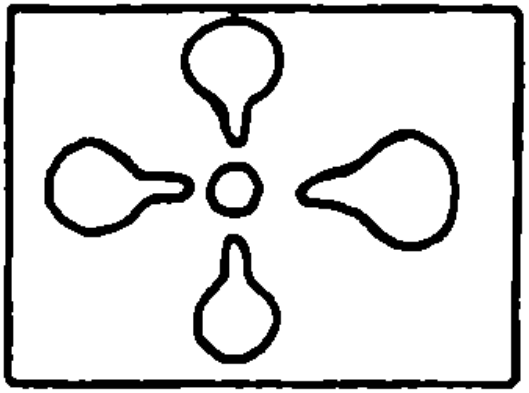
AIII. 2: Droodle used in the team-check



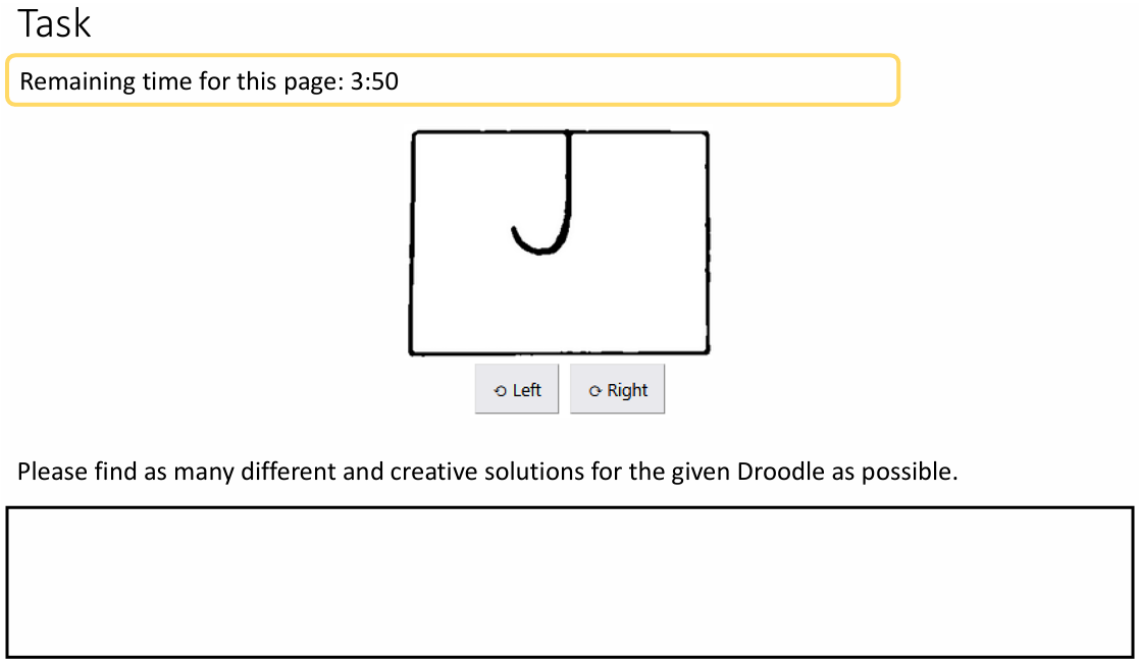
AIII. 3: First Droodle of the incentive phase



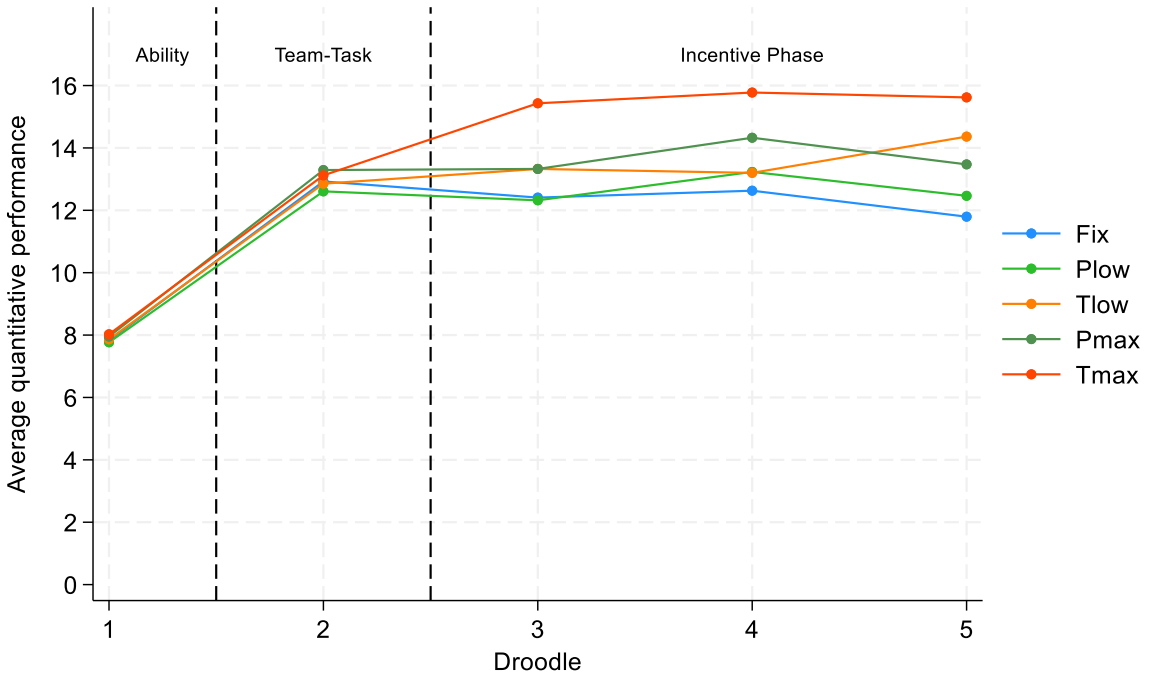
AIII. 4: Second Droodle of the incentive phase



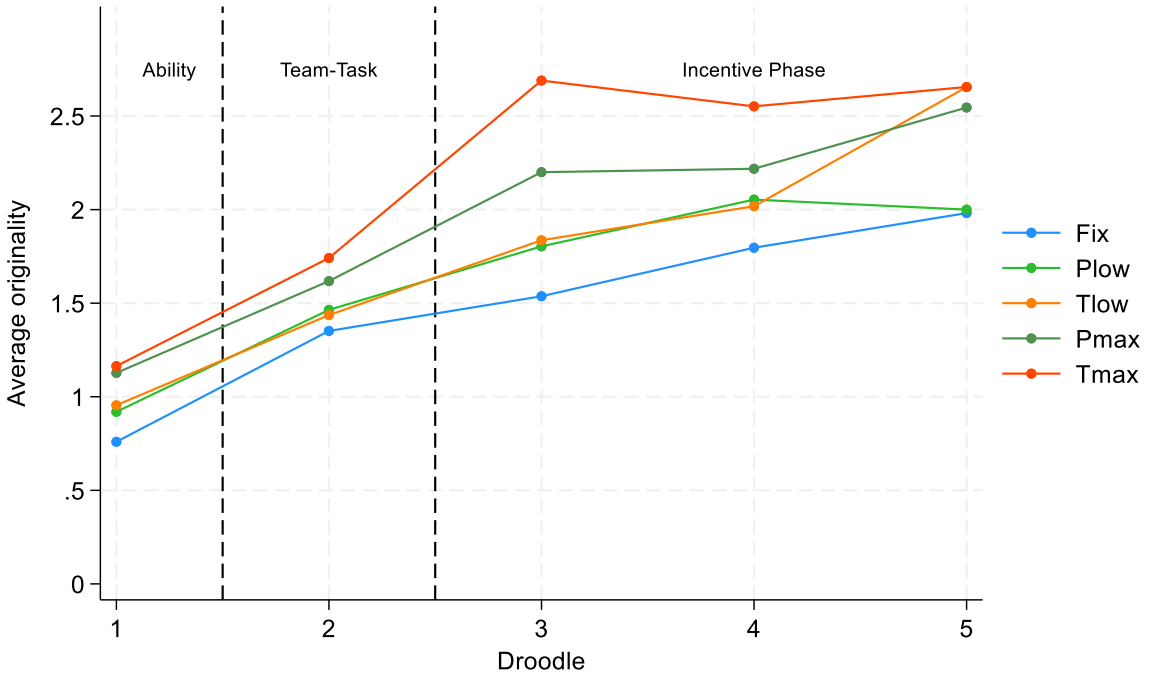
AIII. 5: Third Droodle of the incentive phase



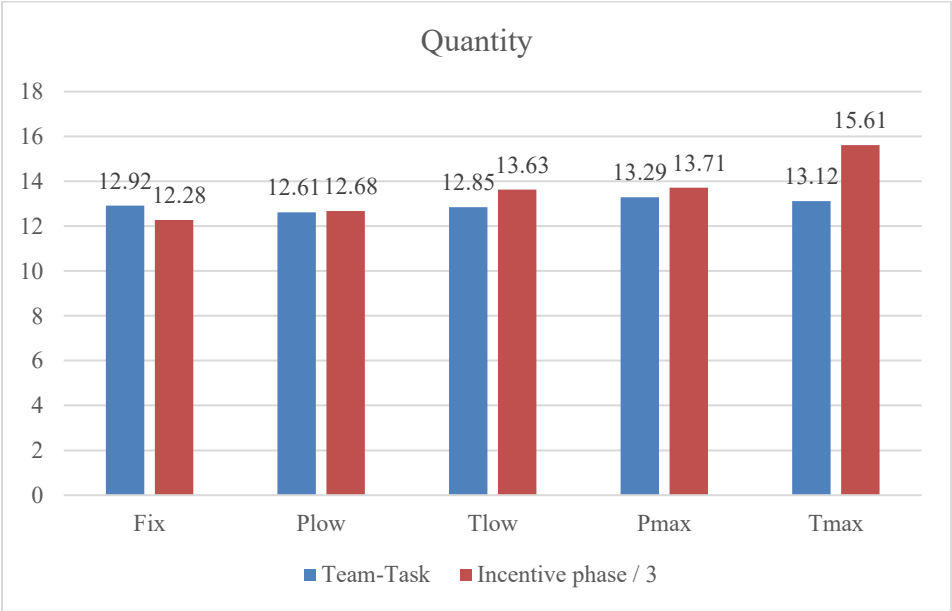
AIII. 6: Example screen



AIII. 7: Average quantitative performance per treatment over the course of the experiment



AIII. 8: Average originality of answers per treatment over the course of the experiment



AIII. 9: Average quantitative performance in the team-check task and the incentive phase averaged over tasks

Tables

AIII. 10: Sample characteristics

Treatment	N	Age	# of Women	Gender	Non-EU	# of Non-Bachelor	# with Econ background
Fix	108	22.94	47	4.84	15	76	56
Plow	112	22.33	50	4.84	12	77	61
Tlow	110	23.06	52	4.88	19	79	68
Pmax	110	22.41	53	4.96	5	88	69
Tmax	116	22.73	51	4.97	24	79	74
Whole Sample	556	22.70	253	4.83	75	399	328

Notes: Age is given as mean and given per year; the other values are absolute numbers. Gender is self-assessed on a scale from 1(very masculine) to 10 (very feminine). Econ background refers to whether participants have an economic background, for example through their field of study.

AIII. 11: Means of different experiment-related controls measured on the individual level

Treatment	Intrinsic Motivation*	Cognitive Flexibility	Effort	Difficulty	Pressure	Leader	High Effort Group	Cohesion	Identification *	Group Fun *	Group Advantages
Fix	4.03	3.815	4.259	3.019	2.565	2.269	3	4.045	3.475	4.361	3.204
	(0.603)	(0.503)	(1.036)	(1.023)	(1.277)	(1.220)	(1.207)	(0.517)	(0.742)	(0.837)	(0.567)
Plow	3.773	3.78	4.205	2.973	2.536	2.357	2.866	4.035	3.335	4.214	3.11
	(0.693)	(0.468)	(1.075)	(1.009)	(1.237)	(1.161)	(1.263)	(0.481)	(0.695)	(0.944)	(0.579)
Tlow	3.922	3.862	4.2 (1.056)	2.873	2.555	2.373	3	4.074	3.555	4.236	3.167
	(0.622)	(0.488)		(1.006)	(1.201)	(1.291)	(1.263)	(0.585)	(0.646)	(0.985)	(0.643)
Pmax	3.965	3.9	4.418	2.964	2.536	2.291	2.636	4.119	3.491	4.255	3.136
	(0.531)	(0.394)	(0.861)	(0.908)	(1.106)	(1.222)	(1.187)	(0.439)	(0.694)	(0.861)	(0.493)
Tmax	3.863	3.946	4.224	2.914	2.707	2.664	2.767	4.111	3.556	4.069	3.124
	(0.656)	(0.501)	(1.080)	(1.068)	(1.224)	(1.285)	(1.260)	(0.427)	(0.767)	(0.921)	(0.564)

Note: The stars (*** p<0.01, ** p<0.05, * p<0.1) denote whether the treatments differ between each other in the corresponding dimension, tested with Kruskal-Wallis Tests.

AIII. 12: Pearson correlations of individual characteristics, perceptions, and total performance measures

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Gender	1.000												
(2) Age	-0.029	1.000											
(3) Knowledge	0.034	-0.125***	1.000										
(4) Intrinsic Motivation	0.089*	-0.037	-0.034	1.000									
(5) Effort	0.006	-0.087*	0.026	0.063	1.000								
(6) Difficulty	-0.041	-0.002	0.012	-0.230***	0.254***	1.000							
(7) Pressure	0.078*	-0.012	0.032	-0.088*	0.203***	0.332***	1.000						
(8) Group advantages	0.022	-0.059	0.044	0.108*	0.006	0.089*	0.128***	1.000					
(9) Group cohesion	0.083*	-0.034	0.204***	0.289***	0.097*	-0.058	-0.015	0.259***	1.000				
(10) Group identification	0.083*	0.014	-0.055	0.364***	0.006	0.009	0.130***	0.172***	0.247***	1.000			
(11) Cognitive flexibility	-0.081*	0.044	-0.061	0.200***	-0.054	-0.204***	-0.143***	0.041	0.142***	0.165***	1.000		
(12) Creative inclination	-0.118***	0.040	-0.022	0.091*	-0.067	-0.120***	-0.043	0.032	0.035	0.141***	0.540***	1.000	
(13) Econ. background	-0.027	-0.032	-0.081*	-0.028	0.037	-0.040	0.023	-0.045	-0.032	-0.056	0.070	-0.018	1.000
(14) Familiarization total	-0.070*	0.022	0.017	0.019	0.132***	-0.012	0.075*	-0.031	0.097*	-0.106*	-0.028	-0.008	0.023
(15) Team-Check total	0.112***	0.039	-0.024	0.035	0.170***	0.072*	0.120***	0.096*	0.086*	-0.054	-0.037	-0.022	0.053
(16) Incentive total	0.074*	0.005	-0.027	0.106*	0.114***	0.005	0.109***	0.105*	0.151***	0.010	0.016	0.043	0.052

*** p<0.01 ** p<0.05 * p<0.1

AIII. 13: OLS regression on quantitative performance during the incentive phase

VARIABLES	(1)	(2)	(3)	(4)	(5)
Treatments (Base: Fix)					
Plow	1.185	1.933	1.934	2.479*	2.299*
	(2.525)	(1.360)	(1.364)	(1.347)	(1.355)
Tlow	4.058*	4.225***	4.227***	4.531***	4.478***
	(2.285)	(1.346)	(1.349)	(1.358)	(1.388)
Pmax	4.294*	3.437***	3.440***	3.723***	3.582***
	(2.243)	(1.263)	(1.265)	(1.203)	(1.222)
Tmax	9.994***	9.537***	9.538***	10.01***	9.864***
	(2.543)	(1.769)	(1.776)	(1.798)	(1.805)
Team-Ability		2.348***	2.348***	2.317***	2.297***
		(0.147)	(0.151)	(0.140)	(0.140)
Gender composition (Base. Only Males)					
Mixed			0.00246	-0.523	-0.500
			(1.231)	(1.203)	(1.199)
Only Females			-0.0454	-0.563	-0.828
			(1.179)	(1.246)	(1.286)
Experience Droodle				6.742**	7.200**
				(2.994)	(2.850)
Effort				-0.799	-0.876
				(0.664)	(0.670)
Difficulty				-0.492	-0.518
				(0.710)	(0.714)
Pressure				1.210**	1.264**
				(0.593)	(0.604)
Intrinsic Motivation				2.405**	2.011*
				(0.936)	(1.054)
Group cohesion					2.156*
					(1.219)
Group identification					-0.538
					(1.087)
Constant	36.83***	6.487***	6.488***	-1.137	-5.839
	(1.753)	(1.896)	(2.000)	(5.365)	(6.559)
Observations	278	278	278	278	278
Adj. R-squared	0.061	0.603	0.601	0.614	0.614

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

AIII. 14: OLS regression on qualitative performance during the incentive phase with stepwise inclusion of the control variables

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatments (Base: Fix)										
Plow	3.610	9.236	9.347	8.992	8.973	8.855	8.843	12.76*	11.45	10.84
	(12.49)	(7.988)	(7.942)	(8.884)	(7.763)	(7.730)	(7.720)	(7.714)	(7.791)	(7.731)
Tlow	25.60**	30.69***	30.94***	31.73***	31.71***	31.22***	31.01***	32.93***	31.87***	33.27***
	(11.77)	(8.138)	(8.189)	(8.931)	(8.147)	(8.207)	(8.209)	(8.215)	(8.380)	(8.457)
Pmax	26.57**	22.66***	23.00***	22.98**	23.08***	22.77***	22.99***	24.25***	22.92***	23.31***
	(11.59)	(7.752)	(7.696)	(8.926)	(7.557)	(7.641)	(7.648)	(7.549)	(7.656)	(7.720)
Tmax	50.21***	52.52***	52.62***	54.70***	54.68***	54.34***	53.41***	55.81***	53.88***	55.53***
	(12.68)	(9.371)	(9.365)	(8.843)	(9.266)	(9.322)	(9.314)	(9.388)	(9.465)	(9.501)
Team-Ability		1.700***	1.717***	1.701***	1.706***	1.709***	1.689***	1.682***	1.669***	1.632***
		(0.136)	(0.141)	(0.119)	(0.135)	(0.137)	(0.137)	(0.135)	(0.134)	(0.135)
Gender composition (Base: Only Males)										
Mixed			-0.767	-1.491	-1.533	-1.721	-2.272	-3.746	-3.546	-3.607
			(7.063)	(6.625)	(6.965)	(6.949)	(6.908)	(6.934)	(6.879)	(6.836)
Only Females			-5.990	-4.904	-4.915	-4.955	-6.534	-8.844	-11.43	-10.41
			(6.934)	(7.018)	(6.874)	(6.900)	(7.132)	(7.325)	(7.536)	(7.513)
Experience Droodle				44.81**	44.75**	45.60***	46.95***	43.67**	45.29***	50.53***
				(18.53)	(17.36)	(17.47)	(17.30)	(17.24)	(16.59)	(16.38)
Effort					-0.683	0.306	-0.329	-1.911	-2.391	-2.757
					(3.629)	(3.655)	(3.680)	(3.772)	(3.844)	(3.766)
Difficulty						-3.918	-5.655	-2.975	-3.351	-2.968
						(3.965)	(3.919)	(3.866)	(3.885)	(3.910)
Pressure							5.457	6.397*	6.339*	7.650**
							(3.384)	(3.436)	(3.417)	(3.483)
Intrinsic Motivation								14.78***	9.783	14.26**
								(5.490)	(6.051)	(6.127)
Group cohesion									17.55**	18.70**
									(7.665)	(7.631)
Group identification										-11.78*
										(6.082)
Constant	183.9***	77.38***	78.16***	76.23***	78.85***	86.28***	81.98***	20.43	-25.79	-8.740
	(9.026)	(9.020)	(9.897)	(10.37)	(17.21)	(18.77)	(18.87)	(29.28)	(34.78)	(36.03)
Observations	278	278	278	278	278	278	278	278	278	278
Adj. R-squared	0.068	0.469	0.467	0.476	0.474	0.474	0.477	0.485	0.493	0.498

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

AIII. 15: OLS regression on qualitative performance including interaction effects between several control variables and treatment effects

VARIABLES	(1) Gender composition	(2) Experience Droodle	(3) Intrinsic Motivation	(4) Knowledge	(5) Cognitive Felxibility	(6) Inclination
Treatments (Base: Fix)						
Plow	15.70	9.372	35.81	3.077	-135.9	-53.59
	(14.52)	(8.091)	(58.19)	(8.569)	(90.56)	(74.27)
Tlow	21.88	34.76***	48.97	28.36***	-104.2	48.86
	(16.60)	(8.366)	(60.80)	(8.956)	(94.37)	(87.97)
Pmax	35.15**	25.49***	56.57	19.06**	-123.4	63.12
	(14.00)	(8.162)	(69.41)	(8.257)	(98.40)	(84.96)
Tmax	51.08***	56.02***	33.73	52.33***	-83.78	11.63
	(18.45)	(9.574)	(56.75)	(10.15)	(110.1)	(80.58)
Team-Ability	1.731***	1.676***	1.694***	1.706***	1.707***	1.712***
	(0.141)	(0.127)	(0.135)	(0.137)	(0.137)	(0.139)
Gender composition (Base: Only Males)						
Mixed	7.038					
	(13.13)					
Only Females	-10.43					
	(12.55)					
*Treatment * Gender composition*						
Plow*Mixed	-12.52					
	(19.89)					
Plow*Only Females	-6.871					
	(18.26)					
Tlow*Mixed	1.188					
	(20.88)					
Tlow*Only Females	29.96					
	(19.90)					
Pmax*Mixed	-33.01*					
	(18.27)					
Pmax*Only Females	-4.106					
	(18.75)					
Tmax*Mixed	4.300					
	(24.03)					
Tmax*Only Females	1.521					
	(21.87)					
Experience Droodle		70.27*				
		(36.91)				
*Treatment * Experience Droodle*						
Plow*Experience Droodle		-9.518				
		(53.03)				
Tlow*Experience Droodle		-61.20				
		(59.29)				
Pmax*Experience Droodle		-42.29				
		(44.76)				
Tmax*Experience Droodle		-11.19				
		(66.99)				
IM			15.87			
			(9.755)			
*Treatment * IM*						
Plow*IM			-5.961			
			(15.22)			
Tlow*IM			-4.227			
			(15.67)			
Pmax*IM			-8.309			
			(17.74)			
Tmax*IM			5.549			
			(14.86)			
Knowledge				-22.86**		
				(10.53)		
*Treatment * Knowledge*						

Plow*Knowledge				51.31***		
				(19.36)		
Tlow*Knowledge				24.98		
				(19.43)		
Pmax*Knowledge				36.32		
				(22.11)		
Tmax*Knowledge				-1.884		
				(19.81)		
Cognitive flexibility					-19.72	
					(18.26)	
*Treatment * Cogn. flexibility*						
Plow*Cogn. Flexibility					38.21	
					(23.62)	
Tlow*Cogn. Flexibility					35.17	
					(24.78)	
Pmax*Cogn. Flexibility					37.86	
					(25.40)	
Tmax*Cogn. Flexibility					35.20	
					(28.17)	
Creative inclination						8.266
*Treatment * Creative inclination*						(15.15)
Plow*Creative Inclination						18.29
Tlow*Creative Inclination						(21.67)
Pmax*Creative Inclination						-5.218
						(25.49)
Tmax*Creative Inclination						-11.59
						(24.79)
						11.41
						(23.69)
Constant	75.68***	74.33***	13.75	79.10***	152.2**	47.77
	(12.51)	(8.959)	(41.14)	(9.261)	(70.78)	(52.89)
Observations	278	278	278	278	278	278
Adj. R-squared	0.494	0.493	0.490	0.488	0.486	0.485

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

AIII. 16: Indirect effects of potential mediators of PfP systems on qualitative creative performance with *Fix* as baseline

	IM	Cohesion	Identification	Leader	Pressure	Difficulty
Plow	-3.327	-8.417	4.587	1.128	-1.898	1.073
	(p=0.568)	(p=0.159)	(p=0.418)	(p=0.748)	(p=0.488)	(p=0.713)
Tlow	-5.636	-0.395	0.149	0.271	-3.909	1.928
	(p=0.339)	(p=0.889)	(p=0.914)	(p=0.800)	(p=0.362)	(p=0.529)
Pmax	-0.472	-0.002	1.581	-0.742	-3.050	0.619
	(p=0.824)	(p=0.999)	(p=0.569)	(p=0.720)	(p=0.425)	(p=0.643)
Tmax	-4.334	0.592	-0.004	-2.276	1.252	6.640
	(p=0.375)	(p=0.885)	(p=0.980)	(p=0.593)	(p=0.642)	(p=0.349)

AIII. 17: OLS regression on qualitative performance including interaction effects between several control variables and tournament effects

VARIABLES	(1) Gender composition	(2) Experience	(3) Intrinsic Motivation	(4) Cognitive Felxibility	(5) Inclination	(6) Team-Task
Tournament	11.66	27.83***	-6.644	27.67	27.41	36.09*
	(11.95)	(6.740)	(46.74)	(67.88)	(64.89)	(18.47)
High Incentive	18.41***	18.66***	17.88***	16.49**	17.68***	18.05***
	(6.494)	(6.490)	(6.619)	(6.664)	(6.499)	(6.553)
Team-Check	1.566***	1.557***	1.528***	1.566***	1.557***	1.631***
	(0.160)	(0.147)	(0.147)	(0.153)	(0.152)	(0.187)
Gender composition (Base: Only Males)						
Mixed	-14.94					
	(9.963)					
Only Females	-14.18					
	(9.903)					
*Tournament*Sex compo*						
Tournament*Mixed	23.41					
	(16.18)					
Tournament*Only Females	20.11					
	(15.38)					
Experience		45.19*				
		(24.39)				
Tournament*Experience		-23.69				
		(45.47)				
IM			8.986			
			(9.040)			
Tournament*IM			8.193			
			(12.33)			
Cognitive flexibility				16.29		
				(11.21)		
Tournament*Cogn. flexibility				-0.822		
				(17.61)		
Inclination					13.05	
					(12.85)	
Tournament*Inclination					-0.742	
					(18.74)	
Tournament*Team-Check						-0.174
						(0.302)
Constant	101.7***	89.85***	60.28*	30.81	47.99	88.59***
	(11.42)	(9.763)	(35.69)	(45.07)	(45.54)	(10.94)
Observations	224	224	224	224	224	224
Adj. R-squared	0.403	0.409	0.411	0.407	0.405	0.404

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

AIII. 18: OLS regression on quantitative and qualitative performance including interaction effects between tournament incentives and tournament related controls

	Qualitative Performance			Quantitative Performance		
VARIABLES	(1) Gender composition	(2) Winning	(3) Competitive Orientation	(1) Gender composition	(2) Winning	(3) Competitive Orientation
Tournament (Base: Tlow)						
Tmax	29.25	-5.841	8.358	19.32	25.19	71.89
	(19.77)	(41.23)	(86.18)	(14.21)	(48.38)	(76.92)
Team-Check	1.465***	1.454***	1.442***	0.716	0.733	0.633
	(0.242)	(0.245)	(0.255)	(0.998)	(1.004)	(0.968)
Gender Composition (Base: Only Males)						
Mixed	7.141			0.396		
	(16.19)			(3.282)		
Only Females	20.83			4.527		
	(15.57)			(2.893)		
Tmax x Mixed	3.315			14.09		
	(25.73)			(23.86)		
Tmax x Only Females	-27.39			10.66		
	(23.64)			(24.94)		
Winning		-0.0221			-0.0828	
		(0.275)			(0.0876)	
Tmax x Winning		0.465			0.0265	
		(0.657)			(0.761)	
Competitive orientation			-14.39			0.974
			(19.22)			(3.873)
Tmax x Competitive orientation			4.499			-13.93
			(26.51)			(22.85)
Constant	113.4***	124.1***	169.4***	30.22**	36.48**	29.64*
	(18.18)	(24.51)	(59.79)	(12.91)	(17.25)	(17.17)
Observations	113	113	113	113	113	113
Adj. R-squared	0.284	0.287	0.288	0.430	0.439	0.435

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

AIII. 19: Mann-Whitney U-Tests comparing different performance ratios between treatments

	Qual / Quan	Originality / Quan	Errors / Quan
Fix < Plow	n.s.	n.s.	n.s.
	(p=0.282)	(p=0.139)	(p=0.185)
Fix < Tlow	**	*	(-) n.s.
	(p=0.034)	(p=0.085)	(p=0.264)
Fix < Pmax	*	***	n.s.
	(p=0.067)	(p=0.009)	(p=0.336)
Fix < Tmax	n.s.	**	(-) n.s.
	(p=0.391)	(p=0.017)	(p=0.342)
Plow < Pmax	**	n.s.	(-) n.s.
	(p=0.042)	(p=0.121)	(p=0.320)
Tlow < Tmax	n.s.	n.s.	(-) n.s.
	(p=0.118)	(p=0.153)	(p=0.141)
Plow < Tlow	**	n.s.	(-) n.s.
	(p=0.025)	(p=0.425)	(p=0.298)
Pmax < Tmax	n.s.	n.s.	(-) n.s.
	(p=0.144)	(p=0.431)	(p=0.232)

AIII. 20: OLS regression on average top-ranked ideas per Droodle during the incentive phase

VARIABLES	(1)
Treatments (Base: Fix)	
Plow	0.120
	(0.0743)
Tlow	0.246***
	(0.0752)
Pmax	0.224***
	(0.0734)
Tmax	0.264***
	(0.0795)
Team-Check top ideas	0.130**
	(0.0541)
Constant	0.268***
	(0.0462)
Observations	278
Adj. R-squared	0.047

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

AIII. 21: OLS regression on quantitative and qualitative performance including interaction effects between ability of teams and treatment effects

VARIABLES	(1) Quantity	(2) Quality
Treatments (Base: Fix)		
Plow	2.065	14.50
	(1.345)	(9.178)
Tlow	6.443***	34.95***
	(1.541)	(9.585)
Pmax	5.222***	37.27***
	(1.270)	(9.775)
Tmax	12.07***	62.57***
	(2.582)	(12.18)
Team-Check	2.494***	1.734***
	(0.229)	(0.226)
High ability teams	1.150	11.83
	(2.200)	(12.43)
*Treatment*High ability teams*		
Plow*High ability teams	-0.0559	-10.19
	(2.841)	(16.17)
Tlow*High ability teams	-4.399	-8.117
	(2.698)	(16.48)
Pmax*High ability teams	-3.977	-29.05*
	(2.562)	(15.33)
Tmax*High ability teams	-5.035	-19.74
	(3.573)	(18.85)
Constant	4.037*	69.33***
	(2.400)	(11.46)
Observations	278	278
Adj. R-squared	0.605	0.465

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

AIII. 22: OLS regression on qualitative performance during the incentive phase split by previous ability of teams

VARIABLES	(1) Qual low ability	(2) Qual high ability
Treatments (Base: Fix)		
Plow	11.27	9.840
	(11.77)	(16.23)
Tlow	35.17***	25.78
	(11.66)	(18.29)
Pmax	35.27***	17.44
	(10.75)	(15.87)
Tmax	60.96***	48.07***
	(12.61)	(17.76)
Experience Droodle	14.10	99.19**
	(20.00)	(44.36)
Effort	1.050	-2.308
	(4.234)	(8.002)
Difficulty	-5.119	-3.351
	(5.187)	(7.635)
Pressure	18.72***	7.535
	(4.241)	(6.298)
IM	22.19**	18.80*
	(8.833)	(10.92)
Group cohesion	20.56**	14.93
	(10.20)	(15.07)
Group identification	-31.15***	1.778
	(7.665)	(11.45)
Gender composition (Base: Only Males)		
Mixed	3.121	-13.35
	(8.292)	(13.58)
Only Females	-13.86	0.0238
	(9.600)	(13.91)
Constant	50.14	74.68
	(44.15)	(77.96)
Observations	141	137
Adj. R-squared	0.279	0.084

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

AIII. 23: OLS regression on quantitative and qualitative performance controlling for heterogeneity in teams

VARIABLES	(1) Quantity	(2) Quality
Treatments (Base: Fix)		
Plow	1.779	8.254
	(1.370)	(8.031)
Tlow	4.133***	30.26***
	(1.348)	(8.153)
Pmax	3.227**	21.35***
	(1.252)	(7.757)
Tmax	9.522***	53.09***
	(1.811)	(9.659)
Team-Check	2.361***	1.716***
	(0.149)	(0.137)
Diff. Equal contribution	-1.269**	-8.047***
	(0.569)	(3.028)
Diff. Group identification	-0.0788	-0.00510
	(0.714)	(4.030)
Diff Group cohesion	0.00149	-0.0140
	(0.00485)	(0.0147)
Diff. Group advantages	0.187	1.736
	(0.961)	(5.291)
Diff. Leader present	0.0188	0.447
	(0.488)	(2.827)
Constant	7.516***	82.78***
	(2.239)	(11.71)
Observations	278	278
Adj. R-squared	0.604	0.471

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

AIII. 24: OLS regression on qualitative performance controlling for socio-demographic differences between team members

VARIABLES	(1) Gender	(2) Age	(3) Nationality	(4) Education	(5) Field	(6) Ability
Treatments (Base: Fix)						
Plow	9.241	8.926	8.649	9.275	10.02	9.267
	(7.989)	(8.013)	(8.038)	(8.013)	(7.810)	(7.957)
Tlow	30.68***	30.54***	30.73***	30.80***	31.73***	30.61***
	(8.142)	(8.139)	(8.237)	(8.224)	(8.163)	(8.265)
Pmax	22.73***	22.66***	20.99***	22.82***	23.89***	22.66***
	(7.774)	(7.754)	(7.786)	(7.835)	(7.666)	(7.768)
Tmax	52.57***	52.35***	52.58***	52.56***	52.05***	52.56***
	(9.370)	(9.403)	(9.310)	(9.405)	(9.422)	(9.332)
Team-Check	1.704***	1.702***	1.672***	1.698***	1.694***	1.695***
	(0.138)	(0.136)	(0.134)	(0.136)	(0.135)	(0.137)
Diff. in Gender	1.737					
	(5.949)					
Diff. in Age		-0.296				
		(0.346)				
Diff. in Nationality			-10.22*			
			(5.711)			
Diff. in Education				1.127		
				(5.658)		
Diff. in Field of Study					-8.952	
					(5.702)	
Diff. in Familiarization						0.0204
						(0.221)
Constant	76.50***	78.87***	82.75***	76.82***	82.21***	77.33***
	(9.746)	(9.399)	(9.182)	(9.642)	(9.699)	(9.086)
Observations	278	278	278	278	278	278
Adj. R-squared	0.467	0.468	0.472	0.467	0.472	0.467

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

AIII. 25: Instructions

1. Start Page

Welcome to the experiment!

Thank you for taking the time to participate. The following pages will present the instructions and explain the task.

Please click „Next“ now and read the following information carefully.

2. Data Protection & Terms of Participation

3. General Instructions and Example Droodle with Explanation

The experiment consists of several tasks. In each task, you will be shown an image of a so-called Droodle. A **Droodle is an abstract visual puzzle** that you are asked to interpret. You can find an example of such a Droodle below. For each task, you will have **4 minutes of working time** to find **as many different and creative solutions as possible** for the respective Droodle shown. You may use single words as well as complete sentences. The perspective on the Droodle is not relevant, meaning it can be viewed from all four sides. Please type your answers below the Droodle und **separate them from one another by a paragraph break.**

All recorded solutions will subsequently be evaluated by a four-member jury. The jury assesses the **validity and creativity** of all solutions. A solution is considered valid if at least 3 of the 4 jury members recognize it as valid and if it **differs** sufficiently from solutions that were previously mentioned. A solution is recognized if, from the jury member’s point of view, it is discernible in the Droodle. A sufficient distinction from previous solutions exists if the solution offers a new interpretation of the Droodle. In the example Droodle below, for instance, „duel“ and „fencing“ would not be considered sufficiently different solutions. In this case, only the solution listed higher would be counted.

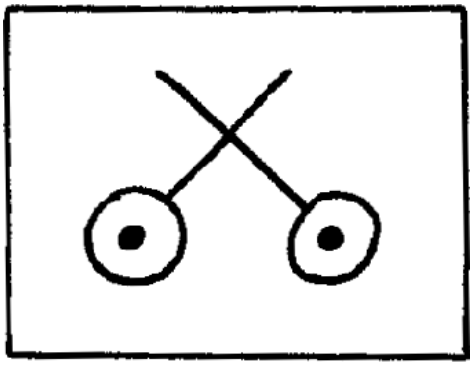
Each **valid solution is awarded 5 points.** All valid solutions are additionally evaluated in terms of their creativity. For this purpose, **each** jury member awards **0–3 points per valid solution**. Here, 0 points represent a particularly uncreative and obvious solution, and 3 points represent a particularly creative and unique solution. Depending on the creativity ratings given by the four jury members, **a total of 5–17 points** can be achieved **per valid solution**. For example, if one of your answers is classified as valid and receives 1 point from one juror and 2 points from each of the other three jurors, you will

achieve a score of 5 + 1 + 2 + 2 + 2 = 12 out of 17 points. The overall performance per task is the sum of the points achieved.

For participating in the experiment as well as for the first two tasks, you will receive a flat compensation of 100 Taler. One Taler corresponds to a payment of 4 cents. For completing additional Droodle, you will receive **additional compensation**, which will be announced before the beginning of each round.

After completing the experiment, we will ask you to answer a questionnaire. Payment will be made via bank transfer to the bank details you provided after all solutions have been evaluated. This may take several weeks, as the evaluation must be carried out manually and the transfers require time.

Please click "Next" once you have understood the instructions and follow the subsequent directions.

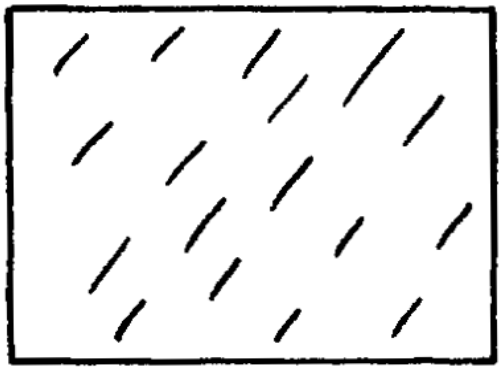


Sample Droodle

4. Start-Droodle

5. "Waiting Page"

6. Droodle **Individual**; automatic redirection after 4 minutes; text field



7. Instructions: Droodle Team-Check

All treatments

You will now work on the next tasks together **in pairs as a group. The person sitting directly next to you is your other group member.**

Agree on which group member will type in the solutions and sit together at the corresponding workstation.

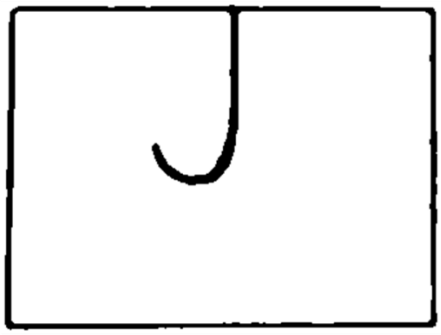
You have 4 minutes to find as many and as creative solutions as possible for the following Droodle.

The communication ban is lifted for collaboration with your group member.

Please click "Next" when you are ready.

8. "Waiting-Page"

9. Droodle-**Team-Check**; automatic redirection after 4 minutes; text field



10. Instructions: 1. Droodle with Treatmentvariation

Fix

You will continue to work on the next three Droodle **together in pairs** as a group. For each Droodle, you have 4 minutes to find as **many and** as **creative** solutions as possible.

For completing the three Droodle, **your group will receive a fixed payment of 300 Taler**.

This amount will be **divided equally between the group members**.

Please click "Next" when you are ready.

Plow

You will continue to work on the next three Droodle **together in pairs** as a group. For each Droodle, you have 4 minutes to find as **many and** as **creative** solutions as possible.

For completing the three Droodle, **your group will receive a fixed payment of 150 Taler.**
In addition, **for one of the three Droodle**, you will receive a **piece-rate payment of 1 Taler per point achieved**. **For each valid solution, 5-17 points can be earned. Only one of the three tasks** will be **randomly** selected after the experiment to be **payout-relevant**. Only the points achieved in this selected task will determine your additional payment. The total amount earned (base payment + any additional payment) will be **divided equally between the group** members after the evaluation.
Please click "Next" when you are ready.

Tlow

You will continue to work on the next three Droodle **together in pairs** as a group. For each Droodle, you have 4 minutes to find as **many and** as **creative** solutions as possible.
For completing the three Droodle, **your group will receive a fixed payment of 150 Taler.**
In addition, **your group may receive an additional payment for one of the 3 Droodle**. This depends on whether your group achieves a higher total score than **another randomly selected group**. The total score is calculated based on the points earned for valid solutions. **For each valid solution, 5-17 points can be earned.** If your group achieves a **higher total score than the other group**, **your group will receive an additional payment of 450 Taler**. If your group achieves a **lower total score**, **no additional payment** will be awarded.
Only one of the 3 tasks will be **randomly** selected after the experiment to be **payout-relevant.** Only the points achieved in this selected task will determine the possible additional payment! The total amount earned (base payment + any additional payment) will be **divided equally between the group** members after the evaluation.
Please click "Next" when you are ready.
Please remember that the group with the higher total score will receive an additional payment of 450 Taler, while the group with the lower total score will receive no additional payment.

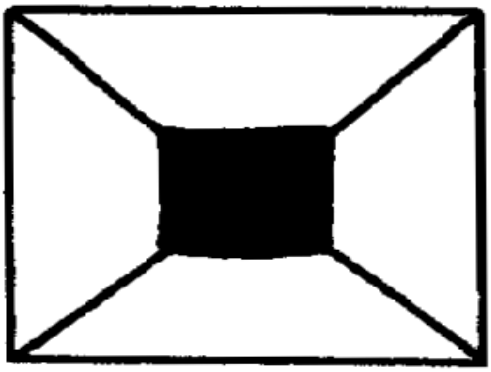
Pmax

You will continue to work on the next three Droodle **together in pairs** as a group. For each Droodle, you have 4 minutes to find as **many and** as **creative** solutions as possible.

Your group will receive a **payment of 2 Taler per point** achieved **for one of the 3 Droodle**. **For each valid solution, 5-17 points can be earned. Only one of the 3 tasks** will be **randomly** selected after the experiment to be **payout-relevant.** Only the points you achieve in this task will determine the payment! The total amount earned (base payment + any additional payment) will be **divided equally between the group** members after the evaluation.
Please click "Next" when you are ready.

Tmax
You will continue to work on the next three Droodle **together in pairs** as a group. For each Droodle, you have 4 minutes to find as **many and** as **creative** solutions as possible. Your group may receive a **payment for one of the 3 Droodle**. This depends on whether your group achieves a higher total score than **another randomly selected group** within that Droodle.
The total score is calculated based on the points earned for valid solutions. **For each valid solution, 5-17 points can be earned.** If your group achieves a **higher total score than the other group**, **your group will receive 600 Taler**. If your group achieves a **lower total score**, **no payment** will be awarded. **Only one of the 3 tasks** will be **randomly** selected after the experiment to be **payout-relevant.** Only the points you achieve in this task will determine the payment! The total amount earned (base payment + any additional payment) will be **divided equally between the group** members after the evaluation.
Please click "Next" when you are ready.
Please remember that the group with the higher total score will receive a payment of 600 Taler, while the group with the lower total score will receive no payment.

11. "Waiting-Page"
12. 1.Droodle-Team; automatic redirection after 4 minutes



13. Instructions: 2. Droodle with Treatmentvariation

14. "Waiting-Page"

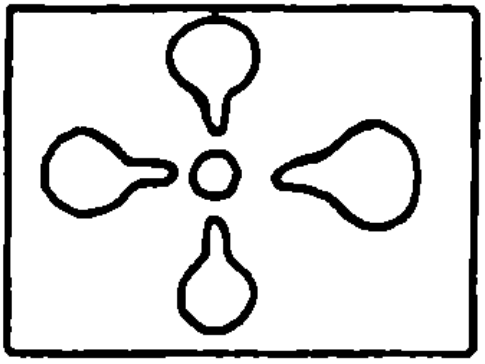
15. 2. Droodle-Incentive, automatic redirection after 4 minutes; text field



16. Instructions: 3. Droodle with Treatmentvariation

17. "Waiting-Page"

18. 3. Droodle-Incentive, automatic redirection after 4 minutes; text field



19. Intermediate Page: Tasks completed

20. Questionnaire

21. Payment Page

AIII. 26: Questionnaires

Italics: Note on the intention of the questions/measured facts

Scale: 1-5 (unless otherwise specified); Strongly disagree; Disagree somewhat; Neither agree nor disagree; Agree somewhat; Strongly agree

Questionnaire 1 – All treatments

Assessment of own performance (Droodles are shown again)

Finding the solution for Droodle 1 was difficult.

My performance in working on Droodle 1 was above average.

Finding the solution for Droodle 2 was difficult.

My performance in working on Droodle 2 was above average.

Finding the solution for Droodle 3 was difficult.

My performance in working on Droodle 3 was above average.

Finding the solution for Droodle 4 was difficult.

My performance in working on Droodle 4 was above average.

Finding the solution for Droodle 5 was difficult.

My performance in working on Droodle 5 was above average.

Effort

I worked hard on editing the Droodle.

Pressure

I felt under pressure while working on the Droodle.

Difficulty

Working on the Droodle was difficult overall.

Intrinsic Motivation (Interest-Enjoyment of the IMI)

I really enjoyed doing the tasks.

These tasks were fun.

I found the tasks boring.

The tasks did not hold my attention at all.

I would describe these tasks as very interesting.

I found these tasks quite enjoyable.

While doing these tasks, I thought about how much I liked them.

Knew each other beforehand

I already knew my other group member beforehand.

No; Yes

Perceived advantages of the group

Working in a group was more fun than working alone.

I felt pressured by the member of the group.

Communicating within the group gave rise to ideas that no member would have come up with on their own.

Group Cohesion & Collaboration (Chen et al. 2012 based on GEQ)

Overall, I liked the member of my group.

Our group was united in our attempt to develop creative ideas.

My group did not give me enough opportunities to contribute my ideas. (R)

I did not like the working style in this group. (R)

I feel connected to the group.

I feel quite similar to the others in my group in terms of general attitudes and opinions.

I think my group had greater cohesion (unity) compared to other groups in this study.

The members of my group were open to each other's ideas.

Group Identification (Chen et al. 2012 based on the Situation Motivation Scale (Guay et al. 2000)

I think this activity is important for me.

I think this task is important for my group.

I thought about my creative performance in the task.

My group thought about their creative performance in the task.

Classification of performance in the group

I worked harder in the group than I would have alone.

All group members contributed equally to the results.

Leader

One group member took on the role of a leader.

Big5 (John & Soto, BFI-2-XS)

Below you will find a series of personal characteristics that may apply to you. Please indicate the extent to which you agree with each of the following statements.

I am rather quiet.

I am empathetic and warm-hearted.

I am rather messy.

I often worry.

I am enthusiastic about art, music, and literature.

I tend to take the lead.

I am sometimes rude and abrupt.

I tend to procrastinate.

I am often depressed and downhearted.

I am not very interested in abstract ideas.

I am full of energy and drive.

I trust others easily and believe in the good in people.

I am reliable; people can count on me.

I am well-balanced and not easily upset.

I am original and develop new ideas.

Cognitive Flexibility (Martin & Rubin, 1995)

I can convey an idea in many different ways.

I can find practical solutions to seemingly unsolvable problems.

I am willing to work on creative solutions to problems.

I am able to act appropriately in any situation.

My behavior is the result of conscious decisions I make.

I have many options for how to behave in a given situation.

I am willing to listen and consider alternatives for dealing with a problem.

I have the confidence to try out different behaviors.

Risk-taking (Single-Item R-1; Beierlein et al., 2014)

How would you personally assess yourself: How willing are you to take risks in general?

Scale 1-7: 1= not at all willing to take risks; 7= very willing to take risks.

Locus of Control (short scale, according to Rotter (1966))

Scale: 1-5; does not apply at all; applies slightly; applies somewhat; applies quite a bit; applies completely

I am in control of my own life.

If I work hard, I will be successful.

Whether in my private life or at work, my life is largely determined by others.

My plans are often thwarted by fate.

Creative inclination (Burk & Wiese, 2018)

I tend to lose track of time when I am engaged in creative work.

I often become enthusiastic when I am working on a new idea.

I enjoy playing with new ideas for the sheer pleasure of it.

When I have a new idea, I am completely absorbed in it until I have pursued it to its conclusion.

I enjoy looking for new ideas outside my field of expertise and trying to apply them to my field.

I like to fantasize about how things could be done better in my everyday work.

When I look at existing products/processes, I usually evaluate them critically to see how I can improve them.

New concepts are essential for good decision-making and should therefore be highly valued.

People who want to constantly develop new ideas often disrupt the workflow in a work group.

If you don't generate enough new ideas, this is the reason you are not competitive.

I would rather pursue an unusual idea than try to please everyone.

When I make a suggestion, it would be particularly bad for me if my colleagues thought it was nothing new or original.

I often come up with ideas or solutions that other people would never have thought of.

I am good at combining different concepts in ways that others have not yet tried.

I regularly succeed in coming up with solutions to problems that no one else could have imagined before.

When I am faced with a problem, a lot of ideas and possible solutions immediately come to mind.

Experience Creativity

Have you previously participated in laboratory experiments in the field of creativity?

Yes; No

Have you previously worked with Droodles?

Yes; No

Questionnaire 2 – Tournament Treatments

Scale 1-5; Strongly disagree; Disagree; Neutral/No opinion; Agree; Strongly agree

Perception of the tournament

I worked harder in the competition between the groups than in previous rounds.

I felt pressured by the competition between the groups.

I preferred working in competition between the groups to previous rounds.

Assessment of own performance (Measure for balanced competition)

What do you think: How high are your chances of winning the tournament?

Slider 0-100%

Competitive Orientation (Burk & Wiese, 2018)

I am hardly affected when my colleagues/fellow students achieve greater success than I do.

It motivates me greatly to compete with colleagues/fellow students and see who comes out on top in the end.

I try to avoid competitive situations at work.

I don't mind giving someone credit for something I could have done just as well or better.

In teamwork, where only the overall result is visible at the end, it bothers me when my individual performance is no longer recognizable.

I think it's good when there is healthy competition within a team.

When another person in my team is honored/praised for their success, it spurs me on to be recognized as the more successful one at the next opportunity.

I like the feeling of strength when others acknowledge that I have presented superior arguments in a discussion.

Questionnaire 3 – All treatments

Socio-demographic

What is your gender?

Male; female; diverse

How do you generally see yourself? Where would you place yourself on this scale from "0 - very masculine" to "10 - very feminine"?

(0-10; very masculine – very feminine)

How old are you?

XXX years old

What is your nationality?

German; EU (outside Germany); Non-EU

What is your highest level of education?

High school diploma; Vocational training; University entrance qualification (Abitur); Bachelor's degree; Master's degree; Doctorate

What subject are you studying / what field are you working in?

Economics; Natural sciences; Language and cultural studies; Teaching; Agricultural and Forestry Sciences; Engineering; Social sciences; Art, Music, Design; Health sciences; Public administration

Does/did your degree program have an economic focus?

Yes/No

PART V: CREATIVITY IN TEAMS – THE ROLE OF GENDER AND INCENTIVES

Information: The following study is an imminent expansion of Essay 3 towards the connection between the gender composition of teams and PfP incentives in context of creativity, whereas both essays developed independently of each other and at different times. In both essays the same experimental design and procedure are used. As the original baseline sample of Essay 3 was not gender-balanced and too small to statistically test gender effects within a treatment, further data collection was necessary to conduct Essay 4. As the following essay does not provide harsh evidence for an interaction effect of incentive schemes with gender composition of teams, no sperate paper for publication on this topic is written. Therefore, the larger sample size induced by this follow-up study is also used for Essay 3 and a shorter paper, combining essays 3 and 4, to strengthen its data base.

This declaration is written to preempt speculation about a non-academically desirable division of one project into several publications.

Abstract: Remarkably, women are underrepresented in the creative field, even though no evidence for creative potential being gender dependent exists. One potential limiting factor can be the competitive setting of the innovative field, often being characterized by tournaments. The present study is the first one which systematically investigates different incentive types and intensities and their potential interaction with gender composition of a team on creative performance. Using a large-scale laboratory experiment, no significant interaction effects of teams being of different gender composition are detected. Interestingly, purely female teams outperform those with at least one male during an unincentivized task. This effect diminishes as soon as incentives are introduced, independent of incentive type or intensity. Thus, the presented study is of highly practical relevance, indicating a beneficial effect of monetary team-based incentives for the whole workforce of an organization, independent of gender composition of a team.

1 Introduction

Non-routine work, i.e., creative thinking, is expected to become one of the most relevant skills in the labor market in the upcoming years (World Economic Forum, 2025). Creativity is here defined as the production of novel and useful ideas (Amabile, 1988). Harnessing the creative potential of the full workforce is therefore not merely a managerial aspiration but an economic imperative. Despite this, the creative and innovative workforce remains markedly skewed by gender. Of all listed inventors worldwide, only 13% are women (WIPO, 2023) and women in academia publish significantly less than men (Iaria et al., 2022). This imbalance raises the question whether structural features of creative work environments may act as barrier for women, and if so, which feature this can be.

One salient characteristic of the creative work environment is its competitive nature. Logo design competitions, public architecture tenders, and innovation challenges are a few settings in which creative output is explicitly incentivized and ranked (Morgan & Wang, 2013; Terwiesch & Ulrich, 2009). Previous research on routine tasks indicates varying preferences for competitive environments induced by tournaments between men and women (e.g., Datta Gupta et al., 2013; Dohmen & Falk, 2011; Gneezy et al., 2003; Niederle & Vesterlund, 2007). Importantly, however, differential preferences for competition do not necessarily translate into differential performance outcomes (e.g., Ivanova-Stenzel & Kübler, 2011; Städter et al., 2022).

Another structural characteristic of many creative settings is the implementation of small teams working on creative tasks. Recent evidence consistently points to a performance advantage of teams over individuals on creative tasks (Grund et al., 2026; Kozlowski & Bell, 2013; Kurtzberg & Amabile, 2001). Teams appear to benefit from shared thoughts during direct exchange of ideas (Feinstein, 2011). Therefore, teams are in focus for studying creative performance.

The present study focuses on divergent thinking, the early generative stage of the creative process in which a broad range of creative ideas is produced. Divergent thinking serves as a starting point for most innovative processes and is therefore central for reaching high levels of innovation (Reiter-Palmon et al., 2019). Divergent thinking is episodic rather than continuous: it constitutes one component of working life that alternates with other activities, and the incentive structures governing it may therefore change across projects or time periods. Employees typically cannot self-select into preferred incentives for several work tasks separately; they must perform under the given compensation rules. Therefore, this study does not aim to focus on

preferences and consequently self-selection process in a competitive creative environment but on performance under exogenously imposed incentive structures

Among the instruments available to organizations for steering creative output, PfP schemes are particularly prominent, like for example piece-rates and tournaments between teams. Both schemes can vary in their intensity, which is defined as the payment that is directly performance dependent, such that a PfP scheme with a high intensity has only a relatively small or no positive fixed component. While the broader literature on incentive effects for routine tasks is well developed (e.g., Bull et al., 1987; Nalbantian & Schotte, 1997; Opitz et al., 2024; DellaVigna & Pope, 2018), evidence on creative tasks is comparatively sparse, and evidence involving teams (e.g., Kachelmeier et al., 2023) is scarcer still.

Against this backdrop, a critical research gap emerges: the interplay between the gender composition of teams and the incentive effects of PfP schemes on creative performance has not been systematically investigated. Gender differences in responses to competitive incentives at the individual level provide theoretical grounds to expect that the gender composition of a team may moderate how different incentive structures affect a team's creative performance. Understanding this relationship is practically significant: it has direct implications for the design of gender-sensitive incentive systems in organizations, like private forms but also universities, that aim to foster creative performance while promoting equal participation.

This study addresses two interrelated research questions:

1. *Do pay-for-performance incentives affect the creative performance of teams differently than fixed pay, depending on the gender composition of the team?*
2. *Does this effect depend on the type and intensity of the incentive?*

By examining piece-rate and tournament structures and their interaction effects across differently composed teams, this paper contributes novel evidence to the literature on gender, incentives, and creativity, and offers practical guidance for the design of work environments that allow all employees to realize their creative potential.

2 Theoretical background and literature review

The following section presents the relevant strands of literature and theory for investigating the role of gender composition in creative teams in context of PfP incentive schemes, focusing specifically on the intersection of gender and creative performance and the interaction between gender and PfP schemes. For a comprehensive review of the literature on creative teams and PfP incentive schemes more broadly, refer to section 2 in Part IV or section 3.2 in Part I.

A fundamental observation for this study is that creativity constitutes a male-stereotypical task (Hora et al., 2021; Proudfoot et al., 2015), despite the absence of significant differences in creative potential between men and women at the individual level (e.g., Baer & Kaufmann, 2008; Runco et al., 2010). This contradiction is best understood through the lens of SRT and RTC.

SRT posits that men and women are ascribed distinct normative characteristics by society, whereas men are associated with agentic traits, i.e., independence, dominance, and assertiveness, while women are associated with communal characteristics, i.e., cooperativeness, caring, and a need for harmony (Eagly, 1997; Franke et al., 1997; Hora et al., 2021). Building on this foundation, RCT further stipulates that individuals are evaluated favorably when their perceived characteristics align with their assigned social role, and face penalties when they deviate from it (del Carmen et al., 2023; Eagly & Karau, 2002). As creative behavior is coded as agentic, for example it is connected with being adventurous, courageous and risk taking (Proudfoot et al., 2015), and therefore male-typed, RCT predicts systematic disadvantages for women in how their creative output is recognized and rewarded (Hora et al., 2021). Consistent with this prediction, women in creative fields must demonstrate significantly greater talent (Lubczyk & Moser, 2025) and exert more effort than comparable men to receive equivalent recognition (Proudfoot et al., 2015; Ritter & Yoder, 2004). Anticipating such penalties for role incongruity, women may strategically withhold effort on creative tasks (Hora et al., 2021), in mixed-gender team settings (Hora et al., 2021; Ritter & Yoder, 2004), and in competitive environments more generally (Schock et al., 2019).

Moving to the team level, team members are expected to complement each other during the creative process, using different perspectives and thought processes (Feinstein, 2011). Consequently, the assumption exist that more diverse team members can complement each other, through for example, different experiences of men and women induced by society and cultural background (Crenshaw, 2013; West & Zimmerman, 1987). Thus, potentially enabling higher creative performance of more diverse teams. Empirical evidence, however, does not consistently support this intuition. Dutcher and Rodet (2022) find no significant effect of team diversity along dimensions of gender, origin, or income on creative output. More strikingly, Rodet (2025) reports that mixed-gender teams exhibit negative effects on originality and uniqueness of creative ideas relative to single-gender teams. Further research yields benefits only when genuine inclusion occurs, such as when a woman is meaningfully incorporated into a male-dominated group rather than remaining a token presence (Vedres & Vásárhelyi, 2023).

The role of SRT and RCT becomes particularly salient in team settings, where gendered social norms are activated through interpersonal dynamics. In mixed-gender groups, men are found to contribute more on male-typed tasks, while this stereotype-driven behavior does not emerge in single-gender groups (Chen & Houser, 2017). Peter et al. (2021) find in unincentivized divergent thinking tasks conducted in groups of three or four that purely female teams exhibit higher fluency. Furthermore, gender differences are most pronounced in "solo" configurations, where a single member is of a different gender, thereby creating faultlines that disrupt team cohesion and performance.

The implications of RCT extend beyond the social dynamics of team composition to the domain of competitive incentive structures. Competitive PfP schemes, such as tournaments, create performance-oriented environments that are stereotypically congruent with male agentic traits, thereby compounding the penalties women face for role incongruity (Eagly & Karau, 2002). It is important to distinguish between preferences for and performance under tournaments. While gender differences in preferences for competition are well-documented, in male-typed domains (Niederle & Vesterlund, 2007, 2011), this does not hold for verbal tasks (Markowsky & Beblo, 2022). Moreover, differences are partly attributable to women's reluctance to compete against men specifically (Buser et al., 2023; Geraldes, 2020). The present study, as already explained earlier, focuses on performance under such incentive schemes, rather than preferences for tournaments, as divergent thinking tasks are found to be short-term work task, thus, not always predictable and therefore employees are not able to opt-out of such a task, even though competitive incentives are used and not preferred.

Evidence from individuals performing routine tasks shows that men outperform women under competitions (Gneezy et al., 2003; Niederle & Vesterlund, 2011), especially in male-stereotyped tasks (Niederle & Vesterlund, 2011; Shurchkov, 2011; Städter et al., 2022), which creativity is (Proudfoot et al., 2015). In the specific context of individual creative performance under competition, creativity of girls is found to be undermined by competition (Conti et al., 2001). Similarly, this is found in context of solving computer mazes (Gneezy et al., 2003). On explanation is that women are found to engage in creative tasks through empathy and men through risk-taking which is coupled to performance under tournaments (Hvide, 2002; Kim et al., 2024).

As creative teams are the center of the present study, it is important to consider the relevance of teams in such competitive environments. On the one hand, teams may serve as a buffer against the pressures of role incongruity, particularly when women can collaborate with other women and are thus shielded from direct comparison with male counterparts (Dargnies, 2012;

del Carmen et al., 2023; Ivanova-Stenzel & Kübler, 2011). Furthermore, female participants have been shown to prefer competing in teams rather than as individuals (Healy & Pate, 2011; Niederle & Vesterlund, 2011), suggesting that the team setting may partially attenuate gender-based competitive disadvantage. On the other hand, teams do not uniformly mitigate these dynamics: even in team contexts, women may underperform if they perceive themselves as competing against men, for instance, when opponents in a laboratory setting include male participants or when one's own team partner is male (Buser et al., 2023; Geraldes, 2020). Notably, Dargnies (2012) finds no evidence that team-based competition increases women's preferences for tournaments, qualifying the more optimistic view of teams as equalizing environments.

Gender-based faultlines pose a further challenge in mixed-gender teams under competitive conditions. Such faultlines may impede the cooperation and coordination that are essential for team creativity (Baer et al., 2014; C.X. Chen et al., 2012; Gächter et al., 2025; Grund et al., 2025). Generally speaking, existing findings on the role of gender composition of teams on their performance remain inconclusive (Apesteguia et al., 2012; Hoogendoorn et al., 2013; Ivanova-Stenzel & Kübler, 2011; Städter et al., 2022). One notable exception is the finding that ranking-based tournament incentives are particularly effective for male-dominated teams (Englmaier et al., 2024b), suggesting that the interaction between team gender composition and incentive scheme type is not uniform across competitive structures.

While the literature on gender and creativity, gender and team composition, and gender and competitive incentives has developed considerably, their intersection remains substantially underexplored. In particular, no study to date has jointly examined how team gender composition moderates' creative performance across different PfP incentive schemes in a controlled experimental setting. The present study addresses this gap. Given the limited and ambiguous state of existing evidence, partly contradicting SDT and RCT, the analysis is predominantly exploratory in nature, and formal hypotheses are replaced by the following three expectations.

First, with respect to mixed-gender teams, two competing mechanisms generate opposing predictions: on one hand, gender diversity may expand the team's available skill and perspective set, potentially positively affecting creative performance (Herring, 2009; Lee et al., 2025; Yang & Konrad, 2011); on the other, it may give rise to interpersonal conflict rooted in dissimilarity (Herring, 2009). Moreover, focusing on gender difference in competitiveness and competitive behavior, assuming a higher level of competitiveness of males (Niederle & Vesterlund, 2011), two options are possible: First, the higher competitiveness of a male could transfer to the female team partner, increasing their level of competitiveness, and thus, the general competitiveness

of a team (Ivanova-Stenzel & Kübler, 2011). Second, gender stereotypes could be reinforced when different levels in competitiveness in a female and a male reveal themselves during active teamwork, potentially inducing lower competitiveness of the female (del Carmen et al., 2023; Eagly & Karau, 2002). Given the counterbalancing nature of these forces, no net effect of gender diversity on creative output is expected relative to single-gender teams.

Second, although creative tasks are generally characterized as male-stereotyped, stereotype-driven dominance effects have not been found in single-gender groups (Chen & Houser, 2017). Accordingly, no systematic difference in creative performance between exclusively male and exclusively female teams is anticipated.

Third, drawing on evidence from individual and team-level research on routine tasks, exclusively male teams are expected to outperform mixed-gender and exclusively female teams under tournament incentives, given men's comparatively stronger response to competitive incentive structures in male-typed tasks (Englmaier et al., 2024b; Geraldes, 2020; Healy & Pate, 2011). By contrast, no significant differences between team gender compositions are expected under piece-rate or fixed-payment incentive schemes.

The present study contributes to the growing literature on creative team performance, and to the ongoing debate concerning the interaction between PfP incentives and creative performance, by extending insights from the well-documented gender gap in routine task performance under competition to the underexplored domain of creative teamwork.

3 Experimental design

The design of the conducted lab experiment, the instructions, and the questionnaires are identical to the design of Essay 3 (refer to Appendix III AIII. 25 for instructions and AIII. 26 questionnaires). That is five treatments with varying incentive intensity and incentive type (i.e., piece-rates and tournaments) are implemented after a first individual and an additional team-task. To measure creative performance in a divergent thinking task, the Droodle-Task is used. Special focus is laid on a gender-balanced sample, such that around 15 observations per treatment and gender constellation, i.e., only male, only female, and gender mixed teams, is targeted. This leads to a 5x3 between-subject design.

Table 28 below shows the experimental design, now considering the gender composition and also containing payment information to echo the content of Essay 3, without giving a full repetition.[68]

Table 28: Experimental procedure

Treatment / Timing	Fix	Plow	Tlow	Pmax	Tmax
	Instructions				
Individual	Individually				
Team-Check	Team				
Incentive Phase (3 Droodles)	Team x Fix (12€)	Team x Piece with fixed component (6€+0.04€*points)	Team x Tournament with loser prize (6€; 18€)	Team x Pure Piece (0.08€*points)	Team x Winner-takes-all Tournament (0€; 24€)
	a) Two females b) Two males c) Mixed	a) Two females b) Two males c) Mixed	a) Two females b) Two males c) Mixed	a) Two females b) Two males c) Mixed	a) Two females b) Two males c) Mixed
	Questionnaire				

The answers provided have been evaluated by four jurors, two males and two females, trained with results of a pre-test. The jurors decide individually whether an answer is valid or not, determining quantity and number of errors made. If an answer gets classified as valid, jurors decide how creative it is by assigning between 0 and 3 points. The sum of those points determines the quality of an idea. Ex-post the frequency with which an answer occurred is determined as measure for originality of ideas.

4 Data collection, sample and descriptive statistics

The data has been collected between March 2025 and February 2026 at the lab of RWTH Aachen University. The sample is the same as in Essay 3 and consists of 278 dyadic teams, whereas 303 males and 253 females participated. Table 29 shows the allocation of pure male, mixed and pure female teams per treatment. In total, pure female teams are relatively less present in the current sample, due to the fact that the university at which the sample has been collected is characterized by a focus on STEM fields, thus, is dominated by male students. Still, results of power analysis using G-Power for Mann-Whitney U-Tests indicate a sufficient sample size of

[68] The specific gender-centered research questions, experimental design, and treatments are preregistered at *AsPredicted* (AsPredicted #273439).

28 pairs for one comparison (effect-size 0.5, error probability=0.05, power=0.8). Thus, the sample size of a minimum of 14 only female teams is sufficient for statistical testing of significant differences.

Table 29: Sample by gender composition and treatment

	Treatment					
Gender composition	Fix	Plow	Tlow	Pmax	Tmax	Whole Sample
Only Males	21	21	19	20	23	104
Mixed	19	20	20	17	19	95
Only Females	14	15	16	18	16	79
Whole Sample	54	56	55	55	58	278

Table AIV. 1 in Appendix IV shows socio-demographic statistics of the sample by gender. No significant difference regarding the age of participants is found (p=0.659). In line with general populational characteristics of Germany (Statistisches Bundesamt, 2026), significantly more male participants are involved in work or studying in the STEM field (p<0.001).

Focusing on several control variables potentially affecting creative performance and team outcomes (Table AIV. 2 in Appendix IV), significant differences between gender shows. Women indicate significantly higher levels of intrinsic motivation (p=0.041), equal contribution during teamwork (p<0.001), group cohesion (p=0.007), group identification (p=0.047). The significantly difference of group cohesion also holds for pure female teams in contrast to mixed and pure male teams (p=0.002). Additionally, pure female teams report a lower presence of a leader (p=0.063). Those findings hint towards a more harmonious and equal team climate in pure female teams.

Interestingly, women show significantly lower levels in cognitive flexibility (p=0.072) and creative inclination (p=0.005), Moreover, qualitative performance during the individual familiarization task amounts to 31.243 points for women and 34.203 points for men on average, whereas this difference is significant (p=0.015). No such difference exists for quantitative performance, whereas the overall average amounts to 7.888 points. Thus, indicating possibly a lower creative interest and potential of women than men participating. Additionally, men show significantly higher risk preferences than women (p<0.001), which is a typical finding (e.g., Siegrist et al., 2002), potentially affecting behavior under competition.

5 Results

The following section presents the key results on the potential interplay between gender composition of teams and PfP incentives on average creative performance and relative changes when introducing incentive schemes.[69] The results are tested on a team-level.

The average qualitative performance in the first unincentivized team task amounts to 61.579 points, not differing between treatments (p=0.895). Interestingly, a significant difference between gender compositions in this task exists, as only female teams show significantly (p=0.005) higher qualitative performance (67.861 points) in contrast to mixed (57.368 points) and only male teams (p=0.041; 60.654 points). Similar holds for quantitative performance during the first team-task (p=0.010; p=0.063).

Table 30: Average creative performance of teams during the incentive phase by treatment and gender composition

	N	Average Creative Performance			
		Quantity	Quality	Originality	Invalid
Whole Sample					
Only Males	104	43.279	205.875	6.702	4.25
Mixed	95	44.663	199.000	6.168	3.347
Only Females	79	48.380	212.595	6.722	3.456
Fix					
Only Males	21	36.952	187.762	6.190	3.476
Mixed	19	36.105	180	4.368	3.895
Only Females	14	37.643	183.429	5.286	2.643
Plow					
Only Males	21	38.333	190.524	6.142	4.571
Mixed	20	37	183.6	5.95	3.35
Only Females	15	38.933	188.533	5.333	2.5333
Tlow					
Only Males	19	39.421	200.947	5.737	3.053
Mixed	20	39.25	202.1	7	2.9
Only Females	16	44.687	228.937	6.812	3.375
Pmax					
Only Males	20	40.9	217.35	7.15	4.3
Mixed	17	37.765	193.176	5.647	1.882
Only Females	18	44.555	219.167	8	5.444
Tmax					
Only Males	23	46.261	230.522	8.087	5.609
Mixed	19	47	236.105	7.789	4.579
Only Females	16	47.438	236.937	7.75	2.875

Table 30 shows average creative performance during the incentive phase, split by gender composition and treatments. Considering the creative performance of different gender compositions

[69] The following analyses will not focus on general effects of PfP incentives on creative team performance, but on the potential role of gender composition on creative performance in context of such incentives. For results on PfP schemes please refer to Essay 3.

independent of treatments, no such effects as in the previous unincentivized team-task are found for qualitative performance (p=0.173), using Kruskal-Wallis test. Surprisingly, significant differences show for quantitative performance (p=0.047) in favor of only female teams. No significant differences in originality (p=0.681) nor number of invalid answers (p=0.299) are found.[70]

Table 31: OLS regression on quantitative and qualitative performance in the team-check task and incentive phase

VARIABLES	(1) Quantity Team-Check	(2) Quantity Incentive	(3) Quality Team-Check	(4) Quality Incentive
Treatments (Base: Fix)				
Plow	-0.325	1.934	-3.354	9.832
	(0.819)	(1.364)	(4.621)	(7.818)
Tlow	-0.0990	4.227***	-3.183	31.43***
	(0.766)	(1.349)	(4.280)	(8.149)
Pmax	0.279	3.440***	1.675	23.79***
	(0.777)	(1.265)	(4.741)	(7.643)
Tmax	0.168	9.538***	-1.552	34.95***
	(0.786)	(1.776)	(4.611)	(8.620)
Team-Ability		2.348***		1.857***
		(0.151)		(0.133)
Gender composition (Base: Only Males)				
Mixed	-0.395	0.024	-3.178	-1.897
	(0.557)	(1.231	(3.261)	(6.761)
Only Females	1.019*	-0.0453	7.141*	-7.810
	(0.614)	(1.179)	(3.675)	(6.828)
Constant	12.80***	6.488**	61.93***	70.25***
	(0.658)	(2.000)	(3.834)	(9.421)
Observations	278	278	278	278
Adj. R-squared	0.002	0.601	0.016	0.501

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

To test whether the observation of a beneficial effect of only female teams on creative performance under the unincentivized team-check holds for the incentivized phase, OLS regression analyses are conducted.[71] Model (1) and (3) in Table 31 depict direct effects of treatments as well as gender composition on quantitative and qualitative performance during the first team-task. Here, no significant treatment effects are found, indicating that the randomization of participants into treatments has been successful. In those two models, a significant positive effect of only female teams in contrast to only male teams working on an unincentivized task shows. This effect is not found in model (2) and (4), assessing direct effects on creative performance

[70] Therefore, the following results will especially focus on quantitative and qualitative performance.
[71] As a dominance of only male teams is expected by theory, only male teams are set as baseline in the following regression analyses, enabling a comparison of mixed and only female teams with those teams.

during the incentive phase, whereas here positive treatment effects show when controlling for a team's ability, especially for the winner-takes-all tournament depicted by *Tmax* by showing the highest coefficients. Thus, the first indication exists that implementing incentives affects creative teams, comparing them to unincentivized scenarios.

As participants receive a fixed payment upfront for participating, the possibility exists that they perceived the team-check task as being part of a fixed payment scheme.[72] Therefore, a separate analysis of the *Fix* treatment is conducted (see Table AIV. 3 in Appendix IV). Similarly to the results above, no significant direct effects of only female or mixed teams in comparison to only male teams on creative performance during the incentive phase are found. Thus, the missing significant effect of only female teams when moving to the incentive phase is also found in a setting with payment that is not performance-dependent, potentially stating a general effect of for example time but also payment on creative performance of dyadic teams with different gender composition.

To detect whether a moderation rather than direct effect of gender composition on the relation between PfP schemes and creative performance exists, interaction effects are calculated (see Table AIV. 4 in Appendix IV). No such general significant robust interaction effects show. Surprisingly, a weakly significant effect shows for the interaction between *Pmax* and mixed gender teams (β= -33.01, p=0.072) on qualitative performance. However, this effect is only weakly significant, and I cannot differentiate between an actual effect and noise as the interaction of mixed gender teams with the other high intensity scheme, i.e., *Tmax*, or the other piece-rate scheme, i.e., *Plow*, is not significant. Moreover, theoretical arguments are not found to support this specific result. Another possible explanation for this potential noise might be power issues when analyzing a 5x3 design. Therefore, regression analyses with new variables of incentive type (piece-rate or tournament) and incentive intensity (low- and high-intensity) with *Fix* being the baseline is introduced. Table 32 shows corresponding results. Similarly, no significant interaction effects for incentive type nor intensity are found. Thus, there is no harsh evidence for the role of the gender composition of teams when implementing PfP incentives in a creative context.

[72] Such upfront payment was necessary, as otherwise behavior of participants under *Tmax* could have been influenced by the perceived risk of leaving the experiment without any payment in case of losing the tournament, which is prohibited by the laboratory's regulations.

Table 32: OLS regression on quantitative and qualitative performance with interaction effects between incentive and gender composition differentiated between incentive type and intensity

	Incentive type		Incentive intensity	
	(1)	(2)	(3)	(4)
VARIABLES	Quantity	Quality	Quantity	Quality
Incentive (Base: Fix)				
Piece-rate	16.12**	25.26**		
	(6.620)	(12.49)		
Tournament	42.11***	37.94**		
	(7.289)	(14.64)		
Low intensity			20.00***	18.47
			(6.847)	(13.19)
High intensity			38.23***	43.56***
			(7.270)	(13.94)
Team-Ability	1.735***	1.744***	1.700***	1.704***
	(0.142)	(0.144)	(0.146)	(0.147)
Gender composition (Base: Only Males)				
Mixed	-1.389	7.154	-0.711	6.812
	(7.169)	(12.95)	(7.225)	(13.05)
Only Females	-6.217	-10.48	-6.331	-10.34
	(7.029)	(12.35)	(7.137)	(12.53)
Incentive x Gender composition				
Piece-rate x Mixed		-22.60		
		(16.29)		
Piece-rate x Only Females		-4.322		
		(15.79)		
Tournament x Mixed		0.874		
		(18.66)		
Tournament x Only Females		14.31		
		(17.23)		
Low intensity x Mixed				-5.347
				(17.05)
Low intensity x Only Females				12.48
				(16.47)
High intensity x Mixed				-13.24
				(18.43)
High intensity x Only Females				-2.331
				(17.07)
Constant	77.25***	74.80***	79.29***	77.39***
	(9.901)	(12.57)	(10.19)	(12.72)
Observations	278	278	278	278
Adj. R-squared	0.454	0.453	0.437	0.432

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

To test potential mechanisms, especially grounded in team behavior, several control variables are tested as moderators. One potential candidate is group cohesion, as women reported significantly higher levels of cohesion than men. Table AIV. 5 and AIV. 6 in Appendix IV show those interaction effects for incentive type and incentive intensity. No significant effects for

team-ability, cohesion, group identification, and felt pressure are found, whereas treatment variables stay constantly significant. Similarly, no significant effects of potential mediations of the effects of mixed and only female teams in comparison to pure males teams on qualitative creative performance are found, focusing on control variables indicating team climate, i.e., group cohesion (0.172, p=0.876; 2.917, p=0.240), group identification (0.396, p=0.819; 1.627, p=0.409), leadership (0.415, p=0.662; 0.514, p=0.701), pressure (0.634, p=0.691; -0.747, p=0.628), difficulty (0.700, p=0.562; -0.003, p=0.990), effort (0.475, p=0.573; -0.171, p=0.734), and intrinsic motivation (0.226, p=0.822; 1.242, p=0. 351).

As incentive schemes do not seem to explain a diminishing significant difference of pure female teams over the course of the experiment, the question arises why some teams change their performance moving from the unincentivized team-check task to the incentive phase. One explanation can be that either purely female teams decrease in performance or teams with at least one male catch-up. To test for this, the measure of relative performance moving from the team-check to the incentive phase is computed.[73] On average, independent of treatments, the increase for only male teams is 1.619 points for quantitative performance and 7.971 points for qualitative performance, for mixed teams 2.488 points and 8.691 points respectively, and for only female teams 2.291 points and 3.004 points respectively. Thus, only female teams do not decrease their average performance, rather the other teams tend to catch-up. Those differences are significant from each other for quality, such that the relative increases are significantly larger for pure male (p=0.036) and mixed teams (p=0.024) in comparison to pure female teams. This does not hold for quantity (p=0.756).

Figure 11 shows the average changes in qualitative performance per gender composition and incentive scheme. Mann-Whitney U-Tests show significant differences between pure male and pure female teams for *Plow* (p=0.017) and *Pmax* (p=0.095), similar for the comparison with mixed gender teams under *Pmax* (p=0.087).

[73] To compute this measure the difference between performance in the team-check task and the average performance per Droodle in the incentive phase is calculated. The average per Droodle is necessary, as the incentive phase consists of three tasks, whereas the team-check task contains only one Droodle.

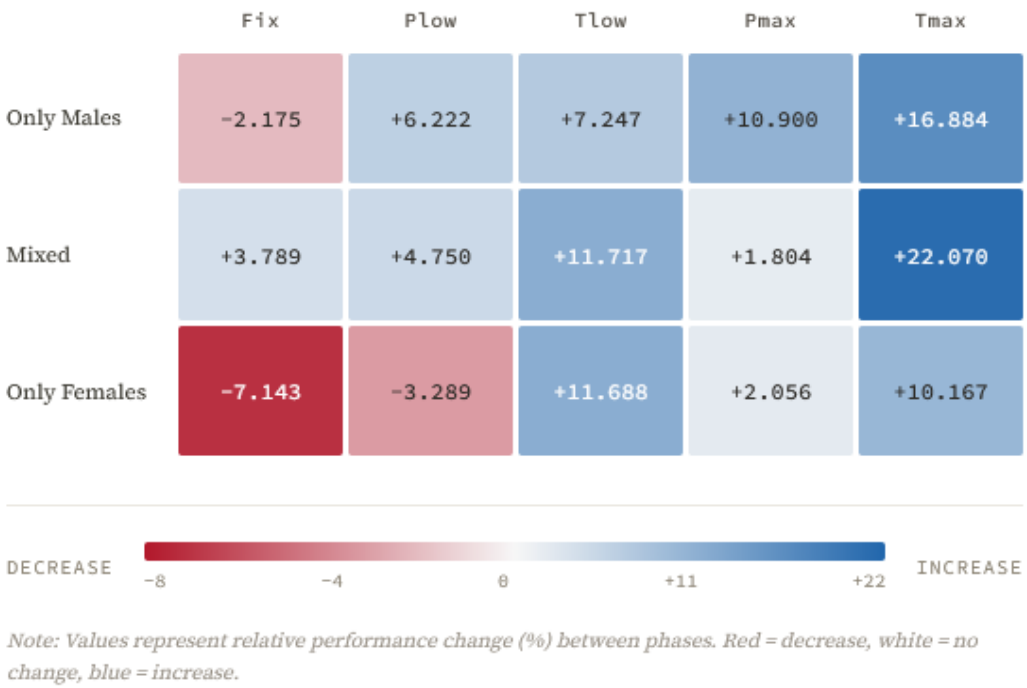


Figure 11: Average qualitative performance differences between the team-check and the incentive phase per Droodle

Testing those findings further using OLS regression analyses and interaction effects between gender composition and incentive types (see Table 33), no significant interaction effects comparing the other gender constellations to pure female teams show, neither for quantity (model (1)) nor quality (model (2)). Conducting the analyses with incentive intensity or treatments as independent variables, significant interaction effects show neither (see Table AIV. 7 and AIV. 8 in Appendix IV). Therefore, regression results do not replicate the findings of Mann-Whitney U-test above that this catching-up behavior might be treatment dependent.

Nonetheless, the result of this catching-up of teams with at least one male is in line with the finding that women tend to show higher performance than men at unincentivized tasks, whereas this difference diminishes if incentives in general are introduced (Segal, 2012). Potentially, as men react stronger towards incentives (McGee & McGee, 2025; Migheli, 2015), whereas women tend to work harder under small rewards (Lewis et al., 2023). Thus, rather than a time-effect, the relative increase can be explained by a general effect of incentives independent of type or intensity, helping men to focus their effort on the task and increase their level of performance.

Table 33: OLS regression on average differences of quantitative and qualitative performance moving from the team-check to the incentive phase with interaction effects between incentive type and gender composition

VARIABLES	(1) Quantity	(2) Quality
Incentive type (Base: Fix)		
Piece-rate	0.779	6.769
	(0.669)	(4.104)
Tournament	2.320***	18.07***
	(0.734)	(4.051)
Gender composition (Base: Only Females)		
Only Males	-0.206	4.968
	(0.805)	(4.270)
Mixed	0.687	10.93***
	(0.734)	(3.869)
Incentive type x Gender composition		
Piece-rate x Only Males	0.993	3.910
	(0.959)	(5.980)
Piece-rate x Mixed	-0.747	-7.162
	(0.929)	(5.777)
Tournament x Only Males	0.180	-3.372
	(1.130)	(6.426)
Tournament x Mixed	-0.240	-5.099
	(1.027)	(5.967)
Constant	-0.810	-7.143***
	(0.561)	(2.741)
Observations	278	278
Adj. R-squared	0.102	0.118

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

To sum up, I find no harsh evidence for an interaction between gender composition of dyadic teams and different PfP incentives on their creative performance. Generally, independent of specific incentives, no significant differences during the main phase of the experiment are found between different gender compositions. Thus, empirical results align with the theoretical and argumentative derived first expectation, which proposed no difference between mixed gender teams and single-gender teams. Interestingly, women perform significantly better than teams with at least one male during the unincentivized task, contradicting the second expectation. Results do not support the third expectation that pure male teams outperform mixed and pure female teams under competitive incentives.

6 Robustness checks

In the following section, several robustness checks for the results above are presented. Firstly, alternative performance measures than average creative performance are analyzed, like the frac-

tion of top-evaluated ideas, variability in creative performance, and produced quality by quantity of ideas. Thereafter, an alternative measure of gender is used. Moreover, a detailed analysis of participants being involved in the STEM field or not is presented, as especially females that select into this field might show different competitive traits than females in non-STEM fields. Lastly, the analyses are completed by a differentiation between low- and high-ability teams and potential differences between those two groups.

Top-evaluated ideas

Next to reaching a high average creative performance level in a divergent thinking task, employers might be interested in detecting team and work characteristics that potentially also produce ideas being especially high in creative quality. In the present essay, such ideas can be identified by looking at the ideas receiving a top score of 11 or 12 points by the jury.[74]

Table 34: Average number of top-evaluated ideas per treatment and gender composition

	Number of Teams with at least one top idea	Top ideas – Team-Check	Top ideas – Incentive Phase	Top ideas – Incentive Phase per Droodle	Top ideas per Droodle over all Teams
Whole Sample					
Fix	36	17	50	16.67	0.309
Plow	41	20	73	24.33	0.434
Tlow	42	11	89	29.67	0.539
Pmax	46	15	87	29	0.527
Tmax	42	14	98	32.67	0.563
Only Males					
Fix	14	10	20	6.67	0.333
Plow	17	6	28	9.33	0.444
Tlow	16	3	30	10	0.474
Pmax	19	6	37	12.33	0.633
Tmax	15	8	35	11.67	0.507
Mixed					
Fix	12	4	18	6	0.316
Plow	12	6	19	6.33	0.333
Tlow	12	3	26	8.67	0.483
Pmax	12	2	21	7	0.412
Tmax	15	2	36	12	0.632
Only Females					
Fix	10	3	12	4	0.262
Plow	12	8	26	8.67	0.556
Tlow	14	2	33	11	0.688
Pmax	15	4	29	9.67	0.519
Tmax	12	4	27	9	0.563

Table 34 shows descriptive statistics of such ideas, split by treatment and gender composition. In total 207 of 278 teams came up with such an idea at least once. Next to the absolute number of top ideas and teams producing such, the number of top ideas per Droodle in the incentive

[74] 12 points are reached if all four jurors rate the creative quality of an idea with 3 points, i.e., the highest score possible. As creativity is at least partly subjective, also 11-point answers are included as being top-evaluated ideas.

phase is given, enabling a comparison between the single team-check task and the incentive phase. Additionally, the fraction of such ideas per Droodle over all teams is presented, thereby correcting for varying sample sizes between pure female, mixed and pure male teams. It gets visible that independent of gender composition, teams under tournament incentives tend to produce the most top-evaluated ideas per Droodle and teams under *Fix* the fewest on average.

Table 35: OLS regression on top-evaluated ideas during the team-check task and incentive phase with interaction effects between incentive type and gender composition

	(1)	(2)	(3)
VARIABLES	Team-check	Incentive phase	Incentive phase
Incentive (Base: Fix)			
Piece-rate	0.00112	0.00872**	0.0119**
	(0.00501)	(0.00340)	(0.00521)
Tournament	-0.00481	0.0131***	0.00875*
	(0.00474)	(0.00345)	(0.00503)
Team-Ability Top Ideas		0.132**	0.139**
		(0.0607)	(0.0615)
Gender composition (Base: Only Males)			
Mixed	-0.00586*	0.000633	0.00267
	(0.00355)	(0.00311)	(0.00522)
Only Females	-0.000951	0.00964**	0.00488
	(0.00424)	(0.00377)	(0.00733)
Incentive x Gender composition			
Piece-rate x Mixed			-0.00907
			(0.00716)
Piece-rate x Only Females			0.000188
			(0.00951)
Tournament x Mixed			0.00385
			(0.00744)
Tournament x Only Females			0.0114
			(0.00950)
Constant	0.0186***	0.0120***	0.0124***
	(0.00458)	(0.00317)	(0.00405)
Observations	278	278	278
Adj. R-squared	0.006	0.070	0.071

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

Table 35 shows regression analyses with the average number of top-evaluated ideas during the team-check task and incentive phase corrected for sub-sample size, e.g., number of only female teams under tournament incentives, as dependent variables.[75] Model (1) and (2) show the direct effects of gender composition in comparison to only male teams on average top-evaluated ideas. No positive direct effect of only female teams in comparison to only male teams in the team-

[75] The dependent variable is a fractional response bounded between 0 and 1, which would generally call for a Fractional Logit estimator (Papke & Wooldridge, 1996). However, given that the mean of the variable in the team-check task (0.148) and in the incentive phase (0.026), thus, well away from the boundaries, OLS with robust standard errors yields consistent and unbiased estimates. Papke & Wooldridge (1996) themselves note that OLS produces similar results in such cases.

check task is found whereas a direct positive effect in the incentive-phase shows (p=0.011) but diminished when interaction effects are added (Model (3)) (p=0.506). Similar holds for low- and high-intensity payment schemes and treatments (see Table AIV. 9 in Appendix IV).

Thus, also for the alternative qualitative measure of creative performance, no evidence for an interplay between incentive type nor intensity and gender composition of teams is found.

Variability in performance

Next to average measures of creative performance, variance in results, potentially indicating steadiness of performance levels, are interesting for interpretation of results. For average qualitative performance independent of incentive scheme, only male teams show significantly higher SDs (70.248 points) than mixed (59.942 points, p=0.060) but not in comparison to only female teams (61.009 points, p=0.567). Regarding the different PfP types, significant differences show between pure male (79.055 points) and mixed teams (51.074 points) under fixed payment (p=0.068). Under tournament incentives, only male teams (75.047 points) differ significantly in comparison to only female teams (44.084 points, p=0.003). The same results hold for quantitative performance, with significant higher SDs for pure male teams (15.648 points) in comparison to mixed teams (9.539, p=0.039) under fixed payment and higher variability in performance of pure male teams (14.610 points) under competitive incentives in comparison to pure female teams (8.77 points, p=0.004).

Thus, some indication exists, that pure male teams show higher variability in qualitative creative performance than other gender compositions if fixed or competitive incentives are into place. This might be of interest if one aims for a high, potentially induced by tournaments, but stable average creative performance.

Creative performance ratios

As only a valid answer can generate qualitative creative points, average qualitative creative performance is partly driven by general quantity. Therefore, in the following the ratio of qualitative performance per generated valid answer is calculated. Additionally, the ratio of originality per quantity as well as the number of errors made per valid answer are computed. Table 36 shows the corresponding results per gender composition and treatment.

Table 36: Average creative performance ratios per gender composition and treatment

	Quality / Quantity	Originality / Quantity	Errors / Quantity
Whole Sample			
Only Males	5.092	0.159	0.096
Mixed	5.084	0.152	0.078
Only Females	4.957	0.154	0.072
Fix			
Only Males	5.073	0.153	0.082
Mixed	4.980	0.123	0.102
Only Females	4.875	0.131	0.058
Plow			
Only Males	4.986	0.159	0.105
Mixed	5.035	0.152	0.075
Only Females	4.840	0.137	0.056
Tlow			
Only Males	5.088	0.139	0.070
Mixed	5.217	0.167	0.070
Only Females	5.113	0.153	0.077
Pmax			
Only Males	5.333	0.172	0.096
Mixed	5.110	0.149	0.049
Only Females	4.898	0.173	0.107
Tmax			
Only Males	4.998	0.168	0.120
Mixed	5.076	0.169	0.089
Only Females	5.049	0.168	0.057

Testing for significant differences on the ratio of quality per quantity using non-parametric tests, weakly significant differences show between pure male and pure female teams under fixed payment ($p=0.085$), pure male and mixed teams ($p=0.096$) and pure female teams ($p=0.002$) under *Pmax*. Whereas for both treatment conditions, pure male teams show higher quality per quantity of answers. Regarding the number of errors made per valid answer on average, significant differences show for pure male and mixed teams under *Pmax* ($p=0.076$) and pure male and pure female teams under *Tmax* ($p=0.027$), whereas only male teams conducted more errors per valid answer. This finding is in line with the results above regarding higher variability in qualitative results of pure male teams, as those teams seem to show at least partially, a higher frequency of errors during the creative process. No significant differences for originality per quantity are found ($p=0.820$).

Self-perceived gender

The reported gender of participants is one of the most important variables for the present study. Until now, a binary variable is used to indicate gender. An alternative measure is the self-perceived gender on a Likert-Scale ranging from 0, very masculine, to 10, very feminine. This self-perceived gender variable correlates strongly with the reported gender by the binary variable ($r=0.856$, $p<0.001$). Only 28 participants diverged from the binary classification, whereas nine

females reported a self-perceived gender lower than five and 19 males reported their self-perceived gender above five. Nonetheless, it is important to notice that the individual understanding of the item might differ. For example, a female participant might understand the question as conditional on the actual gender, i.e., how female does she perceives herself, whereas zero can still indicate a perception of being female but more on the masculine spectrum of females. Additionally, what is understood as being masculine/feminine might differ between individuals.

Table 37: OLS regression on relative difference in qualitative performance moving from the team-check task to the incentive phase with a self-perceived gender measure

	(1)	(2)	(3)
VARIABLES	Type	Intensity	Treatments
Incentive (Base: Fix)			
Piece-rate	17.11***		
	(5.554)		
Tournaments	20.03***		
	(6.173)		
Low intensity		12.81**	
		(5.739)	
High intensity		24.05***	
		(5.803)	
Plow			15.39**
			(6.251)
Tlow			9.273
			(7.510)
Pmax			18.86***
			(6.825)
Tmax			27.80***
			(7.171)
Gender (0-10)	0.228	0.228	0.228
	(0.378)	(0.378)	(0.380)
Incentive x Gender			
Piece-rate x Gender	-1.194**		
	(0.506)		
Tournament x Gender	-0.542		
	(0.537)		
Low intensity x Gender		-0.504	
		(0.512)	
High intensity x Gender		-1.212**	
		(0.518)	
Plow x Gender			-1.133**
			(0.568)
Tlow x Gender			0.216
			(0.641)
Pmax x Gender			-1.255**
			(0.633)
Tmax x Gender			-1.042
			(0.632)
Constant	-3.484	-3.484	-3.484
	(4.278)	(4.278)	(4.310)
Observations	278	278	278
Adj. R-squared	0.099	0.065	0.108

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

Using this alternative measure as gender variable in regression analyses on average qualitative performance (see Table AIV. 10 in Appendix IV), no significant interaction effects are found.[76] Contrary, focusing again on relative performance differences between the team-check task and incentive phase, AIV. 2 shows significant negative interaction effects between increased femininity, i.e., higher values in gender, of a team under piece-rate as well as high-intensity incentive schemes. Thus, replicating indications from above that the catching-up of teams with at least one male or teams identifying as rather masculine on the performance level of pure female teams might be driven by piece-rate incentives.

Participants involved in STEM

As already mentioned earlier, the experiment has been conducted at a top-tier university in STEM fields. Therefore, the sample is characterized by 407 participants being involved in this field of study or work and a potentially higher competitive orientation, especially for females who have actively selected into this non-stereotypical environment. Indeed, females in STEM differ significantly in their competitive orientation (2.30) in comparison to non-STEM females (1.66, p=0.013), no such significant difference exists for males in STEM (2.22) and those that are not (2.32, p=0.796). In total, of 255 females 65.49% and for males 79.73% are involved in STEM. On the team level 8.27% of teams have no member with STEM background, 36.33% have one member, and 55.40% of teams are characterized by both members being involved in STEM.

Table 38: Distribution of teams over gender composition, STEM involvement, and incentive type

	Fix	Piece-rate	Tournament
Whole Sample			
STEM: none	6	13	4
STEM: one	22	44	35
STEM: both	26	54	74
Only Males			
STEM: none	2	4	0
STEM: one	8	11	12
STEM: both	11	16	30
Mixed			
STEM: none	3	3	3
STEM: one	9	17	10
STEM: both	7	17	26
Only Females			
STEM: none	1	6	1
STEM: one	5	16	13
STEM: both	1	13	18

[76] Due to potential differences on what is perceived as being masculine/feminine and the option to interpret this question conditional on ones one gender, as well as such less deviations of the binary indicated gender by participants, no gender-split in pure male, mixed, and female teams is conducted, rather the sum of gender points stated by the team is used.

Table 38 shows the distribution of participants being involved in STEM over gender composition and incentive type, illustrating an unbalanced distribution in the sample, thus no regression analyses with interaction between incentive schemes, team members being involved in STEM, and gender composition can be conducted as the number of observations is partly too little or nonexistent. For example, only one pure female team with no involvement into STEM participated under tournament incentives.

Additional regression analyses (Table 39) on a sample-split between teams with none, one, and both members being part of STEM and their gender composition are conducted both for incentive type (model (1), (3), and (5)). and incentive intensity (model (2), (4), and (6)). No significant direct effects of gender composition are found, whereas compared to only male teams, constant negative coefficients of pure female teams are found, even though not significant. Tournament incentives show constantly positive significant effects on qualitative creative performance, independent of the sample split. Notably, the observations for no team member being part of STEM are rather small with only 23 teams, thus limiting the validity of results.

Table 39: OLS regression on qualitative performance for incentive type and intensity split by STEM occupation of team members

	None in STEM		One in Stem		Both in STEM	
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Type	Intensity	Type	Intensity	Type	Intensity
Incentive (Base: Fix)						
Piece-rate	-5.086		7.798		24.01**	
	(17.72)		(10.54)		(9.889)	
Tournament	65.65***		51.48***		34.36***	
	(22.12)		(12.42)		(10.14)	
Low intensity		13.80		17.44		20.36**
		(21.52)		(11.45)		(9.803)
High intensity		18.54		40.56***		38.29***
		(26.58)		(12.68)		(10.34)
Team-Ability	1.568***	1.706***	1.890***	1.789***	1.659***	1.636***
	(0.344)	(0.430)	(0.229)	(0.255)	(0.201)	(0.200)
Gender composition (Base: Only Males)						
Mixed	-36.78	-14.02	-8.362	-14.77	8.619	11.33
	(22.08)	(23.47)	(12.00)	(13.15)	(9.696)	(9.682)
Only Females	-19.71	-17.97	-5.038	-10.37	-3.497	-0.718
	(24.50)	(31.03)	(11.97)	(12.97)	(9.459)	(9.195)
Constant	118.4***	97.21**	65.49***	75.19***	81.06***	81.02***
	(26.34)	(35.79)	(17.42)	(19.44)	(12.98)	(12.88)
Observations	23	23	101	101	154	154
Adj. R-squared	0.548	0.325	0.530	0.470	0.401	0.412

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

Low- vs. high-ability teams

Lastly, a gender split in low- and high-ability teams might be interesting, as indication exist that pure female teams outperform the other teams in the ability-checking task, whereas regarding the whole sample afterwards during the incentive phase, these effects diminish. Results show that incentive effects maintain especially for the low-ability teams, no significant interaction effects between incentives and gender composition are found in either of both groups (see Table AIV. 11 in Appendix IV). Thus, indicating that especially low-ability teams can benefit from PfP incentive schemes, independent of their gender composition.

7 Discussion and conclusion

In the presented study the interaction between gender composition of teams and PfP incentive schemes, especially tournaments, on creative team performance is analyzed, contributing to the research gap on the role of gender and competitive incentives in context of creative work. The present study complements the findings of Essay 3 that PfP incentives, except a low-intensity piece-rate, have beneficial effects on creative team performance in comparison to fixed pay. These effects increase with incentive intensity and are more pronounced under relative incentive schemes. The consideration of the role of gender composition, especially in context of competitive environments, is crucial to enable a potential increase in creative performance for the whole workforce of an organization and to detect potential detrimental effects for women participating in the creative field. As the present analysis focuses on divergent thinking, the results are of special importance for early stages of the innovation process.

Using a controlled laboratory experiment, not only an arbitrary piece-rate scheme is compared to an arbitrary tournament incentive, but different incentive intensities, i.e., the fraction of payment which is performance dependent, are implemented, providing a more pronounced and informative analyses of those effects. Additionally, the study explicitly focuses on gender, rather than treating it as side effects. Therefore, recruiting participants and designing adequate gender composed teams provides a suitable database. Lastly, the experimental setting is characterized by using a rather unknown divergent thinking task, i.e., Droodles, thus enlarging the task repertoire of creative tasks for researchers conducting creativity research in the lab.

The results provide at first glance an indication of a potential effect of different gender compositions, as pure female teams show a significant positive impact on creative team's performance during the first unincentivized team task in comparison to pure male teams, which diminishes

when moving to the incentive phase. As women showed significantly lower quantitative performance in the individual familiarization task as well as significantly lower levels of creative inclination and cognitive flexibility, potentially indicating lower creative potential in the present sample, the direct positive effect of pure female teams in comparison to pure male teams in the unincentivized team-task is even more remarkable. Nonetheless, further analyses testing for moderation of gender composition and different incentive types or intensities do not detect significant interaction effects. Interestingly, female participants reported higher levels of team cohesion, team identification, intrinsic motivation and pressure than their male counterparts. Even though those variables are found to have a significant influence on creative performance, no significant interaction effects show either. Thus, no harsh evidence for the role of gender composition of dyadic teams on creative performance in a divergent thinking task under PfP incentives is found. Similar results show for using alternative measures of creative performance, gender, separate analyses for participants being involved in the STEM field, and low- and high-ability teams.

Focusing not on average qualitative creative performance as main indicator for creativity but on ideas that received the two highest ratings by the jury, which might be of special interest for managers and employers who aim for a few very creative ideas than average performance, no interaction effects between gender composition and incentive type or intensity are found either.

Further robustness checks examining variability in creative performance provide some indication that pure male teams tend to show higher standard deviations in qualitative performance than other gender compositions, at least under certain conditions. Specifically, significantly higher variability is found in comparison to mixed teams under fixed payment and in comparison to pure female teams under tournament incentives, with similar patterns emerging for quantitative performance. Analyses of performance ratios provide similar results, indicating that pure male teams conduct significantly more errors per valid answer under high-intensity piece-rate and tournament conditions compared to mixed and pure female teams respectively. While these differences do not hold consistently across all conditions, they suggest that under competitive incentive schemes in particular, the results of pure male teams may be less stable.

The diminishing significant effect of only female teams in the incentive phase in comparison to pure male teams can be due to a catching-up of teams with at least one male or a lower performance of pure female teams over the course of the experiment. Thus, relative performance differences between the unincentivized team task and the incentive phase are checked. Results hint towards the relevance of the first mechanism. Potentially, pure female teams seem to have

a lower hurdle already reaching a high-performance level at the unincentivized beginning of cooperation whereas teams with at least one male only seem to reach similar creative performance when incentives are into place. This is in line with the finding of women performing better in unincentivized tasks than men (Segal, 2012), indicating a contingent conscientiousness of men in this specific setting (Minbashian et al., 2010). Similar results are also found for teams performing an unincentivized divergent thinking task in a psychological experiment by Peter et al (2021), where all-female teams showed higher fluency, i.e., quantity, than male counterparts.

7.1 Limitations

The following section addresses limitations of the presented study, with particular focus on those pertaining to its gender-sensitive components. For a discussion of limitations relating to the incentive effects, the reader is referred to Chapter 8.1 of Essay 3.

A first limitation concerns the generalizability of the null result regarding gender composition as a moderator of the relationship between PfP incentives and creative team performance. The present study employs dyadic teams, and it is conceivable that gender composition effects are less salient in pairs than in larger groups. As team size increases, the probability of the emergence of gender-based subgroups within a team rises correspondingly, which in turn heightens both the risk of faultline formation (Peter et al., 2021) and the social pressure on individual members to conform to gender-stereotypical behavior in line with the predictions of RCT (Hora et al., 2021). Those faultlines have been shown to impede the exchange between team members, which is essential to creative team performance (C. X. Chen et al., 2012; Feinstein, 2011; Grund et al., 2026). Accordingly, the null effect observed in the present study should not be generalized to larger team configurations without caution.

Second, the design including five treatments with three gender compositions yields relatively small cell sizes, particularly in the underlying sample for exclusively female teams. This is attributable in part to the comparatively lower proportion of female students enrolled at the institution at which the study has been conducted, and in part to the time and financial constraints inherent to laboratory experimentation. As a consequence, the power available for certain subgroup analyses is limited, even though sufficient power for the main analyses is met.

Third, the composition of the participant pool introduces a potential boundary condition on the external validity of the findings. Given the pronounced STEM orientation of the host institution, female participants in this study may represent a self-selected group whose characteristics differ systematically from those of women who do not select into STEM field or environments, par-

ticularly with respect to competitiveness. It should be noted, however, that supplementary analyses comparing STEM and non-STEM participants do not reveal statistically significant differences.

A fourth limitation concerns the operationalization of gender in the present study. Although participants are given the option to identify as non-binary in the questionnaire, the realized sample is binary in composition, comprising exclusively male and female participants. As a consequence, the findings neither represent nor speak to the experiences of non-binary individuals and cannot capture the full spectrum of gender identities. To partially address this limitation, a continuous, self-perceived measure of gender expression ranging from very masculine to very feminine is employed as an alternative operationalization. The robustness checks so reveal some divergence between participants' binary gender classification and their self-reported position on the scale, yet no significant interaction effects are identified, lending tentative support to the robustness of the main findings. It should be acknowledged, however, that the continuous gender scale is itself subject to interpretive limitations. Assessments are likely made relative to gender-specific rather than universal benchmarks, such that a self-identified female participant perceiving herself as relatively masculine compared to other women may occupy a scale position indistinguishable from that of a male participant with an equivalent self-assessment, despite the qualitatively different referential frame underlying each response.

Finally, the absence of communication data constitutes another limitation. Such data from the outset of the experiment and during task completion would have afforded valuable insight into the social mechanisms and enacted stereotypes underlying observed performance differences across gender compositions. In particular, the finding that female participants report significantly higher levels of team cohesion and identification raises the possibility that qualitatively distinct interpersonal dynamics operate across team types, the nature and consequences of which cannot be fully assessed without such data. Similarly, the presence and intensity of intra-team conflict cannot be evaluated, which is an important mediator of creative exchange.

7.2 Implications for further research

The present study opens several avenues for future research that would both deepen the understanding of gender composition effects in creative teams and address the methodological limitations identified above.

A first and arguably most pressing direction concerns the investigation of the social mechanisms underlying the observed performance patterns. The present study adds to existing evidence suggesting that women outperform men in unincentivized tasks and extends this finding

to the team level. Yet, the mechanisms driving this difference remain opaque. Future studies would therefore benefit from incorporating more fine-grained measures of social interaction prior to and during the creative task, such as structured communication protocols, process tracing, or video-based behavioral coding, in order to identify through which specific interpersonal processes purely female teams are able to mobilize their creative potential more immediately, and what it is about the presence of a male team member that delays this process. Doing so would not only deepen the theoretical understanding of gender composition effects in creative teams but also contribute more broadly to the literature on the conditions under which the individual performance of women translates, or fails to translate, into superior team-level output.

Closely related is the question of communication and its role in mediating gender composition effects on creative output. The observation that females show lower quantitative performance in the individual task and potentially lower individual creative potential, yet nonetheless outperformed male teams in the unincentivized team condition, warrants further investigation. This pattern may reflect either particularly effective communication and coordination processes within female teams, or a comparatively strong suppressive effect of missing (competitive) incentives on male performance, or both. Disentangling these competing explanations represents an important task for future research, ideally through designs that directly observe and quantify communication behavior rather than relying on post-hoc self-report.

A second avenue concerns the distinction between preferences for and performance under competitive incentive schemes. The present study deliberately focuses on the latter, motivated by the nature of divergent thinking tasks in the workplace. However, the findings hint at a potential selection mechanism: rather than reflecting an inherent performance disadvantage under competitive conditions, the underrepresentation of women in creative, competitive professional environments may be driven by differential preferences for such settings in the first place. Future research could therefore examine how preferences for competition evolve and how the presence of a same-gender team member moderates these preferences. In particular, it remains an open question whether the presence of another woman in a team, by shielding individual members from the experience of competing against men alone (e.g., Dargnies, 2012; Ivanova-Stenzel & Kübler, 2011), attenuates the well-documented female aversion to compete against male opponents (Buser et al., 2023; Geraldes, 2020), and whether mixed-gender teams produce analogous buffering effects.

Third, the generalizability of the present findings to larger teams deserves empirical attention. The current design employs dyadic teams, in which gender-based subgroup formation is by

definition impossible. Future studies should vary both team size and the proportional representation of each gender within teams, as the salience of gender-based faultlines is theoretically expected to increase with greater numerical imbalance between gender groups. Examining how the magnitude of faultline effects scales with the degree of gender imbalance would provide important boundary conditions for the present results and connect to broader debates on optimal team composition in organizational settings.

Finally, as noted in the limitations, the operationalization of gender through binary or self-perceived continuous measures each carries its own constraints. Future research would benefit from adopting more theoretically grounded and multi-dimensional measures of gender identity, distinguishing for instance between biological gender, gender identity, and gender expression, for example by using the gender identity scale of Ho & Mussap (2019).

7.3 Practical implications

The null result regarding the moderating role of gender composition in the relationship between PfP incentives and creative team performance carries several implications for practitioners. Most fundamentally, the finding that PfP incentive schemes enhance creative team performance independently of team gender composition suggests that organizations do not have to be concerned that the introduction of such schemes for divergent thinking tasks will systematically disadvantage female employees or erode the creative potential that women bring to the workforce. Practitioners can therefore consider implementing PfP systems when deploying short-term teams for divergent thinking tasks, such as during the early ideation stages of an innovation process, without the risk of generating gender-based performance disparities.

A noteworthy complementary finding, however, is that purely female teams outperform both purely male and mixed-gender teams when the divergent thinking task is unincentivized. This suggests that in organizational contexts where competitive incentives are absent or inappropriate, purely female teams may represent a particularly effective compositional choice, at least for discrete, short-term tasks.

With respect to the choice of specific incentive scheme, winner-takes-all tournament incentives appear to induce the highest average levels of creative performance across team compositions. Practitioners prioritizing peak output may therefore find tournament-style incentives attractive. However, an analysis of performance variability reveals that purely male teams exhibit significantly higher variance in creative output than purely female teams under competitive conditions, alongside elevated error rates. For managers operating in contexts where consistency and

reliability of creative output are paramount, rather than the occasional exceptional result, purely male teams under tournament incentives may therefore not represent the optimal choice.

Finally, the absence of a gender composition effect on performance under competitive incentives has implications beyond team design. It challenges the notion that the well-documented underrepresentation of women in creative, competitive professional environments reflects a performance deficit under such conditions. If women perform comparably to men under PfP schemes in creative tasks, the more likely drivers of their underrepresentation are differential preferences for competitive environments rather than an inability to thrive within them or discriminatory reasons, as women have to act against stereotypes and have to work harder than men if engaging in the creative field (Lubczyk & Moser, 2025; Proudfoot et al., 2015). HR professionals should therefore direct attention toward the design of selection and recruitment processes, considering for instance whether the framing of creative roles as inherently competitive may deter qualified female candidates before performance ever becomes relevant.

7.4 Contribution

The present study investigates the moderating role of team gender composition in the relationship between absolute and relative PfP incentive schemes of varying intensity and creative team performance and offers several contributions to academic literature.

At the broadest level, the study extends the well-established literature on gender differences under PfP incentives (e.g., Niederle & Vesterlund, 2011) in two important respects. First, it moves beyond the routine task settings that dominate this literature into the domain of creative, non-routine performance, where the underlying mechanisms of incentive response may differ substantially. Second, and relatedly, it shifts the unit of analysis from the individual to the team, a context in which existing evidence on the effects of gender composition remains mixed and inconclusive with respect to both the presence and direction of compositional effects (e.g., Apesteguia et al., 2012; Hardt et al., 2025; Ivanova-Stenzel & Kübler, 2011). By employing a clean laboratory design with distinct but comparable absolute and relative incentive schemes, the study contributes to a more systematic understanding of this underexplored intersection.

A further contribution lies in the study's focus on performance outcomes under assigned incentive schemes, rather than on preferences for competitive environments. While the literature on gender differences in competitive preferences is substantial, comparatively little is known about how men and women actually perform once incentive structures are in place and self-selection is removed. This distinction is particularly consequential in the context of divergent thinking tasks, which are short-term and occur rather infrequently as non-major part of working life,

such that employees in organizational settings are unable to opt out of competitive incentive conditions.

The null result regarding the interaction between team gender composition and incentive scheme constitutes a substantive finding in its own right. It indicates that PfP incentives can be introduced to enhance creative team performance without generating systematic disadvantages for purely female teams, including winner-takes-all tournament conditions. This finding contributes to the literature on women in creative, competitive professional contexts, and provides empirical grounding for the argument that the underrepresentation of women in such environments is more plausibly attributable to preference-based selection or discriminatory reasons than to performance deficits under competitive incentives.

Furthermore, the present study offers insights relevant to the theoretical predictions of SRT and RCT in the context of incentivized creative teamwork. While the study does not constitute a direct test of these theories, as both frameworks operate at the individual level whereas performance is measured exclusively at the team level here, the pattern of results nonetheless speaks to their applicability in this context. In particular, the result with respect to interaction effects between team gender composition and incentive scheme tentatively suggests that the role incongruity penalties RCT predicts for women engaging in competitive, male-typed environments may not straightforwardly translate into observable team-level performance differences, at least in the context of short-term dyadic creative tasks.

Finally, by incorporating team gender composition as an additional dimension in the analysis of external monetary incentives in creative team settings, the present study advances a more complete understanding of a relationship that has attracted considerable debate. The observation that purely female teams outperform teams containing at least one male in the absence of incentives adds important nuance to existing accounts of how gender and incentive design jointly shape creative output at the team level.

APPENDIX IV

Acknowledgements

This project was partially funded by the Federal Ministry of Education and Research (BMBF) and the Ministry of Culture and Science of the German State of North Rhine-Westphalia (MKW) under the Excellence Strategy of the Federal Government and the Länder

Tables

AIV. 1: Socio-demographic statistics by gender

Gender	N	Age	Non-EU	# of non-Bachelor	# in STEM	# with Econ background
Male	301	23.02	43	227	240	178
Female	255	22.31	32	172	167	150
Whole Sample	556	22.70	75	399	407	328

AIV. 2: Means of different experiment-related controls by gender measured on the individual level

Gender	N	Intrinsic Motivation **	Cognitive Flexibility *	Effort	Difficulty	Pressure **	Leader **	High Effort Group	Cohesion ***	Identification **	Group Advantages
Males	301	3.856	3.896	4.269	2.960	2.475	2.498	2.914	4.025	3.424	3.132
		(0.641)	(0.457)	(0.982)	(0.976)	(1.232)	(1.256)	(1.235)	(0.494)	(0.738)	(0.568)
Females	255	3.973	3.820	4.253	2.933	2.708	2.265	2.775	4.140	3.550	3.162
		(0.605)	(0.493)	(1.076)	(1.035)	(1.182)	(1.214)	(1.247)	(0.485)	(0.667)	(0.573)

Note: The stars (*** p<0.01, ** p<0.05, * p<0.1) denote whether the genders differ between each other in the corresponding dimension, tested with Mann-Whitney U-Tests.

AIV. 3: OLS regression on creative performance during the team-check task and incentive phase of the *Fix* treatment

VARIABLES	(1) Quantity Team-Check	(2) Quantity Incentive	(3) Quality Team-Check	(4) Quality Incentive
Team-Ability		2.594***		2.342***
		(0.308)		(0.242)
Gender composition (Base: Only Males)				
Mixed	-1.175	2.202	-8.551	12.27
	(1.184)	(2.196)	(6.995)	(12.21)
Only Females	0.0238	0.629	3.524	-12.59
	(1.759)	(2.291)	(9.731)	(11.50)
Constant	13.33***	2.365	64.76***	36.06**
	(0.941)	(3.431)	(5.646)	(15.10)
Observations	54	54	54	54
Adj. R-squared	0.018	0.732	0.041	0.725

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

AIV. 4: OLS regression on quantitative and qualitative creative performance with interaction effects between treatments and gender composition

VARIABLES	(1) Quantity	(2) Quality
Treatments (Base: Fix)		
Plow	4.184*	15.70
	(2.482)	(14.52)
Tlow	3.625	21.88
	(2.650)	(16.60)
Pmax	5.321**	35.15**
	(2.363)	(14.00)
Tmax	10.20***	51.08***
	(3.461)	(18.45)
Team-Ability	2.355***	1.731***
	(0.153)	(0.141)
Gender composition (Base: Only Males)		
Mixed	1.921	7.038
	(2.312)	(13.13)
Only Females	0.634	-10.43
	(2.266)	(12.55)
Treatment x Gender composition		
Plow x Mixed	-3.742	-12.52
	(3.385)	(19.89)
Plow x Only Females	-3.466	-6.871
	(3.170)	(18.26)
Tlow x Mixed	-0.227	1.188
	(3.443)	(20.88)
Tlow x Only Females	2.200	29.96
	(3.224)	(19.90)
Pmax x Mixed	-5.091	-33.01*
	(3.132)	(18.27)
Pmax x Only Females	-0.838	-4.106
	(2.975)	(18.75)
Tmax x Mixed	-0.664	4.300
	(4.497)	(24.03)
Tmax x Only Females	-1.474	1.521
	(4.219)	(21.87)
Constant	5.553**	75.68***
	(2.377)	(12.51)
Observations	278	278
Adj. R-squared	0.596	0.494

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

AIV. 5: OLS regression on qualitative creative performance with interaction effects between incentive type and several control variables

VARIABLES	(1) Team-Ability	(2) Team-Ability	(3) Cohesion	(4) Cohesion	(5) Identification	(6) Identification	(7) Pressure	(8) Pressure
Incentive type (Base: Fix)								
Piece-rate	16.12**	16.48**	15.47**	15.27**	16.17**	15.97**	16.23**	16.65**
	(6.620)	(6.551)	(6.717)	(6.767)	(6.585)	(6.580)	(6.642)	(6.733)
Tournament	42.11***	42.30***	41.11***	40.85***	42.05***	42.03***	41.80***	41.63***
	(7.289)	(7.296)	(7.370)	(7.401)	(7.289)	(7.362)	(7.295)	(7.457)
Gender composition (Base: Only Males)								
Mixed	-1.389	13.77	-1.655	37.66	-1.427	-36.91	-1.671	-26.52
	(7.169)	(22.70)	(7.070)	(73.04)	(7.179)	(47.46)	(7.146)	(20.18)
Only Females	-6.217	-2.992	-10.32	52.32	-6.357	-25.90	-7.416	-3.674
	(7.029)	(19.42)	(7.304)	(73.43)	(7.264)	(48.04)	(7.227)	(20.14)
Team-Ability	1.735***	1.831***	1.709***	1.702***	1.737***	1.738***	1.713***	1.721***
	(0.142)	(0.270)	(0.138)	(0.138)	(0.144)	(0.144)	(0.145)	(0.144)
Mixed x Team-Ability		-0.259						
		(0.396)						
Only Females x Team-Ability		-0.0578						
		(0.310)						
Cohesion			21.78***	29.45**				
			(7.339)	(14.43)				
Mixed x Cohesion				-9.781				
				(18.41)				
Only Females x Cohesion				-15.20				
				(18.00)				
Group identification					0.779	-3.814		
					(5.843)	(8.723)		
Mixed x Group identification						10.29		
						(14.09)		
Only Females x Group identification						5.669		
						(13.67)		
Pressure							4.220	1.160
							(3.360)	(6.001)
Mixed x Pressure								9.909
								(8.266)
Only Females x Pressure								-1.010
								(7.386)
Constant	77.25***	71.24***	-8.053	-38.27	74.49***	90.20***	68.23***	75.23***
	(9.901)	(16.24)	(29.71)	(57.06)	(23.07)	(32.42)	(11.22)	(15.72)
Observations	278	278	278	278	278	278	278	278
Adj. R-squared	0.454	0.452	0.469	0.446	0.452	0.449	0.455	0.456

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

AIV. 6: OLS regression on qualitative creative performance with interaction effects between incentive intensity and several control variables

VARIABLES	(1) Team-Ability	(2) Team-Ability	(3) Cohesion	(4) Cohesion	(5) Identification	(6) Identification	(7) Pressure	(8) Pressure
Incentive intensity (Base: Fix)								
Low intensity	20.00***	20.30***	19.75***	19.77***	20.11***	19.92***	20.04***	20.28***
	(6.847)	(6.834)	(6.932)	(6.984)	(6.842)	(6.834)	(6.857)	(6.983)
High intensity	38.23***	38.60***	36.89***	36.54***	38.11***	37.94***	37.97***	37.85***
	(7.270)	(7.227)	(7.346)	(7.333)	(7.272)	(7.329)	(7.299)	(7.428)
Gender composition (Base: Only Males)								
Mixed	-0.711	17.37	-1.005	26.67	-0.849	-29.14	-1.030	-27.38
	(7.225)	(23.20)	(7.137)	(70.80)	(7.246)	(46.21)	(7.203)	(20.23)
Only Females	-6.331	-0.908	-10.31	34.72	-6.807	-28.42	-7.633	-0.409
	(7.137)	(19.72)	(7.420)	(73.89)	(7.381)	(47.57)	(7.358)	(19.49)
Team-Ability	1.700***	1.821***	1.676***	1.671***	1.706***	1.706***	1.676***	1.687***
	(0.146)	(0.273)	(0.143)	(0.143)	(0.148)	(0.148)	(0.150)	(0.148)
Mixed x Team-Ability		-0.308						
		(0.409)						
Only Females x Team-Ability		-0.0930						
		(0.313)						
Cohesion			21.13***	26.62*				
			(7.168)	(14.35)				
Mixed x Cohesion				-6.890				
				(17.89)				
Only Females x Cohesion				-10.92				
				(18.03)				
Group identification					2.653	-1.416		
					(5.829)	(8.933)		
Mixed x Group identification						8.216		
						(13.75)		
Only Females x Group identification						6.225		
						(13.63)		
Pressure							4.588	1.627
							(3.395)	(5.973)
Mixed x Pressure								10.50
								(8.318)
Only Females x Pressure								-2.270
								(7.280)
Constant	79.29***	71.64***	-3.536	-25.17	69.87***	83.90**	69.45***	76.08***
	(10.19)	(16.44)	(29.29)	(57.11)	(22.65)	(32.61)	(11.26)	(15.60)
Observations	278	278	278	278	278	278	278	278
Adj. R-squared	0.437	0.435	0.450	0.447	0.435	0.432	0.439	0.441

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

AIV. 7: OLS regression on average differences of quantitative and qualitative performance moving from the team-check to the incentive phase with interaction effects between incentive intensity and gender composition

	Base: Only Females	
	(1)	(2)
VARIABLES	Quantity	Quality
Incentive intensity (Base: Fix)		
Low intensity	1.035	11.58***
	(0.676)	(3.872)
High intensity	1.996***	13.02***
	(0.735)	(4.449)
Gender composition (Base: Only Females)		
Only Males	-0.206	4.968
	(0.805)	(4.270)
Mixed	0.687	10.93***
	(0.734)	(3.869)
Incentive intensity x Gender composition		
Low intensity x Only Males	0.456	-2.701
	(0.991)	(5.968)
Low intensity x Mixed	-0.404	-7.140
	(0.931)	(5.548)
High intensity x Only Males	0.749	3.260
	(1.100)	(6.516)
High intensity x Mixed	-0.410	-4.305
	(1.063)	(6.539)
Constant	-0.810	-7.143***
	(0.561)	(2.741)
Observations	278	278
Adj. R-squared	0.049	0.051

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

AIV. 8: OLS regression on average differences of quantitative and qualitative performance moving from the team-check to the incentive phase with interaction effects between treatments and gender composition

	Base: Only Females	
	(1)	(2)
VARIABLES	Quantity	Quality
Treatments (Base: Fix)		
Plow	0.187	3.854
	(0.749)	(4.071)
Tlow	1.830**	18.83***
	(0.755)	(4.628)
Pmax	1.272*	9.198
	(0.761)	(5.730)
Tmax	2.810***	17.31***
	(0.978)	(5.501)
Gender composition (Base: Only Females)		
Only Males	-0.206	4.968
	(0.814)	(4.318)
Mixed	0.687	10.93***
	(0.742)	(3.913)
Treatment x Gender composition		
Plow x Only Males	1.463	4.543
	(1.103)	(6.643)
Plow x Mixed	-0.0812	-2.893
	(1.081)	(6.442)
Tlow x Only Males	-0.516	-9.410
	(1.179)	(7.550)
Tlow x Mixed	-0.674	-10.90
	(1.074)	(6.673)
Pmax x Only Males	0.627	3.876
	(1.082)	(7.757)
Pmax x Mixed	-1.326	-11.18
	(1.101)	(7.757)
Tmax x Only Males	0.670	1.749
	(1.501)	(8.415)
Tmax x Mixed	0.243	0.971
	(1.397)	(8.261)
Constant	-0.810	-7.143**
	(0.567)	(2.772)
Observations	278	278
Adj. R-squared	0.097	0.097

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

AIV. 9: OLS regression on top-evaluated ideas during the team-check task and incentive phase with interaction effects between incentive intensity and gender composition

VARIABLES	(1) Team-check	(2) Incentive phase	(3) Incentive phase
Incentive (Base: Fix)			
Low Intensity	-0.000772	0.00986***	0.00824
	(0.00491)	(0.00342)	(0.00519)
High Intensity	-0.00296	0.0120***	0.0118**
	(0.00486)	(0.00346)	(0.00508)
Team-Ability Top Ideas		0.126**	0.117*
		(0.0611)	(0.0597)
Gender composition (Base: Only Males)			
Mixed	-0.00597*	0.000698	0.00237
	(0.00355)	(0.00311)	(0.00526)
Only Females	-0.000874	0.00957**	0.00456
	(0.00425)	(0.00379)	(0.00724)
Incentive x Gender composition			
Low intensity x Mixed			-0.00446
			(0.00710)
Low intensity x Only Females			0.0119
			(0.00975)
High intensity x Mixed			0.000553
			(0.00762)
High intensity x Only Females			0.000849
			(0.00919)
Constant	0.0186***	0.0121***	0.0130***
	(0.00458)	(0.00316)	(0.00408)
Observations	278	278	278
Adj. R-squared	0.002	0.065	0.068

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

AIV. 10: OLS regression on qualitative performance during the incentive phase with a self-perceived gender measure

VARIABLES	(1) Type	(2) Intensity	(3) Treatment
Incentive (Base: Fix)			
Piece-rate	36.32**		
	(16.48)		
Tournaments	45.30**		
	(18.05)		
Low intensity		21.48	
		(17.25)	
High intensity		58.24***	
		(17.02)	
Plow			28.09
			(19.28)
Tlow			12.81
			(20.18)
Pmax			44.62**
			(18.20)
Tmax			68.46***
			(20.67)
Team-Ability	1.749***	1.711***	1.730***
	(0.143)	(0.146)	(0.142)
Gender	0.468	0.461	0.464
	(1.164)	(1.174)	(1.177)
Incentive x Gender			
Piece-rate x Gender	-2.107		
	(1.405)		
Tournament x Gender	-0.358		
	(1.506)		
Low intensity x Gender		-0.183	
		(1.489)	
High intensity x Gender		-2.124	
		(1.423)	
Plow x Gender			-1.956
			(1.671)
Tlow x Gender			1.794
			(1.692)
Pmax x Gender			-2.251
			(1.555)
Tmax x Gender			-1.707
			(1.718)
Constant	69.96***	72.41***	71.16***
	(15.08)	(15.24)	(15.11)
Observations	278	278	278
Adj. R-squared	0.458	0.441	0.474

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

AIV. 11: OLS regression on qualitative performance during the incentive phase split by previous ability of team members with interaction effects between gender composition and incentive schemes

	Low-ability teams		High-ability teams	
	(1)	(2)	(3)	(4)
VARIABLES	Type	Type	Intensity	Intensity
Incentive (Base: Fix)				
Piece-rate	32.29		10.44	
	(19.79)		(24.09)	
Tournament	51.54***		25.88	
	(19.10)		(29.37)	
Low intensity		31.32		6.309
		(18.94)		(25.98)
High intensity		55.54***		26.25
		(19.84)		(26.75)
Gender composition (Base: Only Males)				
Mixed	24.06	24.06	-13.46	-13.46
	(18.72)	(18.72)	(30.24)	(30.24)
Only Females	-7.778	-7.778	-1.750	-1.750
	(17.35)	(17.35)	(31.00)	(31.00)
Incentive x Gender composition				
Piece-rate x Mixed	-20.88		-14.48	
	(24.30)		(35.81)	
Piece-rate x Only Females	4.440		7.677	
	(24.49)		(35.66)	
Tournament x Mixed	-21.42		24.30	
	(23.90)		(39.52)	
Tournament x Only Females	10.73		1.846	
	(23.18)		(37.16)	
Low intensity x Mixed		-26.32		14.28
		(23.24)		(38.29)
Low intensity x Only Females		7.593		5.060
		(24.77)		(35.14)
High intensity x Mixed		-15.42		-2.336
		(24.52)		(37.71)
High intensity x Only Females		-2.444		8.850
		(23.97)		(36.22)
Constant	137.1***	137.1***	225.8***	225.8***
	(15.12)	(15.12)	(22.36)	(22.36)
Observations	141	141	137	137
Adj. R-squared	0.131	0.160	0.075	0.043

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

PART VI: CONCLUDING REMARKS

1 Summary of findings and contribution

The dissertation at hand contributes to the understanding of workplace factors, namely, place to work at, PfP incentivization, and gender composition of teams, and their effect on the creative performance of newly formed dyadic teams, thereby focusing on divergent thinking tasks as an early stage of the innovative process. Thus, the presented work contributes to the understanding of increasing competitive advantage through innovations in organizations, whose relevance is also stated again by the Nobel Prize in economics in 2025 which has been assigned for the work on sustained growth through creative destruction. Even though by focusing on divergent thinking, the level of creative ideas is not destructive per se, rather it sets the early stage for ideas of which such creative actions might occur (Guilford, 1950; Reiter-Palmon et al., 2019; Weisberg, 2006). By increasing performance of teams in this phase of the creative process either by considering the most productive workplace, i.e., remote or in presence, or the most suitable incentive schemes, i.e., fixed payment, piece-rates, or tournaments, and their respective intensity, as well as gender composition of teams, the present dissertation contributes to an increased understanding of creative teams, enabling practitioners to re-design workplace characteristics, thereby enabling overall corporative performance benefits.

Several key findings can be pointed out, structured along the four essays of the thesis. First of all, before focusing on creative teams, it needs to be investigated whether such teams outperform individuals, justifying their existence form a performance-oriented standpoint. Comparing individuals and creative teams repeatedly working remotely on creative tasks shows that teams significantly outperform individuals in a remote context with repeated interaction. Thus, lending evidence to the implementation of teams working on divergent thinking tasks, even if the exchange which fosters creative performance happens via video conferences.

This key finding leads to follow-up research on whether differences in creative performance occur, depending on the workplace of newly formed teams. Thereby, not only incorporating fully remote or fully presence work, but also hybrid working models and considering self-selection options. Results show that teams which work at least once together in presence outperform fully remote teams. Additionally, indicating that the knowledge of meeting each other at least once is sufficient to achieve greater creative performance than working fully remote. No effects of self-selecting into fully presence or remote workplaces on creative team performance show.

Next to the place to work at, this thesis focuses on another highly relevant component of work design, i.e., incentivizing performance. Thereby, not only comparing absolute and relative performance schemes but also focusing on their intensity, potentially affecting creative team performance through changed risk and informational values of incentives. Results show that teams under PfP incentives, except a low intensity piece-rate, outperform teams under a fixed payment scheme, whereas performance increases with intensity and tournaments outperform comparable piece-rates. Nonetheless, winner-takes-all tournaments tend to create the highest variability in a team's creative performance.

Lastly, in this context, the gender composition of creative teams is considered, as especially competitive incentives might inhibit females from performing in the creative field. No detrimental effects of tournaments for only female or mixed teams are found. Results indicate a higher performance of pure female teams in comparison to the others if the task is unincentivized, whereas teams with at least one male catch-up to the high performance of women if incentives are introduced, independent of incentive type or intensity. Thus, setting monetary incentives for creative team performance induces beneficial effects for the overall workforce of an organization, independent of gender.

By this extensive analysis of creative teams, the thesis has several key contributions. Generally speaking, it advances the understanding of work-related factors affecting creative performance of teams by adding a comprehensive analysis of several contextual factors and systematically comparing them with each other under clean laboratory conditions.

More specifically, the first two parts contribute to research on remote and hybrid working models for teams, including repeated interaction rather than one-shot observation. Thus, they contribute to literature on remote work tendencies and effectiveness over a longer time frame. This is especially valuable in context of the debate around lowering options of remote work for employees (e.g., Lott & Peters, 2026).

Moreover, the second half of this thesis adds to the debate on external monetary incentives potentially crowding-out intrinsic motivation and thus creativity while not only focusing on arbitrary comparisons of one absolute and relative incentive scheme but systematically varying incentive intensity and not only type. Additionally, this is set into context of research on the potential gender gap in creative fields with focus on a structural component of such, i.e., competitive environments induced by tournament incentives. Thereby, adding a comprehensive analysis of the effects of different PfP schemes on creative performance to a literature that mainly focuses on individuals or a gender gap occurring under routine tasks.

Lastly, two different tasks, i.e. UUT and Droodles, to assess creative performance in teams are used throughout this thesis. Both tasks are not commonly used either on team-level or in creative research in general. Thereby, the thesis contributes to the methodological field too, by enlarging the task repertoire for creativity experiments in context of creative teams in the economic laboratory. Additionally, objective and subjective evaluation techniques of creative performance are used for data evaluation, thus enabling guidance for conducting such evaluation using one of the two methods.

2 Future research

The focus on divergent thinking tasks as well as laboratory experiments as the methodology in this dissertation enables a deeper understanding of the early phase of the creative process and the mechanisms through which creative performance of teams on such tasks can be fostered. This focus leads to three avenues for future research, i.e., moving from divergent thinking to convergent thinking, investigating the role of AI for human creativity, and moving from the laboratory to a field setting, which are further discussed in the following section. These are by no means the only paths for future research in the field of creative teams in the workplace, but potentially promising ones.

Convergent thinking

First of all, and as already mentioned during the discussion sections of the presented essays, the focus lays on divergent thinking, which is about finding as many creative ideas as possible. A consequential next step would be to consider convergent thinking, i.e., finding the best solution possible. More specifically, this could be done by switching tasks in future experiments, keeping everything else constant, for example by using the *Remote Association Test* (Mednick, 1962). During this test participants receive three words and have to find the fourth word or overarching category, such that the given words are connected. For example, the words broken, clear and eye, are connected by the word glass. Another alternative is the use of so-called insight problems, which are puzzles that require a shift away from established habits and thinking (Huo, 2020). Next to replacing tasks, another option would be to develop a design including both tasks, such that performance of the same teams in both dimensions of creative thinking can directly be compared. Such a variation in task focus would allow a better understanding of the creative process and how different workplace factors can affect a broader spectrum of creative performance.

Creativity and AI

Another fundamental avenue for future research concerns the relationship between creativity and AI, particularly large language models (LLMs). Current AI systems such as ChatGPT generate outputs by statistically predicting the most likely next words or responses based on patterns learned from large amounts of existing data. While these systems can produce highly sophisticated and often creative-seeming outputs, they fundamentally rely on recombination of already existing information rather than genuine abstraction or original ideation in the human sense. Increasing the "temperature" parameter of AI systems, which raises the randomness and diversity of generated outputs, may enhance originality but simultaneously increase the likelihood of implausible or erroneous suggestions if users fail to evaluate AI-generated ideas. Since LLMs are trained exclusively on existing human-generated content, they may be less capable of producing truly original or paradigm-shifting ideas that require abstraction beyond learned patterns. Human creativity, in contrast, may be characterized precisely by the ability to deviate from established structures and generate novel conceptual connections that are not directly inferable from previous examples.

Recent research demonstrates the growing relevance of AI in creative contexts. Charness and Grieco (2026) show that different versions of ChatGPT outperform individual humans in closed creative tasks, which the UUT and Droodle-Task also belong to as both provide clearly defined constraints and evaluation criteria. Similar findings are reported by Girotra et al. (2023). However, rather than implying a complete substitution of human creativity, these studies point towards the potential of synergies between human and AI creativity. This interpretation is further supported by Boussioux et al. (2024), who find that human-AI collaboration can lead to superior business ideas compared to purely human-generated outcomes. In contrast, Bohren et al. (2024) suggest that such synergies may be less effective than AI-only approaches in text-writing tasks, which are among the primary strengths of contemporary LLMs (Charness & Grieco, 2026).

Building on this discussion, future research could investigate alternative team compositions in creative production processes, for example by comparing traditional human teams with hybrid teams consisting of one human individual supported by AI systems, as well as purely AI-generated outputs.

Another promising direction concerns heterogeneity in AI usage behavior. Existing evidence indicates gender differences in the adoption and effective use of AI tools. Men are generally found to use AI systems more frequently and to achieve higher success rates in prompting tasks (Carvajal et al., 2024; Møgelvang et al., 2024). At the same time, women appear to engage more

critically with AI-generated outputs and tend to formulate more detailed prompts (Matobobo, 2026). These differences may reflect distinct usage strategies, objectives, and evaluation processes when solving creative tasks and could therefore influence the effectiveness of AI support in creative environments.

Finally, AI may also become increasingly important for the evaluation of creativity itself. For instance, Organisciak et al. (2023) employ LLMs trained on subjective human ratings to evaluate creative performance in UUTs. This approach could be particularly valuable for future experimental research, as large experimental datasets such as the one generated in Essay 2 may provide an opportunity to train AI-based evaluation systems alongside existing objective creativity measures. Such approaches could substantially reduce evaluation costs while simultaneously enabling more nuanced and scalable assessments of creative performance.

Moving to the field

Moreover, all essays presented use data from controlled laboratory experiments, thereby providing a starting point for rather explorative research (Falk & Heckman, 2009) on creative teams in the workplace, performing divergent thinking tasks. Nonetheless, some might consider this setting as a limitation, due to missing long-term data and the generalizability of results towards managers in real organizations, given the artificial nature of the laboratory environment. Thus, further research should consider adjusting the empirical approach, either by utilizing existing panel datasets such as the Socio-Economic Panel or the Linked Personnel Panel, or by conducting a natural field experiment.

Considering the characteristics of defining and especially evaluating creative work, an occurring problem for both variants of research is the consistent evaluation of what creative performance is, but also in identifying creative tasks. Therefore, a field experiment enabling ex-ante training of evaluators and more control over the tasks under consideration appears more applicable to creativity research. Optimally, such a study would be conducted across several organizations from different industries over an extended period of time, which would not only increase the available sample size but also allow for testing whether findings generalize beyond a specific organization or sector.

Employee suggestion programs offer an ideal context for such an endeavor, as they naturally accommodate the ex-ante training of evaluators while providing a structured setting for idea submission. These programs allow employees, sometimes anonymously, to submit ideas spanning a broad range of topics, from innovation processes and new product development to activ-

ities for reducing a company's environmental impact (Lasrado et al., 2016). Importantly, suggestion programs are typically open to all employees, allowing individuals to contribute ideas beyond their immediate area of responsibility and thus reflecting the nature of creativity in organizational life (Fairbank & Williams, 2001). Furthermore, many programs are not restricted in duration, enabling longitudinal data collection.

This setting is particularly well suited for extending the findings of the four essays of the present dissertation. By expanding the suggestion scheme to include both individual and team submissions, one could directly compare creativity levels between individuals and teams (Essays 1). Additionally, data could be gathered on whether individuals and teams work fully in presence, remote or using hybrid working models (Essay 2), as well as on the gender composition of teams (Essay 4). Since suggestion programs are frequently tied to profit-sharing arrangements for submitted ideas (Fairbank & Williams, 2001), they provide a framework for examining the effect of incentives on creative performance (Essay 3). Beyond the rated creativity of submissions, whether ideas are ultimately implemented by the organization could serve as an additional outcome variable, capturing convergent thinking.

Nonetheless, such a field experiment still has some limitations to consider. First, there is a risk of self-selection as employees who voluntarily participate in a suggestion scheme may systematically differ from non-participants in terms of baseline creativity and motivation, potentially confounding the results (Freyer et al., 2025). Second, training of the evaluators of those creative ideas is also necessary, whereas in such an environment it cannot be ruled out that evaluators of ideas pursue their own interests when evaluating ideas. Third, if anonymity of employees submitting ideas cannot be fully ensured, social desirability bias might alter submissions.

Despite those limitations, moving from a laboratory setting to the field in context of employee suggestion systems is a promising avenue for future research, aiming for a comprehensive understanding of the influence of workplace factors on creativity, thus, shaping the creative performance of teams.

REFERENCES

Abbink, K., Brandts, J., Herrmann, B., & Orzen, H. (2010). Intergroup Conflict and Intra-Group Punishment in an Experimental Contest Game. *American Economic Review, 100*(1), 420-447. https://doi.org/10.1257/aer.100.1.420

Abi Saad, E., & Agogué, M. (2023). Creativity in virtual teams: Systematic review, synthesis and research agenda. *Creativity and Innovation Management, 32*(1), 117-140. https://doi.org/10.1111/caim.12540

Abra, J. (1994). Collaboration in creative work: An initiative for investigation. *Creativity Research Journal, 7*(1), 1-20. https://doi.org/10.1080/10400419409534505

Acar, S., Ogurlu, U., & Zorychta, A. (2023). Exploration of discriminant validity in divergent thinking tasks: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts, 17*(6), 705-724. https://doi.org/10.1037/aca0000469

Ai, W., Chen, Y., Mei, Q., Ye, J., & Zhang, L. (2023). Putting teams into the gig economy: A field experiment at a ride-sharing platform. *Management Science, 69*(9), 5336-5353. https://doi.org/10.1287/mnsc.2022.4624

Alchian, A., & Demsetz, H. (2009). Production, information costs, and economic organization. In *The economic nature of the firm* (3rd ed., pp. 173–196). Cambridge University Press. https://doi.org/10.1017/cbo9780511817410.015

Ales, L., Cho, S.-H., & Körpeoğlu, E. (2017). Optimal Award Scheme in Innovation Tournaments. *Operations Research, 65*(3), 693-702. https://doi.org/10.1287/opre.2016.1575

Ali, S., Moroso, T., & Breazeal, C. (2019). Can Children Learn Creativity from a Social Robot? In *Proceedings of the 2019 on Creativity and Cognition*. ACM, 359-368. https://doi.org/10.1145/3325480.3325499

Alsharo, M., Gregg, D., & Ramirez, R. (2017). Virtual team effectiveness: The role of knowledge sharing and trust. *Information & Management, 54*(4), 479-490. https://doi.org/10.1016/j.im.2016.10.005

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997-1013. https://doi.org/10.1037/0022-3514.43.5.997

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology, 45*(2), 357-376. https://doi.org/10.1037/0022-3514.45.2.357

Amabile, T. M. (1988). A model of creativity and innovation in organizations. *Research in Organizational Behavior, 10*(1), 123-167.

Amabile, T. M., & Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior, 36*, 157-183. https://doi.org/10.1016/j.riob.2016.10.001

Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A., & Leonardi, S. (2012). Online team formation in social networks. *Proceedings of the 21st International Conference on World Wide Web*, 839-848. Lyon, France. https://doi.org/10.1145/2187836.2187950

Anderson, A. J., Kaplan, S. A., & Vega, R. P. (2015). The impact of telework on emotional experience: When, and for whom, does telework improve daily affective well-being? *European Journal of Work and Organizational Psychology, 24*(6), 882-897. https://doi.org/10.1080/1359432X.2014.966086

Anderson, N., Potočnik, K., & Zhou, J. (2014). Innovation and Creativity in Organizations. *Journal of Management, 40*(5), 1297-1333. https://doi.org/10.1177/0149206314527128

Apesteguia, J., Azmat, G., & Iriberri, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science, 58*(1), 78-93. https://doi.org/10.1287/mnsc.1110.1348

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large Stakes and Big Mistakes. *The Review of Economic Studies, 76*(2), 451-469. https://doi.org/10.1111/j.1467-937X.2009.00534.x

Attanasi, G., Curci, Y., Llerena, P., & Urso, G. (2019). *Intrinsic vs. extrinsic motivators on creative collaboration: The effect of sharing rewards* (No. 2019-20). Bureau d'Economie Théorique et Appliquée, UDS, Strasbourg.

Auer, T., Hoppe, J. A., & Thommes, K. (2024). Time perspectives and collaborative performance in creative tasks. *Journal of Organizational Effectiveness: People and Performance, Ahead of Print, 11*(4), 1023-1042. https://doi.org/10.1108/joepp-07-2023-0285

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics, 118*(4), 1279-1333. https://doi.org/10.1162/003355303322552801

Baard, P. P., Deci, E. L., & Ryan, R. M. (2004). Intrinsic need satisfaction: a motivational basis of performance and well-being in two work settings. *Journal of Applied Social Psychology, 34*(10), 2045-2068. https://doi.org/10.1111/j.1559-1816.2004.tb02690.x

Baer, M., Leenders, R. T. A. J., Oldham, G. R., & Vadera, A. K. (2010). Win or Lose the Battle for Creativity: The Power and Perils of Intergroup Competition. *Academy of Management Journal, 53*(4), 827-845. https://doi.org/10.5465/amj.2010.52814611

Baer, M., Vadera, A. K., Leenders, R. T., & Oldham, G. R. (2014). Intergroup competition as a double-edged sword: How sex composition regulates the effects of competition on group creativity. *Organization Science, 25*(3), 892-908. https://doi.org/10.1287/orsc.2013.0878

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science, 23*(10), 1117-1122. https://doi.org/10.1177/0956797612446024

Barczak, G., Lassk, F., & Mulki, J. (2010). Antecedents of team creativity: An examination of team emotional intelligence, team trust and collaborative culture. *Creativity and Innovation Management, 19*(4), 332-345. https://doi.org/10.1111/j.1467-8691.2010.00574.x

Barrero, J. M., Bloom, N., & Davis, S. J. (2021). Why working from home will stick. *National Bureau of Economic Research, Working Paper*. https://doi.org/10.3386/w28731

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic distance and the Alternate Uses Task: Recommendations for reliable automated assessment of originality. C*reativity Research Journal, 34*(3), 245-260. https://doi.org/10.1080/10400419.2022.2025720

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2014). Eine single-item-Skala zur Erfassung von risikobereitschaft: die Kurzskala risikobereitschaft-1 (R-1). https://www.ssoar.info/ssoar/bitstream/handle/document/42670/ssoar-2014-beierlein_et_al-eine_single-item-skala_zur_erfassung_von.pdf?sequence=1

Bellmann, L., & Hübler, O. (2021). Working from home, job satisfaction and work–life balance–robust or heterogeneous links? *International Journal of Manpower, 42*(3), 424-441. https://doi.org/10.1108/IJM-10-2019-0458

Benz, M., & Frey, B. S. (2008). The value of doing what you like: Evidence from the self-employed in 23 countries. *Journal of Economic Behavior & Organization, 68*(3-4), 445-455. https://doi.org/10.1016/j.jebo.2006.10.014

Birney, D. P., Beckmann, J. F., & Seah, Y. Z. (2016). More than the eye of the beholder: The interplay of person, task, and situation factors in evaluative judgements of creativity. *Learning and Individual Differences, 51*, 400-408. https://doi.org/10.1016/j.lindif.2015.07.007

Bishop, J. (2006). Drinking from the fountain of knowledge: Student incentive to study and learn–externalities, information problems and peer pressure. *Handbook of Economics of Education, 2*, 909-944. https://doi.org/10.1016/S1574-0692(06)02015-0

Blair, M. M. (1999). Firm-specific human capital and theories of the firm. In M. M. Blaire & M. J. Roe (Eds.) Employees and Corporate Governance (pp. 58-90). The Brookings Institute. https://doi.org/10.2139/ssrn.167848

Blaufus, K., Jarzembski, J. K., Reineke, J., & Trenn, I. (2025). Public disclosure of tax strategies and firm's actual tax policy. *Journal of Accounting, Auditing & Finance, 40*(3), 863-887. https://doi.org/10.1177/0148558X231200913

Blaufus, K., Piehl, K., & Schröder, M. (2026). Input versus output incentives in idea generation – An experimental analysis. *European Economic Review*, 105329. https://doi.org/10.1016/j.euroecorev.2026.105329

Blay, T., Schöbel, S., Heubeck, T., Reinwald, M., Zeller-Lanzl, J., Froese, F. J., ... & Kolbe, L. (2026). The future of global virtual work: Toward an interdisciplinary and integrative research agenda. *German Journal of Human Resource Management*. https://doi.org/10.1177/23970022261436919

Bloom, N., Liang, J., Roberts, J., & Ying, Z. J. (2015). Does working from home work? Evidence from a Chinese experiment. *The Quarterly Journal of Economics, 130*(1), 165-218. https://doi.org/10.1093/qje/qju032

Bohren, N., Hakimov, R., & Lalive, R. (2024). Creative and strategic capabilities of generative AI: Evidence from large-scale experiments. IZA Discussion Paper, 17302, 1-38.

Boies, K., Fiset, J., & Gill, H. (2015). Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. *The Leadership Quarterly, 26*(6), 1080-1094. https://doi.org/10.1016/j.leaqua.2015.07.007

Boland, J. E., Fonseca, P., Mermelstein, I., & Williamson, M. (2022). Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General, 151*(6), 1272-1282. https://doi.org/10.1037/xge0001150

Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Accounting, Organizations and Society, 27*(4-5), 303-345. https://doi.org/10.1016/S0361-3682(01)00052-6

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2024). The crowdless future? Generative AI and creative problem-solving. *Organization Science*, *35*(5), 1589-1607. https://doi.org/10.1287/orsc.2023.18430

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. https://doi.org/10.1086/702649

Brenan, M. (2020, October 13). *COVID-19 and Remote Work: An Update*. Retrieved from Gallup. https://news.gallup.com/poll/321800/covid-remote-work-update.aspx

Brosig, J., & Weimann, J. (2003). The effect of communication media on cooperation. *German Economic Review, 4*(2), 217-241. https://doi.org/10.1111/1468-0475.00080

Brown, H. G., Poole, M. S., & Rodgers, T. L. (2004). Interpersonal traits, complementarity, and trust in virtual collaboration. *Journal of Management Information Systems, 20*(4), 115-138. https://doi.org/10.1080/07421222.2004.11045785

Brucks, M. S., & Levav, J. (2022). Virtual communication curbs creative idea generation. *Nature, 605*(7908), 108-112. https://doi.org/10.1038/s41586-022-04852-5

Brüggen, A., & Strobel, M. (2007). Real effort versus chosen effort in experiments. *Economics Letters, 96*(2), 232-236. https://ssrn.com/abstract=816007

Bull, C., Schotter, A., & Weigelt, K. (1987). Tournaments and Piece Rates: An Experimental Study. *Journal of Political Economy, 95*(1), 1-33. https://doi.org/10.1086/261439

Burk, C. L., & Wiese, B. S. (2018). Professor or manager? A model of motivational orientations applied to preferred career paths. *Journal of Research in Personality, 75*, 113-132. https://doi.org/10.1016/j.jrp.2018.06.002

Buser, T., van den Assem, M. J., & van Dolder, D. (2023). Gender and willingness to compete for high stakes. *Journal of Economic Behavior & Organization, 206*, 350-370. https://doi.org/10.1016/j.jebo.2022.12.016

Byron, K., & Khazanchi, S. (2012). Rewards and creative performance: a meta-analytic test of theoretically derived hypotheses. *Psychological Bulletin, 138*(4), 809-830. https://doi.org/10.1037/a0027652

Byron, K., Keem, S., Darden, T., Shalley, C. E., & Zhou, J. (2023). Building blocks of idea generation and implementation in teams: A meta-analysis of team design and team creativity and innovation. *Personnel Psychology, 76*(1), 249-278. https://doi.org/10.1111/peps.12501

Cabra, J. F., & Guerrero, C. D. (2022). Regional creativity: Cultural and socio-economic differences. *Journal of Creativity, 32*(2)*,* 100022. https://doi.org/10.1016/j.yjoc.2022.100022

Carpenter, J., & Huet-Vaughn, E. (2019). 19. Real-effort tasks. *Handbook of Research Methods and Applications in Experimental Economics*, 368-381.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal, 17*(1), 37-50. https://doi.org/10.1207/s15326934crj1701_4

Carvajal, D., Franco, C., & Isaksson, S. (2025). Will artificial intelligence get in the way of achieving gender equality?. *NHH Dept. of Economics Discussion Paper, 03*. https://dx.doi.org/10.2139/ssrn.4759218

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: a 40-year meta-analysis. *Psychological Bulletin, 140*(4), 980-1008. https://doi.org/10.1037/a0035661

Chang, H. H., Chuang, S. S., & Chao, S. H. (2011). Determinants of cultural adaptation, communication quality, and trust in virtual teams' performance. *Total Quality Management, 22*(3), 305-329. https://doi.org/10.1080/14783363.2010.532319

Charalampous, M., Grant, C. A., Tramontano, C., & Michailidis, E. (2019). Systematically reviewing remote e-workers' well-being at work: A multidimensional approach. *European Journal of Work and Organizational Psychology, 28*(1), 51-73. https://doi.org/10.1080/1359432X.2018.1541886

Charness, G., & Chen, Y. (2020). Social identity, group behavior, and teams. *Annual Review of Economics, 12*(1), 691-713. https://doi.org/10.1146/annurev-economics-091619-032800

Charness, G., & Grieco, D. (2026). Creativity and AI. *The Economic Journal, ueag015*. https://doi.org/10.1093/ej/ueag015

Charness, G., & Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association, 17*(2), 454-496. https://doi.org/10.1093/jeea/jvx055

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization, 149*, 74-87. https://doi.org/10.1016/j.jebo.2018.02.024

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. https://doi.org/10.2308/accr-50232

Chen, J., & Houser, D. (2017). Gender composition, stereotype and the contribution of ideas. *Available at SSRN*. http://dx.doi.org/10.2139/ssrn.2989049

Chen, D., Schonger, M., & Wickens, C. (2016). oTree - An open-source platform for laboratory, online and field experiments. *Journal of Behavioral and Experimental Finance, 9*, 88-97. https://doi.org/10.1016/j.jbef.2015.12.001

Chen, R., & Gong, J. (2018). Can self selection create high-performing teams?. *Journal of Economic Behavior & Organization, 148*, 20-33. https://doi.org/10.1016/j.jebo.2018.02.004

Chen, Y., Yu, C., Yuan, Y., Lu, F., & Shen, W. (2021). The influence of trust on creativity: A review. *Frontiers in Psychology, 12*, 1-10. https://doi.org/10.3389/fpsyg.2021.706234

Choudhury, P., Foroughi, C., & Larson, B. (2021). Work-from-anywhere: The productivity effects of geographic flexibility. *Strategic Management Journal, 42*(4), 655-683. https://doi.org/10.1002/smj.3251

Clark, L. A., Karau, S. J., & Michalisin, M. D. (2012). Telecommuting attitudes and the 'big five'personality dimensions. *Journal of Management Policy and Practice, 13*(3), 31-46.

Coenen, M., & Kok, R. A. (2014). Workplace flexibility and new product development performance: The role of telework and flexible work schedules. *European Management Journal, 32*(4), 564-576. https://doi.org/10.1016/j.emj.2013.12.003

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences 2nd ed.* Hillsdale, NJ: Erlbaum. https://doi.org/10.4324/9780203771587

Conti, R., Collins, M. A., & Picariello, M. L. (2001). The impact of competition on intrinsic motivation and creativity: Considering gender, gender segregation and gender role orientation. *Personality and individual differences, 31*(8), 1273-1289. https://doi.org/10.1016/S0191-8869(00)00217-8

Condly, S. J., Clark, R. E., & Stolovitch, H. D. (2003). The effects of incentives on workplace performance: A meta-analytic review of research studies 1. *Performance Improvement Quarterly, 16*(3), 46-63. https://doi.org/10.1111/j.1937-8327.2003.tb00287.x

Crenshaw, K. W. (2013). Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The Public Nature of Private Violence* (pp. 93-118). Routledge.

Croson, R. (2005). The Method of Experimental Economics. *International Negotiation*, *10*(1), 131-148. https://doi.org/10.1163/1571806054741100

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571. https://doi.org/10.1287/mnsc.32.5.554

Danzer, N. (2019). Job satisfaction and self-selection into the public or private sector: Evidence from a natural experiment. *Labour Economics*, *57*, 46-62. https://doi.org/10.1016/j.labeco.2019.01.002

Dargnies, M. P. (2012). Men too sometimes shy away from competition: The case of team competition. Management Science, 58(11), 1982-2000. https://doi.org/10.1287/mnsc.1120.1542

Datta Gupta, N., Poulsen, A., & Villeval, M.C. (2013). Gender Matching and competitiveness Experimental Evidence. *Economic Inquiry, 51*, 816-835. https://doi.org/10.1111/j.1465-7295.2011.00378.x

De Guinea, A. O., Webster, J., & Staples, D. S. (2012). A meta-analysis of the consequences of virtualness on team functioning. *Information & Management, 49*(6), 301-308. https://doi.org/10.1016/j.im.2012.08.003

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627-68; discussion 692-700. https://doi.org/10.1037/0033-2909.125.6.627

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior.* Springer Science & Business Media. https://doi.org/10.1007/978-1-4899-2271-7

Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology, 53*(6), 1024-1037. https://doi.org/10.1037//0022-3514.53.6.1024

del Carmen Triana, M., Song, R., Um, C. T., & Huang, L. (2024). Stereotypical perception in management: A review and expansion of role congruity theory. *Journal of Management, 50*(1), 188-215. https://doi.org/10.1177/01492063231180836

DellaVigna, S., & Pope, D. (2018). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies, 85*(2), 1029-1069. https://doi.org/10.1093/restud/rdx033

Dennis, A. R., & Williams, M. L. (2005). A meta-analysis of group side effects in electronic brainstorming: More heads are better than one. *International Journal of e-Collaboration (IJeC), 1*(1), 24-42. https://doi.org/ 10.4018/jec.2005010102

de Vries, H. B., & Lubart, T. I. (2019). Scientific creativity: divergent and convergent thinking and the impact of culture. *The Journal of Creative Behavior, 53*(2), 145-155. https://doi.org/10.1002/jocb.184

Dewett, T. (2006). Exploring the Role of Risk in Employee Creativity. *The Journal of Creative Behavior, 40*(1), 27-45. https://doi.org/10.1002/j.2162-6057.2006.tb01265.x

Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., & Rayner, M & Zahidi, S. (2023). Future of jobs report 2023. In *World Economic Forum*. https://www.weforum.org/reports/the-future-of-jobs-report-2023.

Di Mascio, R., Kalyuga, S., & Sweller, J. (2018). The effect of wording and placement of task instructions on problem-solving creativity. *The Journal of Creative Behavior, 52*(4), 335-353. https://doi.org/10.1002/jocb.157

Diehl, M., & Stroebe, W. (1991). Productivity loss in idea-generating groups: Tracking down the blocking effect. *Journal of Personality and Social Psychology, 61*(3), 392-403. https://doi.org/10.1037/0022-3514.61.3.392

Doherty, M. J., & Wimmer, M. C. (2005). Children's understanding of ambiguous figures: Which cognitive developments are necessary to experience reversal?. *Cognitive Development, 20*(3), 407-421. https://doi.org/10.1016/j.cogdev.2005.05.003

Dutcher, E. G. (2012). The effects of telecommuting on productivity: An experimental examination. The role of dull and creative tasks. *Journal of Economic Behavior & Organization, 84*(1), 355-363. https://doi.org/10.1016/j.jebo.2012.04.009

Dutcher, E. G., & Rodet, C. S. (2021). Remotely Creative? What Happens When Creative Teams Work Apart? *Available at SSRN*. https://dx.doi.org/10.2139/ssrn.3969124

Dutcher, G., & Rodet, C. S. (2022). Which two heads are better than one? Uncovering the positive effects of diversity in creative teams. *Journal of Economics & Management Strategy, 31*(4), 884-897. https://doi.org/10.1111/jems.12482

Dutcher, G., & Rodet, C. S. (2023). Creative capital generation: The role of diverse teams. Experience and Communication. *Available at SSRN.* http://dx.doi.org/10.2139/ssrn.4672457

Dutcher, E. G., Salmon, T. C., & Saral, K. J. (2024). Is “real” effort more real?. *Experimental Economics, 27*(5), 1001-1032. https://doi.org/10.1007/s10683-024-09844-y

Eagly, A. H. (1997). Sex differences in social behavior: comparing social role theory and evolutionary psychology. *American Psychologist, 52*(12), 1380-1383. https://doi.org/10.1037/0003-066X.52.12.1380.b

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3), 573-598. https://doi.org/10.1037/0033-295X.109.3.573

Eckartz, K., Kirchkamp, O., & Schunk, D. (2013). How Do Incentives Affect Creativity? (2013). *CESifo Working Paper Series No. 4049*. http://dx.doi.org/10.2139/ssrn.2198760

Eckel, C. C., & Grossman, P. J. (2005). Managing diversity by creating team identity. *Journal of Economic Behavior & Organization, 58*(3), 371-392. https://doi.org/10.1016/j.jebo.2004.01.003

Eisenberger, R. (1992). Learned industriousness. *Psychological Review*, *99*(2), 248-267. https://doi.org/10.1037/0033-295X.99.2.248

Eisenberger, R., & Shanock, L. (2003). Rewards, Intrinsic Motivation, and Creativity: A Case Study of Conceptual and Methodological Isolation. *Creativity Research Journal, 15*(2-3), 121-130. https://doi.org/10.1080/10400419.2003.9651404

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., & Schudy, S. (2024a). The Effect of Incentives in Nonroutine Analytical Team Tasks. *Journal of Political. Economy, 132*(8), 2695-2747. https://doi.org/10.1086/729443

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., & Schudy, S. (2024b). The Efficacy of Tournaments for Nonroutine Team Tasks. *Journal of Labor Economics, 42*(4), 921-948. https://doi.org/10.1086/725553

Erat, S., & Gneezy, U. (2016). Incentives for creativity. *Experimental Economics, 19*, 269-280. https://doi.org/10.1007/s10683-015-9440-5

Eriksson, T., Teyssier, S., & Villeval, M. C. (2009). Self-selection and the efficiency of tournaments. *Economic Inquiry, 47*(3), *530-548*. https://doi.org/10.1111/j.1465-7295.2007.00094.x

Exadaktylos, F., Espín, A. M., & Branas-Garza, P. (2013). Experimental subjects are not different. *Scientific reports, 3*(1), 1213. https://doi.org/10.1038/srep01213

Fairbank, J. F., & Williams, S. D. (2001). Motivating creativity and enhancing innovation through employee suggestion system technology. *Creativity and innovation management, 10*(2), 68-74. https://doi.org/10.1111/1467-8691.00204

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535-538. https://doi.org/10.1126/science.1168244

Feinstein, J. S. (2011). Optimal learning patterns for creativity generation in a field. *American Economic Review, 101*(3), 227–232. https://doi.org/10.1257/aer.101.3.227

Feinstein, J. S. (2023). *Creativity in Large-Scale Contexts: Guiding Creative Engagement and Exploration.* Stanford University Press.

Feist, G. (2019). The Function of Personality in Creativity: Updates on the Creative Personality. In J. Kaufman, & S. R. (Eds.), *The Cambridge Handbook of Creativity (Cambridge Handbooks in Psychology)* (pp. 353-373). Cambridge: Cambridge University Press.

Felstead, A., & Henseke, G. (2017). Assessing the growth of remote working and its consequences for effort, well-being and work-life balance. *New Technology, Work and Employment, 32*(3), 195-212. https://doi.org/10.1111/ntwe.12097

Ficapal-Cusí, P., Torrent-Sellens, J., Palos-Sanchez, P., & González-González, I. (2023). The telework performance dilemma: exploring the role of trust, social isolation and fatigue. *International Journal of Manpower, 45*(1), 156-168. https://doi.org/10.1108/IJM-08-2022-0363

Fischer, C., Malycha, C. P., & Schafmann, E. (2019). The Influence of Intrinsic Motivation and Synergistic Extrinsic Motivators on Creativity and Innovation. *Frontiers in Psychology, 10*, 137. https://doi.org/10.3389/fpsyg.2019.00137

Ford, C. M. (1996). A theory of individual creative action in multiple social domains. *Academy of Management Review, 21*(4), 1112-1142. https://doi.org/10.2307/259166

Forthmann, B., Szardenings, C., & Holling, H. (2020). Understanding the confounding effect of fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. *Psychology of Aesthetics, Creativity, and the Arts, 14*(1), 94-112. https://doi.org/10.1037/aca0000196

Franke, G. R., Crown, D. F., & Spake, D. F. (1997). Gender differences in ethical perceptions of business practices: A social role theory perspective. *Jornal of Applied Psychology, 82*(6), 920-934. https://doi.org/10.1037/0021-9010.82.6.920

Frey, B. S., Gibbs, M., Neckermann, S., Bradler, C., Warnke, A. J., & Siemroth, C. (2013). Incentives and creativity (ZEW policy brief 4/2013). *Mannheim: Zentrum für Europäische Wirtschaftsforschung (ZEW)*. https://www.econstor.eu/handle/10419/108745

Freyer, T., Radbruch, J., Schaube, S., & Strang, L. (2025). The effect of task (mis) matching and self-selection on intrinsic motivation and performance. *Journal of Economic Behavior & Organization*, 234, 107010. https://doi.org/10.1016/j.jebo.2025.107010

Friedman, D., & Cassar, A. (2004). *Economics Lab*. London: Routledge. https://doi.org/10.4324/9780203356845

Fullerton, R. L., & McAfee, R. P. (1999). Auctionin Entry into Tournaments. *Journal of Political Economy, 107*(3), 573-605. https://doi.org/10.1086/250072

Gächter, S., Starmer, C., & Tufano, F. (2025). Measuring group cohesion to reveal the power of social relationships in team production. *Review of Economics and Statistics, 107*(2), 539-554. https://doi.org/10.1162/rest_a_01283

Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior, 26*(4), 331-362. https://doi.org/10.1002/job.322

Gajendran, R. S., & Harrison, D. A. (2007). The good, the bad, and the unknown about telecommuting: meta-analysis of psychological mediators and individual consequences. *Journal of Applied Psychology, 92*(6), 1524-1541. https://doi.org/10.1037/0021-9010.92.6.1524

Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., & Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 529-536). https://doi.org/10.1145/642611.642703

Garro-Abarca, V., Palos-Sanchez, P., & Aguayo-Camacho, M. (2021). Virtual teams in times of pandemic: Factors that influence performance. *Frontiers in Psychology, 12*, 624-637. https://doi.org/10.3389/fpsyg.2021.624637

Gavoille, N., & Hazans, M. (2022). Personality traits, remote work and productivity. *IZA Discussion Paper, 15486*, 1-49. https://doi.org/10.16995/ilr.18853

Georganas, S., Tonin, M., & Vlassopoulos, M. (2015). Peer pressure and productivity: The role of observing and being observed. *Journal of Economic Behavior & Organization, 117*, 223-232. https://doi.org/10.1016/j.jebo.2015.06.014

Geraldes, D. (2020). Women dislike competing against men. *Available at SSRN 3741649*. https://dx.doi.org/10.2139/ssrn.3741649

Gerhart, B., & Fang, M. (2015). Pay, Intrinsic Motivation, Extrinsic Motivation, Performance, and Creativity in the Workplace: Revisiting Long-Held Beliefs. *Annual Review of Organizational Psychology and Organizational Behavior, 2*(1), 489-521. https://doi.org/10.1146/annurev-orgpsych-032414-111418

Gerrard, B., & Lockett, A. (2018). Team-specific human capital and performance. *British Journal of Management, 29*(1), 10-25. https://doi.org/10.1111/1467-8551.12173

Giambatista, R. C., & Bhappu, A. D. (2010). Diversity's harvest: Interactions of diversity sources and communication technology on creative group performance. *Organizational Behavior and Human Decision Processes, 111*(2), 116-126. https://doi.org/10.1016/j.obhdp.2009.11.003

Gibbons, R. (1998). Incentives in organizations. *Journal of Economic Perspectives, 12*(4), 115-132. https://doi.org/10.1257/jep.12.4.115

Gibbs, M., Neckermann, S., & Siemroth, C. (2017). A field experiment in motivating employee ideas. *Review of Economics and Statistics, 99*(4), 577-590. https://doi.org/10.1162/rest_A_00631

Girotra, K., Meincke, L., Terwiesch, C., & Ulrich, K. T. (2023). Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN*, *4526071*.

Girotra, K., Terwiesch, C., & Ulrich, K. T. (2010). Idea generation and the quality of the best idea. *Management Science, 56*(4), 591-605. https://doi.org/10.2139/ssrn.1082392

Gneezy, U., Laske, K., & Schröder, M. (2022). Teams do not outperform individuals in a simple creative task. *Applied Economics Letters, 29*(15), 1393-1398. https://doi.org/10.1080/13504851.2021.1934386

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics, 118*(3), 1049-1074. https://doi.org/10.1162/00335530360698496

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics, 115*(3), 791-810. https://doi.org/10.1162/003355300554917

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26-42. https://doi.org/10.1037/1040-3590.4.1.26

Gough, H. G. (1979). A creative personality scale for the Adjective Check List. *Journal of Personality and Social Psychology, 37*(8), 1398-1405. https://doi.org/10.1037/0022-3514.37.8.1398

Gross, D. P. (2020). Creativity under fire: The effects of competition on creative production. *Review of Economics and Statistics, 102*(3), 583-599. https://doi.org/10.1162/rest_a_00831

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. https://doi.org/10.1016/j.jebo.2020.09.009

Grund, C., Harbring, C., & Klinkenberg, L. (2025). An experiment on creativity in virtual teams. *Journal of Economic Behavior & Organization, 231*, 106926. https://doi.org/10.1016/j.jebo.2025.106926

Grund, C., Harbring, C., & Klinkenberg, L. (2026). Creative Performance: A comparison of individuals and virtual teams – an Experimental Study. *Journal of Business Economics*. https://doi.org/10.1007/s11573-025-01254-7

Guala, F. (2005*). The methodology of experimental economics*. Cambridge University Press.

Guilford, J. P. (1950). Creativity. *American Psychologist, 5*, 444-454. https://doi.org/10.1037/h0063487

Guilford, J. P. (1959). *Personality.* New York: McGraw-Hill.

Guilford, J. P. (1960). Basic conceptual problems of the psychology of thinking. *Proceedings of the New York Academy of Sciences, 91*, 6-21. https://doi.org/10.1111/j.1749-6632.1961.tb50913.x

Guilford, J. P. (1967). *The Nature of Human Intelligence.* New York: McGraw-Hill.

Gunnthorsdottir, A., & Rapoport, A. (2006). Embedding social dilemmas in intergroup competition reduces free-riding. *Organizational Behavior and Human Decision Processes, 101*(2), 184-199. https://doi.org/10.1016/j.obhdp.2005.08.005

Hacker, J. V., Johnson, M., Saunders, C., & Thayer, A. L. (2019). Trust in virtual teams: A multidisciplinary review and integration. *Australasian Journal of Information Systems, 23*. https://doi.org/10.3127/ajis.v23i0.1757

Hakanen, M., & Soudunsaari, A. (2012). Building trust in high-performing teams. *Technology Innovation Management Review, 2*(6), 38-41. http://doi.org/10.22215/timreview/567

Harbring, C., & Irlenbusch, B. (2003). An experimental study on tournament design. *Labour Economics, 10*(4), 443-464. https://doi.org/10.1016/s0927-5371(03)00034-4

Harbring, C., & Irlenbusch, B. (2008). How many winners are good to have?: On tournaments with sabotage. *Journal of Economic Behavior & Organization, 65*(3-4), 682-702. https://doi.org/10.1016/j.jebo.2006.03.004

Hardt, D., Mayer, L., & Rincke, J. (2025). Who does the talking here? The impact of gender composition on team interactions. *Management Science, 71*(5), 4131-4152. https://doi.org/10.1287/mnsc.2023.03411

Harrington, D. M. (1975). Effects of explicit instructions to "be creative" on the psychological meaning of divergent thinking test scores . 1. *Journal of Personality, 43*(3), 434-454. https://doi.org/10.1111/j.1467-6494.1975.tb00715.x

Harris, A. M., Williamson, R. L., & Carter, N. T. (2019). A conditional threshold hypothesis for creative achievement: On the interaction between intelligence and openness. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(3), 322-337. https:// doi.org/10.1037/aca0000182

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009-1055. DOI: 10.1257/0022051043004577

Harvey, S., & Berry, J. W. (2023). Toward a meta-theory of creativity forms: How novelty and usefulness shape creativity. *Academy of Management Review, 48*(3), 504-529. https://doi.org/10.5465/amr.2020.0110

Hennessey, B. A. (2010). The creativity-motivation connection. *The Cambridge Handbook of Creativity*, 342-365. https://doi.org/10.1017/CBO9780511763205.022

Herring, C. (2009). Does diversity pay?: Race, gender, and the business case for diversity. *American Sociological Review, 74*(2), 208-224. https://doi.org/10.1177/000312240907400203

Hill, N. S., Bartol, M., K., Tesluk, P. E., & Langa, G. A. (2009). Organizational context and face-to-face interaction: Influences on the development of trust and collaborative behaviors in computer-mediated groups. *Organizational Behavior and Human Decision Processes, 108*(2), 187-201. https://doi.org/10.1016/j.obhdp.2008.10.002

Ho, F., & Mussap, A. J. (2019). The Gender Identity Scale: Adapting the Gender Unicorn to measure gender identity. *Psychology of Sexual Orientation and Gender Diversity, 6*(2), 217-231. https://doi.org/10.1037/sgd0000322

Hocevar, D. (1979). Ideational fluency as a confounding factor in the measurement of originality. *Journal of Educational Psychology, 71*(2), 191-196. https://doi.org/10.1037/0022-0663.71.2.191

Hoever, I. J., Van Knippenberg, D., Van Ginkel, W. P., & Barkema, H. G. (2012). Fostering team creativity: perspective taking as key to unlocking diversity's potential. *Journal of Applied Psychology, 97*(5), 982-996. https://doi.org/10.1037/a0029159

Hoffmann, C., & Thommes, K. (2022). Combining egalitarian and proportional sharing rules in team tournaments to incentivize energy-efficient behavior in a principal-agent context. *Organization & Environment, 35*(2), 307-331. https://doi.org/10.1177/1086026620945343

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 74-91. https://doi.org/10.2307/3003320

Holmström, B. (2017). Pay for performance and beyond. *American Economic Review, 107*(7), 1753-1777. https://doi.org/10.1257/aer.107.7.1753

Holmström, B., & Milgrom, P. (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization, 7*(special_issue), 24-52. https://doi.org/10.1093/jleo/7.special_issue.24

Hoogendoorn, S., Oosterbeek, H., & Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science, 59*(7), 1514-1528. https://doi.org/10.2139/ssrn.1826024

Hora, S., Lemoine, G. J., Xu, N., & Shalley, C. E. (2021). Unlocking and Closing the gender gap in creative performance: A multilevel model. *Journal of Organizational Behavior, 42*(3), 297-312. https://doi.org/10.1002/job.2500

Horwitz, F. M., Bravington, D., & Silvis, U. (2006). The promise of virtual teams: identifying key factors in effectiveness and failure. *Journal of European Industrial Training, 30*(6), 472-494. https://doi.org/10.1108/03090590610688843

Hughes, D. J., Lee, A., Tian, A. W., Newman, A., & Legood, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly, 29*(5), 549-569. https://doi.org/10.1016/j.leaqua.2018.03.001

Hunter, P. (2019). Remote working in research: An increasing usage of flexible work arrangements can improve productivity and creativity. *EMBO reports, 20*(1) *e47435*, 1-4. https://doi.org/10.15252/embr.201847435

Huo, K. (2020). Performance incentives, divergent thinking training, and creative problem solving. *Journal of Management Accounting Research*, *32*(1), 159-176. https://doi.org/10.2308/jmar-52479

Iaria, A., Schwarz, C., & Waldinger, F. (2024). Gender gaps in academia: Global evidence over the twentieth century. *Available at SSRN.* https://doi.org/10.2139/ssrn.4150221

Iqbal, K. M. J., Khalid, F., & Barykin, S. Y. (2021). Hybrid workplace: The future of work. In *Handbook of Research on Future Opportunities for Technology Management Education* (pp. 28-48). IGI Global. https://doi.org/10.4018/978-1-7998-8327-2.ch003

Isaksen, S. G., & Gaulin, J. P. (2005). A reexamination of brainstorming research: Implications for research and practice. *Gifted Child Quarterly, 49*(4), 315-329. https://doi.org/10.1177/001698620504900405

Ivanova-Stenzel, R., & Kübler, D. (2011). Gender differences in team work and team competition. *Journal of Economic Psychology, 32*(5), 797-808. https://doi.org/10.1016/j.joep.2011.05.011

Jang, E., & Kim, Y. C. (2025). Autonomy Constrained: The Dynamic Interplay Among Job Autonomy, Work Engagement, and Innovative Behavior Under Performance Pressure. *Administrative Sciences, 15*(3), 97. https://doi.org/10.3390/admsci15030097

Janis, I. (1972). *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes.* Boston: Houghton Mifflin. https://doi.org/10.2307/3791464

Jaravel, X., Petkova, N., & Bell, A. (2018). Team-specific capital and innovation. *American Economic Review, 108*(4-5), 1034-1073. https://doi.org/10.1257/aer.20151184

Jenkins Jr, G. D., Mitra, A., Gupta, N., & Shaw, J. D. (1998). Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology*, *83*(5), 777-787. https://doi.org/10.1037/0021-9010.83.5.777

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics, 3*(4), 305-360. https://doi.org/10.1016/0304-405X(76)90026-X

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. https://doi.org/10.1111/j.1475-679X.2008.00277.x

Kachelmeier, S. J., Wang, L. W., & Williamson, M. G. (2019). Incentivizing the Creative Process: From Initial Quantity to Eventual Creativity. *The Accounting Review, 94*(2), 249-266. https://doi.org/10.2308/accr-52196

Kachelmeier, S. J., Webb, R. A., & Williamson, M. G. (2024). Do Performance-Contingent Incentives Help or Hinder Divergent Thinking? *The Accounting Review, 99*(2), 229-248. https://doi.org/10.2308/TAR-2020-0751

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. https://doi.org/10.2308/accr.2010.85.5.1669

Kachelmeier, S. J., & Williamson, M. G., & Zhang, X. (2023). We're in this together: The motivational effects of tangible rewards in a group setting. *Contemporary Accounting Research, 49*(2), 842-867. https://doi.org/10.1111/1911-3846.12844

Kaduk, A., Genadek, K., Kelly, E. L., & Moen, P. (2019). Involuntary vs. voluntary flexible work: insights for scholars and stakeholders. *Community, Work & Family, 22*(4), 412-442. https://doi.org/10.1080/13668803.2019.1616532

Kahn, P. H., Friedman, B., Gill, B., Hagman, J., Severson, R. L., Freier, N. G., Feldman, E. N., Carrère, S., & Stolyar, A. (2008). A plasma display window?—The shifting baseline problem in a technologically mediated natural world. *Journal of Environmental Psychology, 28*(2), 192-199. https://doi.org/10.1016/j.jenvp.2007.10.008

Kaivola-Bregenhøj, A. (2001). *Riddles: Perspectives on the use, function and change in a folklore genre.* Finnish Literature Society. https://doi.org/10.21435/sff.10

Kandel, E., & Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of political Economy, 100*(4), 801-817. https://doi.org/10.1086/261840

Kennedy, D. M., Vozdolska, R. R., & McComb, S. A. (2010). Team decision making in computer-supported cooperative work: How initial computer-mediated or face-to-face meetings set the stage for later outcomes. *Decision Sciences, 41*(4), 933-954. https://doi.org/10.1111/j.1540-5915.2010.00293.x

Kerr, D. S., & Murthy, U. S. (2004). Divergent and convergent idea generation in teams: A comparison of computer-mediated and face-to-face communication. *Group Decision and Negotiation, 13*, 381-399. https://doi.org/10.1023/B:GRUP.0000042960.38411.52

Kim, J. H. (2025). Experimental examination of the incentive and sorting effects of pay-for-performance on creative performance. *Journal of Applied Psychology, 110*(4), 598-617. https://doi.org/10.1037/apl0001245

Kim, J. H., Gerhart, B., & Fang, M. (2022). Do financial incentives help or harm performance in interesting tasks?. *Journal of Applied Psychology, 107*(1), 153-167. https://doi.org/10.1037/apl0000851

Kim, J. J., Vaulont, M. J., Zhang, Z., & Byron, K. (2024). Looking inside the black box of gender differences in creativity: A dual-process model and meta-analysis. *Journal of Applied Psychology, 109*(12), 1861-1900. https://doi.org/10.1037/apl0001205

King, L. A., Walker, L. M., & Broyles, S. J. (1996). Creativity and the five-factor model. *Journal of Research in Personality, 30*(2), 189-203. https://doi.org/10.1006/jrpe.1996.0013

Ko, S., & Butler, J. E. (2007). Creativity: A key link to entrepreneurial behavior. *Business Horizons, 50*(5), 365-372. https://doi.org/10.1016/j.bushor.2007.03.002

Kohn, N. W., Paulus, P. B., & Choi, Y. (2011). Building on the ideas of others: An examination of the idea combination process. *Journal of Experimental Social Psychology, 47*(3), 554-561. https://doi.org/10.1016/j.jesp.2011.01.004

Kozlowski, S. W. J., & Bell, B. S. (2013). Work groups and teams in organizations: Review update.

Krishna, V., & Morgan, J. (1998). The winner-take-all principle in small tournaments. *Advances in Applied Microeconomics* (7), 61-74.

Kube, S., Maréchal, M. A., & Puppe, C. (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review, 102*(4), 1644-1662. https://doi.org/10.1257/aer.102.4.1644

Kuhn, P. (2018). *Personnel Economics.* Oxford: Oxford University Press Inc.

Kuhn, P., & Villeval, M. C. (2015). Are women more attracted to co-operation than men? *The Economic Journal, 125*(582), 115-140. https://doi.org/10.1111/ecoj.12122

KPMG. (2023, October 11). *Global CEOs grappling with multiple challenges, including geopolitical pressures, but majority remain confident on three-year outlook*. Retrieved from https://kpmg.com/xx/en/home/media/press-releases/2023/09/ceo-outlook.html

Kurland, N. B., & Bailey, D. E. (1999). The advantages and challenges of working here, there, anywhere, and anytime. *Organizational Dynamics, 28*(2), 53-67. https://doi.org/10.1016/S0090-2616(00)80016-9

Kurtzberg, T. R. (1998). Creative thinking, a cognitive aptitude, and integrative joint gain: A study of negotiator creativity. *Creativity Research Journal, 11*(4), 283-293. https://doi.org/10.1207/s15326934crj1104_2

Kurtzberg, T. R. (2005). Feeling creative, being creative: An empirical study of diversity and creativity in teams. *Creativity Research Journal, 17*(1), 51-65. https://doi.org/10.1207/s15326934crj1701_5

Kurtzberg, T. R., & Amabile, T. M. (2001). From Guilford to creative synergy: Opening the black box of team-level creativity. *Creativity Research Journal, 13*(3-4), 285-294. https://doi.org/10.1207/S15326934CRJ1334_06

Labrado Antolín, M., Rodríguez-Ruiz, Ó., & Fernández Menéndez, J. (2024). A time after time effect in telework: an explanation of willingness to telework and self-reported productivity. *International Journal of Manpower, 45*(1), 200-214. https://doi.org/10.1108/IJM-05-2022-0238

Laske, K., Römer, N., & Schröder, M. (2024). Piece-Rate Incentives and Idea Generation–An Experimental Analysis. *CESifo Working Paper Series 11594*. https://dx.doi.org/10.2139/ssrn.5097046

Lasrado, F., Arif, M., Rizvi, A., & Urdzik, C. (2016). Critical success factors for employee suggestion schemes: a literature review. *International Journal of Organizational Analysis, 24*(2), 315-339. https://doi.org/10.1108/IJOA-04-2014-0753

Lazear, E. P. (2000). Performance Pay and Productivity. *American Economic Review, 90*(5), 1346-1361. https://doi.org/10.1257/aer.90.5.1346

Lazear, E. P. (2018). Compensation and incentives in the workplace. *Journal of Economic Perspectives, 32*(3), 195-214. https://doi.org/10.1257/jep.32.3.195

Lazear, E. P., & Rosen, S. (1981). Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy, 89*(5), 841-864. https://doi.org/10.1086/261010

Lee, H. W., Kim, J., Lee, H., Triana, M. D. C., & Choi, J. N. (2025). When employee gender diversity benefits collective performance: The importance of the proportion of women in top management teams. *Personnel Psychology*, 78(3), 305-330. https://doi.org/10.1111/peps.12677

Leroy, H., Buengeler, C., Veestraeten, M., Shemla, M., & Hoever, I. J. (2022). Fostering team creativity through team-focused inclusion: The role of leader harvesting the benefits of diversity and cultivating value-in-diversity beliefs. *Group & Organization of Management, 47*(4), 789-839. https://doi.org/10.1177/10596011211009683

Lewandowski, P., Lipowska, K., & Smoter, M. (2022). Working from Home During a Pandemic-A Discrete Choice Experiment in Poland. *IZA Discussion Paper Series*, 1-40. https://doi.org/10.2139/ssrn.4114827

Lewis, C. A., Grahlow, M., Kühnel, A., Derntl, B., & Kroemer, N. B. (2023). Women compared with men work harder for small rewards. *Scientific Reports, 13*(1), 5456. https://doi.org/10.1038/s41598-023-32391-0

Li, F., Chen, T., & Lai, X. (2018). How Does a Reward for Creativity Program Benefit or Frustrate Employee Creative Performance? The Perspective of Transactional Model of Stress and Coping. *Group & Organization Management, 43*(1), 138-175. https://doi.org/10.1177/1059601116688612

Lim, V. K., & Teo, T. S. (2000). To work or not to work at home-An empirical investigation of factors affecting attitudes towards teleworking. *Journal of Managerial Psychology, 15*(6), 560-586. https://doi.org/10.1108/02683940010373392

Liu, F., Li, P., Taris, T. W., & Peeters, M. C. (2022). Creative performance pressure as a double-edged sword for creativity: The role of appraisals and resources. *Human Resource Management, 61*(6), 663-679. https://doi.org/10.1002/hrm.22116

Liu, Y., Keller, R. T., & Shih, H. A. (2011). The impact of team-member exchange, differentiation, team commitment, and knowledge sharing on R&D project team performance. *R&D Management, 41*(3), 274-287. https://doi.org/10.1111/j.1467-9310.2011.00636.x

Long, H., & Wang, J. (2022). Dissecting reliability and validity evidence of subjective creativity assessment: A literature review. *Educational Psychology Review, 34*(3), 1399-1443. https://doi.org/10.1007/s10648-022-09679-0

Lott, Y., & Peters, E. (2026). Zurück ins Büro? - Verbreitung, Hintergründe und Folgen von Return-to-Office-Initiativen. *WSI Policy Brief Nr. 96*.

Lu, L. (2015). Building trust and cohesion in virtual teams: the developmental approach. *Journal of Organizational Effectiveness: People and Performance, 2*(1), 55-72. https://doi.org/10.1108/JOEPP-11-2014-0068

Lubart, T. I. (2001). Models of the Creative Process: Past, Present and Future. *Creative Research Journal, 13*(3–4), 295-308. https://doi.org/10.1207/S15326934CRJ1334_07

Lubart, T. I., & Sternberg, R. J. (1995). An investment approach to creativity: Theory and data. In T. B. S. M. Smith, *The Creative Cognition Approach* (pp. 271-302). The MIT Press.

Lubczyk, M., & Moser, P. (2024). The Ms. Allocation of Talent. *NBER Working Paper*.

Lurey, J. S., & Raisinghani, M. S. (2001). An empirical study of best practices in virtual teams. *Information & Management, 38*(8), 523-544. https://doi.org/10.1016/S0378-7206(01)00074-X

Majerczyk, M., Sheremeta, R., & Tian, Y. (2019). Adding tournament to tournament: Combining between-team and within-team incentives. *Journal of Economic Behavior & Organization, 166*, 1-11. https://doi.org/10.1016/j.jebo.2019.09.002

Malik, M. A. R., Butt, A. N., & Choi, J. N. (2015). Rewards and employee creative performance: Moderating effects of creative self-efficacy, reward importance, and locus of control. *Journal of Organizational Behavior, 36*(1), 59-74. https://doi.org/10.1002/job.1943

Manske, M. E., & Davis, G. A. (1968). Effects of simple instructional biases upon performance in the unusual uses test. *The Journal of General Psychology, 79*(1), 25-33. https://doi.org/10.1080/00221309.1968.9710449

Martens, Y. (2011). Creative workplace: instrumental and symbolic support for creativity. *Facilities, 29*(1/2), 63-79. https://doi.org/10.1108/02632771111101331

Martin, M. M., & Rubin, R. B. (1995). A New Measure of Cognitive Flexibility. *Psychological Reports, 76*(2), 623-626. https://doi.org/10.2466/pr0.1995.76.2.623

Mas, A., & Pallais, A. (2017). Valuing alternative work arrangements. *American Economic Review, 107*(12), 3722-3759. https://doi.org/10.1257/aer.20161500

Matobobo, C. (2026). A systematic review of gender differences in students' use of AI tools for learning in higher education. Discover Education, 5(1), 90. https://doi.org/10.1007/s44217-026-01116-6

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52*(6), 1258-1265. https://doi.org/10.1037/0022-3514.52.6.1258

McEvily, B., Perrone, V., & Zaheer, A. (2003). Trust as an organizing principle. *Organization Science, 14*(1), 91-103. https://doi.org/10.1287/orsc.14.1.91.12814

McGee, A. D., & McGee, P. J. (2025). Gender and race differences on incentivized personality measures. *Frontiers in Behavioral Economics*, *4*. 1499464. https://doi.org/10.3389/frbhe.2025.1499464

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220-232. https://doi.org/10.1037/h0048850

Memmert, D., Plessner, H., Hüttermann, S., Froese, G., Peterhänsel, C., & Unkelbach, C. (2015). Collective fit increases team performances: Extending regulatory fit from individuals to dyadic teams. *Journal of Applied Social Psychology, 45*(5), 274-281. https://doi.org/10.1111/jasp.12294

Messenger, J. C., & Gschwind, L. (2016). Three generations of Telework: New ICT s and the (R) evolution from Home Office to Virtual Office. *New Technology, Work and Employment, 31*(3), 195-208. https://doi.org/10.1111/ntwe.12073

Methot, J. R., Lepine, J. A., Podsakoff, N. P., & Christian, J. S. (2016). Are workplace friendships a mixed blessing? Exploring tradeoffs of multiplex relationships and their associations with job performance. *Personnel Psychology, 69*(2), 311-355. https://doi.org/10.1111/peps.12109

Meyerson, D., Weick, K. E., & Kramer, R. M. (1996). Swift trust and temporary groups. In R. M. Kramer, & T. R. Tyler, *Trust in organizations: Frontiers of Theory and Research* (pp. 166-195). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781452243610.n9

Migheli, M. (2015). Gender at work: Incentives and self-sorting. *Journal of Behavioral and Experimental Economics, 55*, 10-18. https://doi.org/10.1016/j.socec.2014.12.005

Minbashian, A., Wood, R. E., & Beckmann, N. (2010). Task-contingent conscientiousness as a unit of personality at work. *Journal of Applied Psychology, 95*(5), 793–806. https://doi.org/10.1037/a0020016

Møgelvang, A., Bjelland, C., Grassini, S., & Ludvigsen, K. (2024). Gender differences in the use of generative artificial intelligence chatbots in higher education: Characteristics and consequences. *Education Sciences, 14*(12), 1363. https://doi.org/10.3390/educsci14121363

Mohnen, A., & Ostermaier, A. (2013). Incentives for creativity: Limits of objective performance evaluation. *Available at SSRN 2210321*. http://dx.doi.org/10.2139/ssrn.2210321

Morgan, J., & Wang, R. (2010). Tournaments for Ideas. *California Management Review, 52*(2), 77-97. https://doi.org/10.1525/cmr.2010.52.2.77

Mullen, B., Johnson, C., & Salas, E. (1991). Productivity Loss in Brainstorming Groups: A Meta-analytic Integration. *Basic and Applied Social Psychology, 12* (1), 3-23. https://doi.org/10.1207/s15324834basp1201_1

Mumford, M. D. (2003). Where have we been, where are we going? Taking stock in creativity research. *Creativity Research Journal, 15*(2-3), 107-120. https://doi.org/10.1080/10400419.2003.9651403

Mumford, M. D., & Gustafson, S. B. (1988). Creativity syndrome: Integration, application, and innovation. *Psychological Bulletin, 103*(1), 27-43. https://psycnet.apa.org/doi/10.1037/0033-2909.103.1.27

Nakrošienė, A., Bučiūnienė, I., & Goštautaitė, B. (2019). Working from home: characteristics and outcomes of telework. *International Journal of Manpower, 40*(1), 87-101. https://doi.org/10.1108/IJM-07-2017-0172

Nalbantian, H. R., & Schotter, A. (1997). Productivity under group incentives: An experimental study. *The American Economic Review*, 87(3), 314-341. https://www.jstor.org/stable/2951348

Naotunna, N. P., & Zhou, E. (2022). Telecommuting and Creativity of Professional Employees in Software Developing Industry in Sri Lanka. *Asian Journal of Management Studies, 2*(1), 28-45. http://doi.org/10.4038/ajms.v2i1.42

Naqshbandi, M. M., Kabir, I., Ishak, N. A., & Islam, M. Z. (2023). The future of work: work engagement and job performance in the hybrid workplace. *The Learning Organization, 31*(1), 5-26. https://doi.org/10.1108/TLO-08-2022-0097

Neininger, A., Lehmann-Willenbrock, N., Kauffeld, S., & Henschel, A. (2010). Effects of team and organizational commitment–A longitudinal study. *Journal of Vocational Behavior, 76*(3), 567-579. https://doi.org/10.1016/j.jvb.2010.01.009

Nelson, J., & Guegan, J. (2022). Creative collaboration in groups. In *Homo Creativus: The 7 C's of Human Creativity* (pp. 87-105). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-99674-1

Nelson, T., & McFadzean, E. (1998). Facilitating problem-solving groups: facilitator competences. *Leadership & Organization Development Journal, 19*(2), 72-82. https://doi.org/10.1108/01437739810208647

Nemiro, J. E. (2002). The creative process in virtual teams. *Communication Research Journal, 14*(1), 69-83. https://doi.org/10.1207/S15326934CRJ1401_6

Nguyen, M. H. (2021). Factors influencing home-based telework in Hanoi (Vietnam) during and after the COVID-19 era. *Transportation, 48*(6), 3207-3238. https://doi.org/10.1007/s11116-021-10169-5

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much?. *The Quarterly Journal of Economics, 122*(3), 1067-1101. https://doi.org/10.1162/qjec.122.3.1067

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review Economics*, *3*(1), 601-630. https://doi.org/10.1146/annurev-economics-111809-125122

Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review, 10*(3), 186-213. https://doi.org/10.1207/s15327957pspr1003_1

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2014). Ready, set, create: What instructing people to "be creative" reveals about the meaning and mechanisms of divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts, 8*, 423-432. https://psycnet.apa.org/doi/10.1037/a0036549

Opitz, S., Sliwka, D., Vogelsang, T., & Zimmermann, T. (2024). The algorithmic assignment of incentive schemes. *Management Science*, *71*(2), 1546-1563. https://doi.org/10.1287/mnsc.2022.03362

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity, 49*, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Osborn, A. F. (1953). *Applied Imagination.* NY: Scribners.

Paoletti, J., Gilberto, J. M., Beier, M. E., & Salas, E. (2020). The role of aging, age diversity, and age heterogeneity within teams. *Current and Emerging Trends in Aging and Work*, 319-336. https://doi.org/10.1007/978-3-030-24135-3_16

Papke, L.E. and Wooldridge, J.M. (1996), Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics, 11*, 619-632. https://doi.org/10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1

Paulus, P. B., (2000). Groups, Teams, and Creativity: The Creative Potential of Idea-generating Groups. *Applied Psychology, 49*(2), 237-262. https://doi.org/10.1111/1464-0597.00013

Paulus, P. B., & Brown, V. R. (2007). Toward more creative and innovative group idea generation: A cognitive-social-motivational perspective of brainstorming. *Social and Personality Psychology Compass, 1*(1), 248-265. https://doi.org/10.1111/j.1751-9004.2007.00006.x

Paulus, P. B., & Dzindolet, M. T. (1993). Social influence processes in group brainstorming. *Journal of Personality and Social Psychology, 64*(4), 575-586. https://doi.org/10.1037/0022-3514.64.4.575

Paulus, P. B., & Dzindolet, M. (2008). Social influence, creativity and innovation. *Social Influence, 3*(4), 228-247. https://doi.org/10.1080/15534510802341082

Paulus, P. B., Dzindolet, M., & Kohn, N. W. (2012). Collaborative creativity-Group creativity and team innovation. In P. B. Paulus, M. Dzindolet, & N. W. Kohn, *Handbook of Organizational Creativity* (pp. 327-357). Academic Press. https://doi.org/10.1016/B978-0-12-374714-3.00014-8

Paulus, P. B., & Kenworthy, J. B. (2022). Research findings on ideational creativity in groups. In *Creativity and innovation: Cognitive, social, and computational approaches* (pp. 47-67). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-77198-0_3

Paulus, P. B., Kohn, N. W., & Arditti, L. E. (2011). Effects of quantity and quality instructions on brainstorming. *The Journal of Creative Behavior, 45*(1), 38-46. https://doi.org/10.1002/j.2162-6057.2011.tb01083.x

Penagos-Corzo, J. C., & Saucedo, A. (2021). Development and Validation of a Questionnaire to Assess Creative Potential Traits. *Frontiers in Psychology, 12*, 756-779. https://doi.org/10.3389/fpsyg.2021.756079

Peñarroja, V., Orengo, V., Zornoza, A., & Hernández, A. (2013). The effects of virtuality level on task-related collaborative behaviors: The mediating role of team trust. *Computers in Human Behavior, 29*(3), 967-974. https://doi.org/10.1016/j.chb.2012.12.020

Pescosolido, A. T., & Saavedra, R. (2012). Cohesion and Sports Teams. *Small Group Research, 43*(6), 744-758. https://doi.org/10.1177/1046496412465020

Peter, L., Michinov, N., Besancon, M., Michinov, E., Juhel, J., Brown, G., … & Desrichard Oliver. (2021). Revisiting the effects of gender diversity in small groups on divergent thinking: A large-scale study using synchronous electronic brainstorming. *Frontiers in Psychology*, 12, 723235. https://doi.org/10.3389/fpsyg.2021.723235

Pfister, J. A., & Lukka, K. (2019). Interrelation of controls for autonomous motivation: A field study of productivity gains through pressure-induced process innovation. *The Accounting Review, 94*(3), 345-371. https://doi.org/10.2308/accr-52266

Pillemer, J., & Rothbard, N. P. (2018). Friends without benefits: Understanding the dark sides of workplace friendship. *Academy of Management Review, 43*(4), 635-660. https://doi.org/10.5465/amr.2016.0309

Pirola-Merlo, A., & Mann, L. (2004). The relationship between individual creativity and team creativity: Aggregating across people and time. *Journal of Organizational Behavior, 25*(2), 235-257. https://doi.org/10.1002/job.240

Plucker, J. A. (2023). The patient is thriving! current issues, recent advances, and future directions in creativity assessment. *Creativity Research Journal, 35*(3), 291-303. https://doi.org/10.1080/10400419.2022.2110415

Plucker, J. A., Qian, M., & Wang, S. (2011). Is originality in the eye of the beholder? Comparison of scoring techniques in the assessment of divergent thinking. *Journal of Creative Behavior, 45*, 1-22. https://doi.org/10.1002/j.2162-6057.2011.tb01081.x

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37*(1), 7-63. https://doi.org/10.1257/jel.37.1.7

Price, R., & Lovka, B. (2000). *Droodles: The classic collection.* Tallfellow Press.

Prokofieva, V., Fenouillet, F., & Romero, M. (2024). The effects of assessment instruction and test anxiety on divergent components of creative problem-solving tasks. *Frontiers in Education* (Vol. 9, p. 1440248). https://doi.org/10.3389/feduc.2024.1440248

Proudfoot, D., Kay, A. C., & Koval, C. Z. (2015). A gender bias in the attribution of creativity: Archival and experimental evidence for the perceived association between masculinity and creative thinking. *Psychological Science, 26*(11), 1751-1761. https://doi.org/10.1177/0956797615598739

Qi, S., Bai, X., & Li, Y. (2022). Competing within and beyond the boundary: Joint effect of intergroup and intragroup competitions on group creativity. *Thinking Skills and Creativity, 46*, 101194. https://doi.org/10.1016/j.tsc.2022.101194

Ramzaninezhad, R., Keshtan, M. H., Shahamat, M. D., & Kordshooli, S. S. (2009). The relationship between collective efficacy, group cohesion and team performance in professional volleyball teams. *Brazilian Journal of Biomotricity, 3*(1), 31-39. https://www.redalyc.org/pdf/930/93012686005.pdf

Reiter-Palmon, R. (2018). Creative Cognition at the Individual and Team Levels. In *The Nature of Human Creativity* (pp. 184-208). Cambridge University Press. https://doi.org/10.1017/9781108185936.015

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 144-152. https://doi.org/10.1037/aca0000227

Reiter-Palmon, R., Kramer, W., Allen, J. A., Murugavel, V. R., & Leone, S. A. (2021). Creativity in virtual teams: A review and agenda for future research. Creativity. *Theories–Research-Applications, 8*(1), 165-188. https://doi.org/10.2478/ctra-2021-0011

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2006). Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection. *Journal of Experimental Social Psychology, 42*(2), 244-251. https://doi.org/10.1016/j.jesp.2005.04.005

Ritter, B. A., & Yoder, J. D. (2004). Gender differences in leader emergence persist even for dominant women: An updated confirmation of role congruity theory. *Psychology of Women Quarterly*, *28*(3), 187-193. https://doi.org/10.1111/j.1471-6402.2004.00135.x

Rodeghero, P., Zimmermann, T., Houck, B., & Ford, D. (2021, May). Please turn your cameras on: Remote onboarding of software developers during a pandemic. *In 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 41-50). https://doi.org/10.1109/ICSE-SEIP52600.2021.00013

Rodet, C. S. (2021). Experiments on creativity and work design. *Journal of Economics & Management Strategy, 30*(3), 600-613. https://doi.org/10.1111/jems.12417

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied* (80), Article 1, 1–28. https://doi.org/10.1037/h0092976

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers, 3*(2), 135-146. https://www.jstor.org/stable/2662082

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66-75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A., Millar, G., Acar, S., & Cramond, B. (2010). Torrance tests of creative thinking as predictors of personal and public achievement: A fifty-year follow-up. *Creativity Research Journal, 22*(4), 361-368. https://doi.org/10.1080/10400419.2010.523393

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*(3), 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54-67. https://doi.org/10.1006/ceps.1999.1020

Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology, 45*(4), 736-750. https://doi.org/10.1037/0022-3514.45.4.736

Santanen, E. L., Briggs, R. O., & Vreede, G. J. D. (2004). Causal relationships in creative problem solving: Comparing facilitation interventions for ideation. *Journal of Management Information Systems, 20*(4), 167-198. https://doi.org/10.1080/07421222.2004.11045783

Sawyer, R. K. (2011). *Explaining Creativity: The Science of Human Innovation.* New York, NY: Oxford university press. https://doi.org/10.1093/oso/9780195161649.001.0001

Schock, A. K., Gruber, F. M., Scherndl, T., & Ortner, T. M. (2019). Tempering agency with communion increases women's leadership emergence in all-women groups: Evidence for role congruity theory in a field setting. *The Leadership Quarterly, 30*(2), 189-198. https://doi.org/10.1016/j.leaqua.2018.08.003

Schumpeter, J. A. (2008). *Capitalism, Socialism, Democracy.* New York: HarperCollins.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58*(8), 1438-1457. https://doi.org/10.1287/mnsc.1110.1509

Selten, R. (1989). Evolution, learning and economic behavior. *Games and Economic Behavior, 3*, 3-24.

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach. *Journal of Economic Behavior & Organization, 7*(1), 47-70. https://doi.org/10.1016/0167-2681(86)90021-1

Sheng, C. W., Tian, Y. F., & Chen, M. C. (2010). Relationships among teamwork behavior, trust, perceived team support, and team commitment. *Social Behavior and Personality: An International Journal, 38*(10), 1297-1305. https://doi.org/10.2224/sbp.2010.38.10.1297

Short, J. A., Williams, E., & Christie, B. (1976). *The Social Psychology of Telecommunications.* London: Wiley.

Shurchkov, O. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association, 10*(5), 1189-1213. https://doi.org/10.1111/j.1542-4774.2012.01084.x

Siegrist, M., Cvetkovich, G., & Gutscher, H. (2002). Risk preference predictions and gender stereotypes. *Organizational Behavior and Human Decision Processes, 87*(1), 91-102. https://doi.org/10.1006/obhd.2001.2956

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity, 6*(1), 24-30. https://doi.org/10.1016/j.tsc.2010.06.001

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., & 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), 68-85. https://doi.org/10.1037/1931-3896.2.2.68

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69-81. https://doi.org/10.1016/j.jrp.2017.02.004

Städter, S., Roider, A., Dowling, M., & Helm, R. (2022). Differences in team performance: Gender versus ability. *Social Science Quarterly, 103*(4), 926-933. https://doi.org/10.1111/ssqu.13164

Stanworth, C. (1998). Telework and the information age. *New Technology, Work and Employment, 13*(1), 51-62. https://doi.org/10.1111/1468-005X.00038

Statistisches Bundesamt. (2026). Studierende in Mathematik, Informatik, Naturwissenschaft und Technik-Fächern (MINT) [Data set]. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/studierende-mint-faechern.html

Sternberg, R. J., Glăveanu, V., & Kaufman, J. C. (2024). In quest of creativity: Three paths toward an Elusive Grail. *Creativity Research Journal, 36*(1), 155-175. https://doi.org/10.1080/10400419.2022.2107299

Stokols, D., Clitheroe, C., & Zmuidzinas, M. (2002). Qualities of work environments that promote perceived support for creativity. *Creativity Research Journal, 14*(2), 137-147. https://doi.org/10.1207/S15326934CRJ1402_1

Sutter, M., & Strassmair, C. (2009). Communication, cooperation and collusion in team tournaments-an experimental study. *Games and Economic Behavior, 66*(1), 506-525. https://doi.org/10.1016/j.geb.2008.02.014

Tauer, J. M., & Harackiewicz, J. M. (2004). The effects of cooperation and competition on intrinsic motivation and performance. *Journal of Personality and Social Psychology, 86*(6), 849-861. https://doi.org/10.1037/0022-3514.86.6.849

Terwiesch, C., & Ulrich, K. (2009). *Innovation Tournaments: Creating and Selecting Exceptional Opportunities*. Harvard Business Press.

Thatcher, S. M., & Brown, S. A. (2010). Individual creativity in teams: The importance of communication media mix. *Decision Support Systems, 49*(3), 290-300. https://doi.org/10.1016/j.dss.2010.03.004

Thorpe, S. J. (2016). Online facilitator competencies for group facilitators. *Group Facilitation: A Research & Applications Journal, 13*, 79-90.

Torrance, E. P. (1966). *The Torrance Tests of Creative Thinking: Norms-Technical Manual. Research Edition. Verbal Tests, Forms A and B-Figural Tests, Forms A and B.* Princeton, NJ: Personnel Press.

Torrance, E. P. (1974). *The Torrance Tests of Creative Thinking-Norms-Technical Manual Research Edition-Verbal Tests, Forms A and B- Figural Tests, Forms A and B.* Princeton, NJ: Personnel Press.

Treffinger, D. J., & Isaksen, S. G., Stead-Dorval, K. B. (2006). *Creative problem solving: An Introduction*. New York: Routledge. https://doi.org/10.4324/9781003419327

Vedres, B., & Vásárhelyi, O. (2023). Inclusion unlocks the creative potential of gender diversity in teams. *Scientific Reports, 13*(1), 13757. https://doi.org/10.1038/s41598-023-39922-9

Vij, A., Souza, F. F., Barrie, H., Anilan, V., Sarmiento, S., & Washington, L. (2023). Employee preferences for working from home in Australia. *Journal of Economic Behavior & Organization, 214*, 782-800. https://doi.org/10.1016/j.jebo.2023.08.020

Villeval, M. C. (2016). Can lab experiments help design personnel policies?. *IZA World of Labor*. https://doi.org/10.15185/izawol.318

Vroom, V. H. (1964). *Work and Motivation*. New York, NY: Wiley & Sons

Vuchkovski, D., Zalaznik, M., Mitręga, M., & Pfajfar, G. (2023). A look at the future of work: The digital transformation of teams from conventional to virtual. *Journal of Business Research, 163,* 113912. https://doi.org/10.1016/j.jbusres.2023.113912

Walls, M., Safirova, E., & Jiang, Y. (2007). What drives telecommuting? Relative impact of worker demographics, employer characteristics, and job types. *Transportation Research Record, 2010*(1), 111-120. https://doi.org/10.3141/2010-13

Wardale, D. (2013). Towards a model of effective group facilitation. *Leadership & Organization Development Journal, 34*(2), 112-129. https://doi.org/10.1108/01437731311321896

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science, 49*(4), 400-415. https://doi.org/10.1287/mnsc.49.4.400.14430

Wegge, J., Jungmann, F., Liebermann, S., Shemla, M., Ries, B. C., Diestel, S., & Schmidt, K.-H. (2012). What makes age diverse teams effective? Results from a six-year research program. *WORK: A Journal of Prevention, Assessment & Rehabilitation*, *41*(S1), 5145-5151. https://doi.org/10.3233/WOR-2012-0084-5145

Wegge, J., & Schmidt, K. H. (2009). The Impact of Age Diversity in Teams on Group Performance. In *Handbook of Managerial Behavior and Occupational Health* (pp.79-94). Edward Elgar Publishing.

Weisberg, R. W. (2006). *Creativity: Understanding Innovation in Problem Solving, Science, Invention, and the Arts.* Hoboken, NJ: John Wiley & Sons.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society, 1*(2), 125-151. https://doi.org/10.1177/0891243287001002002

West, M. A. (1990). The social psychology of innovation in groups. In M. A. West, & J. L. Farr, *Innovation and creativity at work: Psychological and Organizational Strategies* (pp. 309-333). Chichester: John Wiley & Sons.

West, M. A., & Sacramento, C. A. (2023). Creativity and innovation: The role of team and organizational climate. In *Handbook of Organizational Creativity* (pp. 317-337). Academic Press. https://doi.org/10.1016/B978-0-323-91840-4.00024-4

WIPO (2023). *The Global Gender Gap in Innovation and Creativity: An International Comparison of the Gender Gap in Global Patenting over Two Decades.* Development Studies Series. World Intellectual Property Organization, Geneva, Switzerland

Wittchen, M., Krimmel, A., Kohler, M., & Hertel, G. (2013). The two sides of competition: Competition-induced effort and affect during intergroup versus interindividual competition. *British Journal of Psychology, 104*(3), 320-338. https://doi.org/10.1111/j.2044-8295.2012.02123.x

Woodman, R. W., Sawyer, J. E., & Griffin, R. W. (1993). Toward a theory of organizational creativity. *Academy of Management Review, 18*(2), 293-321. https://doi.org/10.2307/258761

Wooten, J. O., & Ulrich, K. T. (2017). Idea generation and the role of feedback: Evidence from field experiments with innovation tournaments. *Production and Operations Management, 26*(1), 80-99. https://doi.org/10.1111/poms.12613

World Economic Forum. (2025). *The future of jobs report 2025*. Geneva: World Economic Forum.

Wróbel, A. E., Johansen, M. K., Jørgensen, M. S., & Cash, P. (2021). Facilitating creativity: Shaping team processes. *Creativity and Innovation Management, 30*(4), 742-762. https://doi.org/10.1111/caim.12465

Yang, Y., & Konrad, A. M. (2011). Diversity and organizational innovation: The role of employee involvement. *Journal of Organizational Behavior, 32*(8). 1062-1083. https://doi.org/10.1002/job.724

Yong, K., Sauer, S. J., & Mannix, E. A. (2014). Conflict and creativity in interdisciplinary teams. *Small Group Research, 45*(3), 266-289. https://doi.org/10.1177/1046496414530789

Yuan, Y., Humphrey, S. E., & van Knippenberg, D. (2022). From individual creativity to team creativity: A meta-analytic test of task moderators. *Journal of Occupational and Organizational Psychology, 95*(2), 358-404. https://doi.org/10.1111/joop.12380

Zhang, S., Moeckel, R., Moreno, A. T., Shuai, B., & Gao, J. (2020). A work-life conflict perspective on telework. *Transportation Research Part A: Policy and Practice, 141*, 51-68. https://doi.org/10.1016/j.tra.2020.09.007

Zhang, W., Jex, S. M., Peng, Y., & Wang, D. (2017). Exploring the effects of job autonomy on engagement and creativity: The moderating role of performance pressure and learning goal orientation. *Journal of Business and Psychology, 32*(3), 235-251. https://doi.org/10.1007/s10869-016-9453-x

Zhang, Q., Zhou, K., Zhang, Y., Zhou, P., & Zhang, Z. (2024). Reexamining the relationship between extrinsic rewards and creativity: a study from the creative process engagement perspective. *The International Journal of Human Resource Management, 35*(15), 2514-2541. https://doi.org/10.1080/09585192.2024.2347645

Zhou, J. (2003). When the presence of creative coworkers is related to creativity: role of supervisor close monitoring, developmental feedback, and creative personality. *Journal of Applied Psychology, 88*(3), 413-422. https://doi.org/10.1037/0021-9010.88.3.413

Ziegler, R., Diehl, M., & Zijlstra, G. (2000). Idea production in nominal and virtual groups: Does computer-mediated communication improve group brainstorming?. *Group Processes & Intergroup Relations, 3*(2), 141-158. https://doi.org/10.1177/1368430200032003